CLARIN

Selected paper from the

CLARIN Annual Conference 2018

Pisa, 8-10 October 2018

Selected papers from the

CLARIN Annual Conference 2018

Pisa, 8-10 October 2018

edited by Inguna Skadiņa, Maria Eskevich

Front cover illustration:
picture composition by CLARIN ERIC

Linköping Electronic Conference Proceedings 159
eISSN 1650-3740 (Online) • ISSN 1650-3686 (Print) 2019
ISBN 978-91-7685-034-3

Introduction

Franciska de Jong
Executive Director CLARIN ERIC
Universiteit Utrecht, The Netherlands
f.m.g.dejong@uu.nl

Inguna Skadiņa
Institute of Mathematics and
Computer Science
University of Latvia
Riga, Latvia
Programme Committee Chair
inguna.skadina@lumii.lv

This volume presents the highlights of the 7^{th} CLARIN Annual Conference 2018 held in Pisa, Italy, on 8^{th}—10^{th} October 2018.

CLARIN ERIC[1] is the European Research Infrastructure for Language Resources and Technology aimed at supporting researchers from the Social Sciences and Humanities (SSH) and beyond in their use of language data and technologies. CLARIN works towards lowering barriers in doing research by giving access to language resources distributed across the countries involved in the infrastructure and by offering advanced, user-friendly and effective applications that enable the analysis of textual data, speech recordings, as well as multimodal data in different research tasks.

Since the establishment of the ERIC in 2012, CLARIN has considerably grown in size. Currently there are 20 member countries and more than 100 associated research institutions who are all encouraged and supported to be represented at the annual conference which is meant to be a central event for CLARIN community. At the conference consortia from all participating countries and the various communities of use meet in order to exchange ideas, experiences and best practices in using the CLARIN infrastructure. The conference covers a wide range of topics, including the design, construction and operation of the CLARIN infrastructure, the data, tools and services that are or could be on offer, its actual use by researchers, its relation to other infrastructures and projects, and the CLARIN Knowledge Sharing Infrastructure. The aim is to attract researchers from all the various SSH fields that work with language materials, i.e. the people who are the raison d'être for CLARIN.

For the 7^{th} edition of the CLARIN Annual Conference the special topic of the CLARIN Annual Conference was "Multimedia, Multimodality and Speech". Early in 2018 a call[2] was issued for which 75 abstracts were submitted. The number of submissions was much higher than in previous years, demonstrating the increase of interest in CLARIN activities, language resources and tools, and of their use by scholars from the humanities and social sciences. At least one abstract was submitted by each of the CLARIN ERIC countries. There were also submissions from countries outside the consortium, including Morocco, Russia, South Africa and the United States, underlining the relevance of the CLARIN infrastructure outside Europe.

The three topics that attracted the most of proposals were (a) design and construction of the CLARIN infrastructure, (b) the use of the CLARIN infrastructure, and (c) the special topic of this year: multimedia, multimodality, speech. For the latter topic 16 abstracts were accepted into the final programme, which is underlining the broad interest in the various methods and techniques for the collection, annotation and processing of audio, visual or multimedia data with language as an important part of the content.

All submissions were reviewed anonymously by three reviewers (PC members and reviewers invited by PC members). Out of the 75 submitted abstracts 44 submissions were accepted for presentation at the conference (acceptance rate 0.58). The accepted contributions were published in the online Proceedings of the Conference[3]. A novelty introduced at the 2018 edition of the CLARIN Annual Conference was a well received student poster session with 16 presentations by PhD students who were selected by the National Coordinator of their country. The abstracts of the student presentations were published in the

[1] http://clarin.eu
[2] https://www.clarin.eu/news/call-papers-clarin-annual-conference-2018
[3] https://office.clarin.eu/v/CE-2018-1292-CLARIN2018_ConferenceProceedings.pdf

Franciska de Jong and Inguna Skadiņa 2019. Introduction. *Selected papers from the CLARIN Annual Conference 2018*. Linköping Electronic Conference Proceedings 159: i–iii.

online CLARIN 2018 Book of Abstracts[4].

The conference hosted two invited talks related to the special topic of the conference - Multimedia, Multimodality and Speech. Professor James Pustejovsky (Brandeis University, USA) presented CLAMS (Computational Linguistic Applications for Multimedia Services), a platform for archivists. He demonstrated the functionality of the platform on a subset of the American Archive of Public Broadcasting[5] data through the available tools for speech processing and computational linguistic analysis.Dr. Costanza Navarretta (University of Copenhagen, Denmark) presented studies of multimodal communication from a computational linguistic point of view, focusing on the collection and annotation of multimodal corpora and research conducted on these data at the Centre for Language Technology (Denmark).

In addition, on the event page[6] CLARIN published a rich set of materials related to the conference :

- Complete conference programme and most of the slides presented: `https://www.clarin.eu/content/programme-clarin-annual-conference-2018`
- Recordings of most talks; the two invited lectures and several other video materials are available on a dedicated channel of VideoLectures: `http://videolectures.net/clarinannualconference2018_pisa/`

After the conference, the authors of the accepted papers and student submissions were invited to submit full versions of their papers to be considered for the post-conference proceedings volume. The papers were anonymously reviewed, each by three PC members. We received 26 (including 3 student papers) full length submissions, out of which 23 (including 2 student papers) were accepted for this volume. All the main topics addressed at the conference are covered in this volume as well.

We would like to thank all PC members and reviewers for their efforts in evaluating and re-evaluating the submissions, Maria Eskevich from CLARIN Office for her indispensable support in the process of preparing these proceedings, and Peter Berkesand at Linköping University Electronic Press, who (as usual) has ensured that the digital publication of this volume came about smoothly.

Members of the Programme Committee for the CLARIN Annual Conference 2018:

- Lars Borin, Språkbanken, University of Gothenburg, Sweden
- António Branco, Universidade de Lisboa, Portrugal
- Griet Depoorter, Institute for the Dutch Language, The Netherlands/Flanders
- Koenraad De Smedt, University of Bergen, Norway
- Jens Edlund, KTH Royal Institute of Technology, Sweden
- Tomaž Erjavec, Dept. of Knowledge Technologies, Jožef Stefan Institute, Slovenia
- Francesca Frontini, University of Montpellier, France
- Eva Hajičová, Charles University, Czech Republic
- Erhard Hinrichs, University of Tübingen, Germany
- Nicolas Larrousse, Huma-Num, France
- Krister Lindén, University of Helsinki, Finland
- Bente Maegaard, University of Copenhagen, Denmark
- Monica Monachini, Institute of Computational Linguistics “A. Zampolli”, Italy

[4] `https://www.clarin.eu/clarin-annual-conference-2018-abstracts`
[5] `http://americanarchive.org`
[6] `https://www.clarin.eu/event/2018/clarin-annual-conference-2018-pisa-italy`

- Karlheinz Mörth, Institute for Corpus Linguistics and Text Technology, Austrian Academy of Sciences, Austria
- Jan Odijk, Utrecht University The Netherlands
- Maciej Piasecki, Wrocław University of Science and Technology, Poland
- Stelios Piperidis, Institute for Language and Speech Processing (ILSP), Athena Research Center, Greece
- Kiril Simov, IICT, Bulgarian Academy of Sciences, Bulgaria
- Inguna Skadiņa, Institute of Mathematics and Computer Science, University of Latvia Latvia (Chair)
- Marko Tadič , University of Zagreb, Croatia
- Jurgita Vaičenonienė, Vytautas Magnus University, Lithuania
- Tamás Váradi, Research Institute for Linguistics, Hungarian Academy of Sciences, Hungary
- Kadri Vider, University of Tartu, Estonia
- Martin Wynne, University of Oxford, United Kingdom

Additional reviewers of this volume:

- Aleksei Kelli, University of Tartu, Estonia
- Neeme Kahusk, University of Tartu, Esdtonia
- Vincent Vandeghinste, the Dutch Language Institute, The Netherlands
- Bob Boelhouwer, the Dutch Language Institute, The Netherlands

Contents

Lärka: From Language Learning Platform to Infrastructure for Research on Language Learning

David Alfter Lars Borin Ildikó Pilán
Språkbanken, University of Gothenburg, Sweden
david.alfter|lars.borin|ildiko.pilan@gu.se

Therese Lindström Tiedemann
Department of Finnish, Finno-Ugrian
and Scandinavian Studies
University of Helsinki, Finland
therese.lindstromtiedemann@helsinki.fi

Elena Volodina
Språkbanken
University of Gothenburg, Sweden
elena.volodina@gu.se

Abstract

Lärka is an Intelligent Computer-Assisted Language Learning (ICALL) platform developed at Språkbanken, as a flexible and a valuable source of additional learning material (e.g. via corpus-based exercises) and a support tool for both teachers and L2 learners of Swedish and students of (Swedish) linguistics. Nowadays, Lärka is being adapted into a building block in an emerging second language research infrastructure within a larger context of the text-based research infrastructure developed by the national Swedish Language bank, Språkbanken, and SWE-CLARIN.

Lärka has recently received a new responsive user interface adapted to different devices with different screen sizes. Moreover, the system has also been augmented with new functionalities. These recent additions aim at improving the usability and the usefulness of the platform for pedagogical purposes. The most important development, though, is the adaptation of the platform to serve as a component in an e-infrastructure supporting research on language learning and multilingualism. Thanks to Lärka's *service-oriented architecture*, most functionalities are also available as web services which can be easily re-used by other applications.

1 Introduction

Lärka[1] is an Intelligent Computer-Assisted Language Learning (ICALL) platform developed at the CLARIN B Center Språkbanken Text (University of Gothenburg, Sweden). Lärka development started in the project *A system architecture for ICALL* (Volodina et al., 2012), the initial goal being to re-implement a previous tool, ITG, used up until then for teaching grammar (Borin and Saxena, 2004) with modern technology. The new application, Lärka, gradually developed into a platform for language learning covering two groups of learners – second/foreign language learners of Swedish and students of (Swedish) linguistics. Lärka is an openly available web-based tool that builds on a variety of existing SWE-CLARIN language resources such as Korp (Borin et al., 2012) for querying corpora, Karp (Borin et al., 2013b) for querying lexical resources and language technology tools (Borin et al., 2017). Thanks to its service-oriented architecture, Lärka functionalities can be re-used in other applications (Volodina et al., 2014b).

In parallel to exercise generation functionalities, Lärka has been evolving into a research tool with a number of supportive modules for experimentation and visualization of research results, such as for selection of best corpus examples for language learners, for readability analysis of texts aimed at or produced by language learners, for prediction of single-word lexical difficulty, as well as for facilitating text-level annotation of language learner corpora, but also to collect data from exercises where learner interaction with the platform and their input have been used in research on metalinguistic awareness. Lärka is actively used in teaching grammar to university students, where we can report only those uses that we have explicitly been told about. As we do not require login to the platform, we do not know who our users are, but we can deduce from the logs that Lärka is being used beyond the reported schools and universities.

[1] https://spraakbanken.gu.se/larka

David Alfter, Lars Borin, Ildikó Pilán, Therese Lindström Tiedemann and Elena Volodina 2019. From Language Learning Platform to Infrastructure for Research on Language Learning. *Selected papers from the CLARIN Annual Conference 2018*. Linköping Electronic Conference Proceedings 159: 1–14.

Nowadays, Lärka is being adapted into a building block in an emerging second language research infrastructure SweLL (Volodina et al., 2018), within a larger context of the text-based research infrastructure developed by the national Swedish Language bank, Språkbanken, and SWE-CLARIN. This addresses an obvious need within CLARIN, as evidenced by the interest in the recent CLARIN workshop on "Interoperability of Second Language Resources and Tools".[2]

The current paper describes the new version of Lärka that was released in 2016, replacing the 2013 version, and illustrates improved and newly added functionalities.

2 Related work

There have been some attempts to combine exercise platforms with different types of data collection. The *Writing Mentor* Google Docs add-on, for example, allows users to get feedback on their writing in different categories such as coherence, topic development or use of sources to back up claims. The application uses natural language processing tools to provide users with feedback but at the same time collects the texts and all subsequent modifications to the texts that have been analyzed (Madnani et al., 2018). However, accessibility of the data for SLA research is limited.

The FeedBook project (Rudzewitz et al., 2017) is based on an English text book and presents the text book in a digitized interactive web platform that has been enriched with natural language processing to provide immediate fine-grained feedback to the students concerning both form and meaning errors. Teachers can also see their students' progress and provide individual feedback. The data is logged and is used iteratively for further improvement of the system, the data access so far being limited to the researchers involved in the project.

Most applications, however, are purely pedagogical. An outstanding example is the *Language Muse* Activity Palette (Burstein and Sabatini, 2016; Burstein et al., 2017). It allows teachers to upload texts and automatically generates exercises based on these texts. Texts are analyzed using natural language processing algorithms to identify different linguistic features such as multi-word expressions, syntactic relations and discourse structure. Based on the analysis, the platform creates over twenty different activities for the teacher to choose from, such as antonym exercises, homonym exercises or verb tense exercises. Teachers have full control over which texts are used, and are offered a possibility to edit automatically suggested exercise items. In that way, teachers can build a 'palette' of activities from the original text that best suits their and their students' needs.

Perez and Cuadros (2017) propose a framework for automatic exercise generation from user-specified texts that works with Spanish, Basque, English and French. Users can use texts of their own choosing in four different languages. The framework can generate three different kinds of tasks, namely gap exercises, multiple-choice exercises and sentence rearrangement exercises. Furthermore, the framework automatically generates hints for the gap exercise and allows for the adjustment of the number of distractors for multiple-choice exercises. Exercises are also exportable in Moodle's CLOZE[3] format, increasing its appeal.

On the other hand, there are multiple examples of SLA and psycholinguistic experiments that are staged through exercises that elicit certain types of data from language learners – data that helps researchers to address particular research questions, e.g. Andersson et al. (2018) investigating the influence of the native language on the processing of the word order in Swedish or Kerz and Wiechmann (2017) studying individual differences in L2 processing of multi-word phrases.

We argue that exercise generation platforms/applications have a capacity to mediate between language learners and researchers, bringing interests of the two groups together. We aim to foster this collaboration through the Lärka platform.

Lärka started as an exercise generation platform for learners of Swedish, and later it was extended to support the development and visualization of new algorithms in support of language learning. Now we are taking a new direction, combining research interests from Second Language Acquisition (SLA),

[2] See https://sweclarin.se/eng/workshop-interoperability-l2-resources-and-tools

[3] https://docs.moodle.org/23/en/Embedded_Answers_(Cloze)_question_type

Learner Corpus Research (LCR) and language learning into one and building an infrastructure supporting the collection of L2 data through exercises.

In the next sections, we delineate how Lärka can be and is used as a pedagogical tool in teaching students of Swedish linguistics (Sections 3.1 and 4), the different exercises in support of research aimed at L2 Swedish (Section 3.2), and the various components that constitute the research infrastructure facet of the platform (Section 5).

3 Lärka for learning and teaching

One of the main functionalities of Lärka is the automatic generation of exercises based on real-life authentic language examples from corpora. Exercise generation is aimed at two groups of learners: students of (Swedish) linguistics and learners of Swedish as a second or foreign language (L2).

3.1 Exercises for students of Linguistics

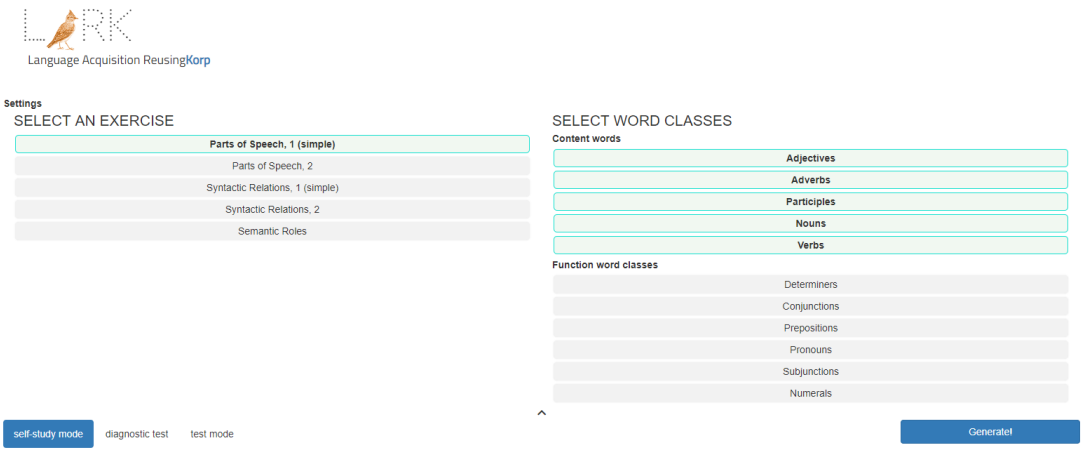


Figure 1: Exercises for linguists

Students learning grammatical analysis are in constant need of exercises and feedback on their analysis. Lärka offers exercises for linguistic analysis of parts of speech (word classes), syntactic relations and semantic roles. The exercises are based on authentic texts, which can make them more difficult than textbook examples. However, they are authentic examples of the type of texts the students are expected to be able to analyze in the future. Through on-the-spot feedback, students' learning is enhanced, especially if exercises are done at least partly in a class room setting with the possibility of consulting a teacher and/or the possibility of discussing one's analysis with a fellow student and together trying to make sense of why the automatic feedback said that they got it right or wrong (Lindström Tiedemann et al., 2016).

As mentioned above, Lärka offers students 3 types of exercises: parts of speech, syntactic relations and semantic roles. The first two offer two levels of difficulty (beginner and intermediate), whereas the third exercise, semantic roles, is only available as one level (Figure 1).

Exercises are presented as shown in Figure 2 with a sentence and a word or phrase highlighted in another colour. The learner then has to select from a multiple choice drop down box which answer is correct given the highlighted word or phrase. In part-of-speech exercises for example, learners have to select the correct part of speech for the highlighted word.

The exercises are available in three different modes: self-study, diagnostic test or test. Students can choose whichever mode they want to use. In self-study mode, answers can be revised as often as desired and needed, even after submitting the answer. In this way, if the answer was incorrect, it is possible to find the correct answer. In test mode, answers cannot be changed after submitting and the correct answer will be shown immediately after submitting. In diagnostic mode, as in test mode, answers cannot be changed after submitting. In addition, the number of exercise items is limited to three items of each main category, e.g. the part-of-speech exercise type covers eleven part-of-speech categories, resulting in a total of 33 diagnostic exercise items. Exercise generation stops after completion of all items in diagnostic

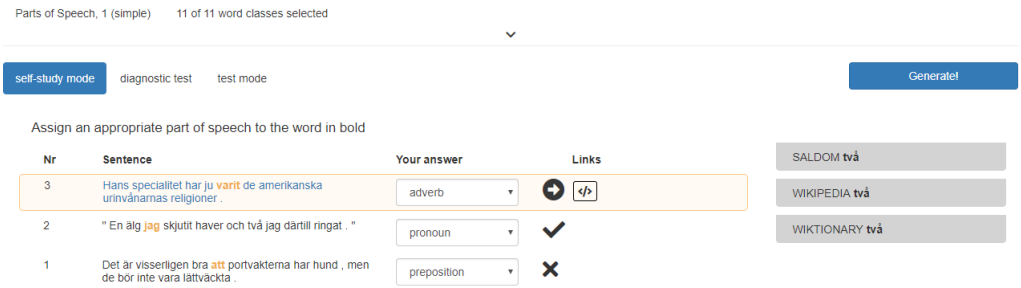


Figure 2: Execises for linguists

mode and a summary is provided which can be emailed to the teacher for further comments or to oneself in order to study the examples further or to be able to track one's learning. In contrast, the other two modes generate exercises infinitely. In both self-study and test mode the actual categories practiced can also be chosen (e.g. one can select to only practice adjectives and adverbs for part-of-speech exercises), whereas the diagnostic test automatically selects all available categories.

In order to avoid exercise item repetition, a sentence will be shown only once during the same session.

3.2 Exercises for language learners

Lärka offers a number of exercises for learners of L2 Swedish as illustrated in the following paragraphs. For all learner exercises, target vocabulary items are sampled from SVALex (François et al., 2016) and SweLLex (Volodina et al., 2016b). SVALex presents a list of lemmata occurring at the different CEFR (Common European Framework of Reference for Languages (Council of Europe, 2001)) levels in the textbook corpus COCTAILL (Volodina et al., 2014a). Similarly, SweLLex is based on the pilot SweLL corpus (Volodina et al., 2016a), a corpus of learner essays. We map each distribution to a single CEFR level according to two approaches, namely *first-occurrence* (Gala et al., 2013; Gala et al., 2014) and *threshold* (Alfter et al., 2016).

The exercises target lexical knowledge of Swedish L2 learners, and speaking pedagogically, train lexical knowledge from various points of view, namely: listening and spelling of lexical items, recognition of an appropriate item for a given context, morphological inflectional behaviour of individual lexical items, and linking definitions/translations with words. There are certainly a many other conceivable exercises that target different word knowledge aspects that we have not implemented. While even the exercise types that we currently offer are still in need of evaluation with teachers and learners, we do believe that they are useful. The session logs for the listening and spelling and word guess exercises show that there is interest in these types of exercises.

3.2.1 Vocabulary and inflection

Vocabulary exercises and inflection exercises have a multiple-choice format. Each item consists of a sentence containing a gap, as well as a list of five answer alternatives, of which one is correct and four are *distractors*, i.e. incorrect options (Figure 3). For vocabulary, distractors are chosen of the same word class as the target word. This morphological selection is further restricted by requiring that distractors be of the same number and/or definiteness as the target item for nouns or the same voice and/or tense for verbs. In case the restriction on the distractors returns too few results, these constraints can be relaxed or dropped.

For inflection exercises, we look up all morphological forms of the target word in Saldo's morphology (Borin et al., 2013a) and use a subset of those as distractors. Figures 3 and 4 show the vocabulary and inflection multiple choice exercise respectively.

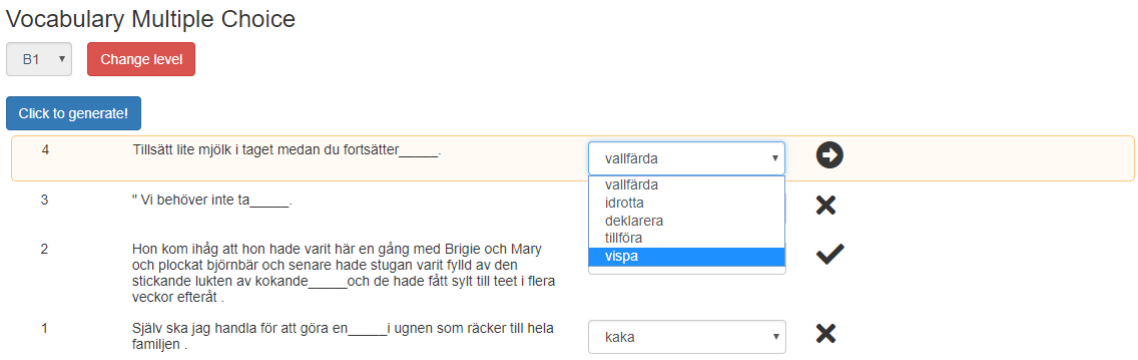


Figure 3: Vocabulary multiple choice

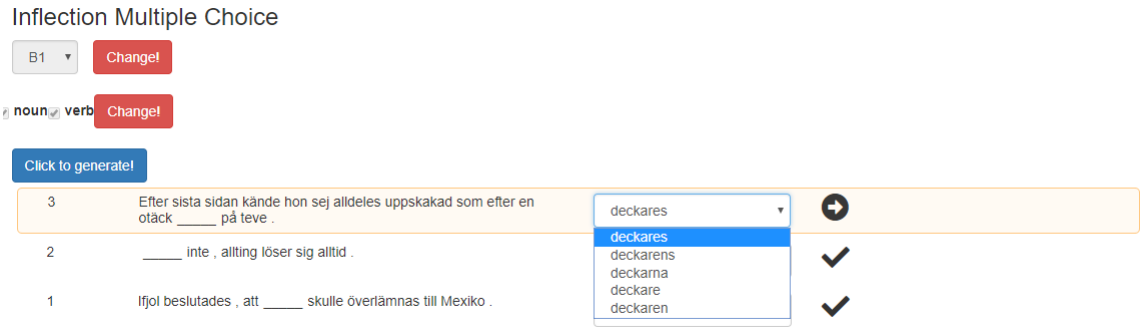


Figure 4: Inflection multiple choice

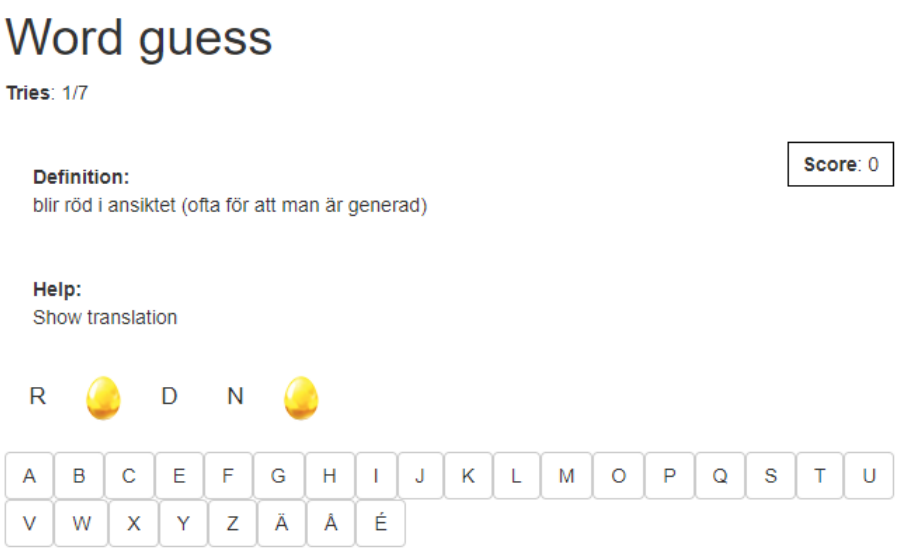


Figure 5: Word guess

3.2.2 Word guess

A recent addition to our platform is a simple word-level exercise, *Word guess*, that takes a step towards gamified learning. Word guess re-implements the well-known Hangman game format: users are presented with a number of hidden characters and the definition of the word in Swedish, and their task is to guess letters contained in the word, which eventually helps them guess the word itself, as shown in Figure 5.

Every time the guessed character is not in the word, users receive penalty points. In our learning-oriented version of the game, users can choose to receive clues such as the translation of the word (into a range of different languages). Both the definition and translations are retrieved from *Lexin*, a core-vocabulary lexicon for immigrants (Gellerstam, 1999). This game is a simple example of reusing information from lexical resources for gamified language learning activities.

3.2.3 Liwrix

Another exercise is the listening exercise *Liwrix* (Volodina and Pijetlovic, 2015). This exercise makes use of Text-to-Speech (TTS) technology by SitePal[4] to dynamically generate audio of single words and multi-word expressions. In the future, we also intend to include phrases and sentences, as was done in the previous version of Lärka. The delay is caused by the newly introduced hint system which needs to be modified in order to work with phrases and sentences.

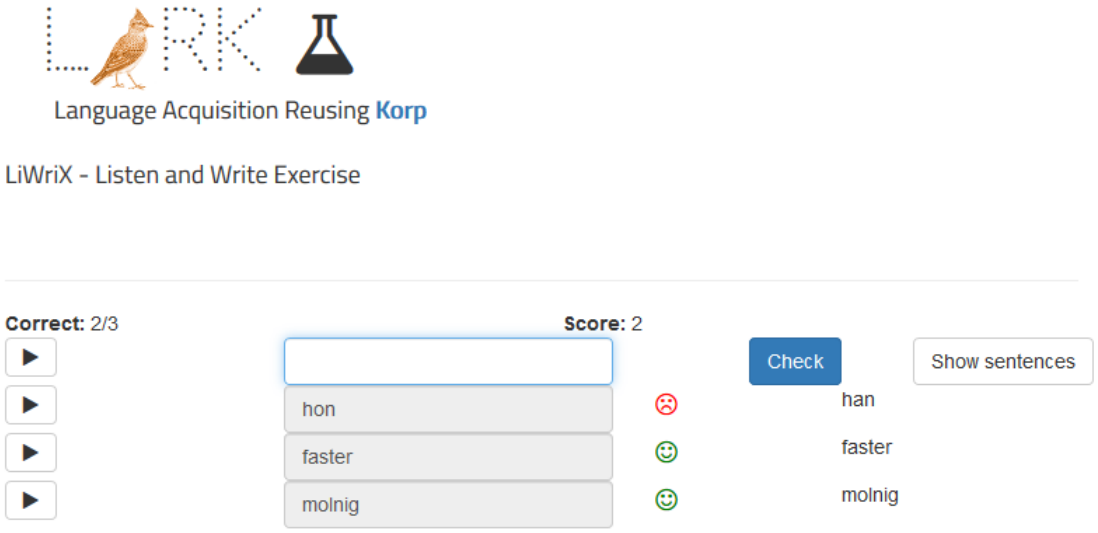


Figure 6: Liwrix

Figure 6 shows the exercise. By clicking on the button on the left, a word or multi-word expression is played and the answer is to be entered into the textfield. In addition, hints are available: As a first hint, but also to avoid problems with homonyms or possible mispronunciations, users can get “clues" in the form of a number of sentences in which the word(s) to be guessed appear in context. As a second hint, learners can choose to have the initial letter of the target word revealed.

Feedback is given in the form of a green smiley if the answer was correct and a red smiley if the answer was incorrect. In test mode (as in Figure 6) the correct answer is also shown irrespective of the correctness of the learner input.

4 Lärka in practice

Lärka for linguists has been used in introductions to grammar and linguistics in Sweden and Finland (Volodina et al., 2014b; Lindström Tiedemann et al., 2016). In Uppsala the platform was often used in lab sessions first so that students had a chance to consult a teacher when they had questions and they were also encouraged to discuss their analysis and the automatic feedback they got with their fellow students.

In Helsinki students have sometimes been encouraged to use it independently on courses in Swedish grammar where they have then been asked to hand in some of their analysis to their teacher or simply been told to use it to get more practice which is something they clearly cannot get too much of in learning grammatical analysis. Some exercise books might not even come with a key, which means that all exercises must be treated in class if the students are to find out what they did right or wrong.

[4] sitepal.com

In comparison, Lärka material is better suited in this case than many exercise books since it provides authentic texts accompanied by immediate automatic feedback.

The students felt that this was of great use and definitely thought that the platform should be used in the future. In a study with 45 students, Lärka was generally well received. Figure 7 shows that the majority of students were in favor of keeping Lärka as part of lab sessions with 34 students (78%) responding strongly in favor of keeping Lärka (scores 5–6), while 10 students (22%) showed more reservation (scores 3–4). No students voted against keeping Lärka (scores 1–2). Similarly, Figure 8 shows that 80% of students would recommend Lärka to a fellow student while 20% showed reservations.

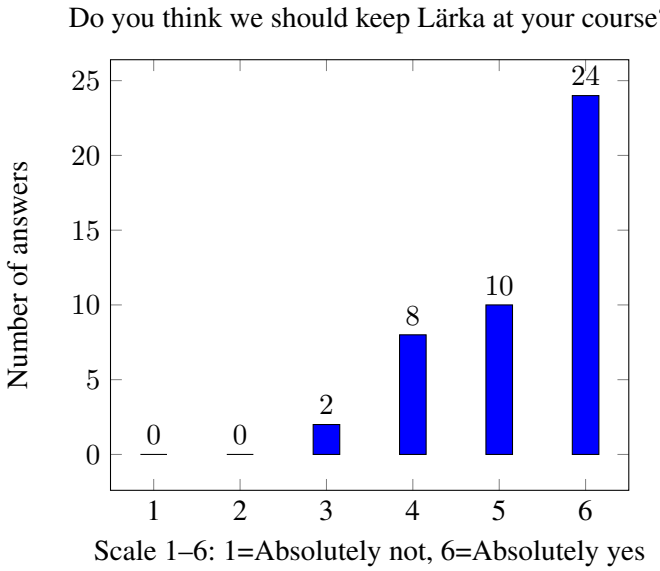


Figure 7: Evaluation results 1

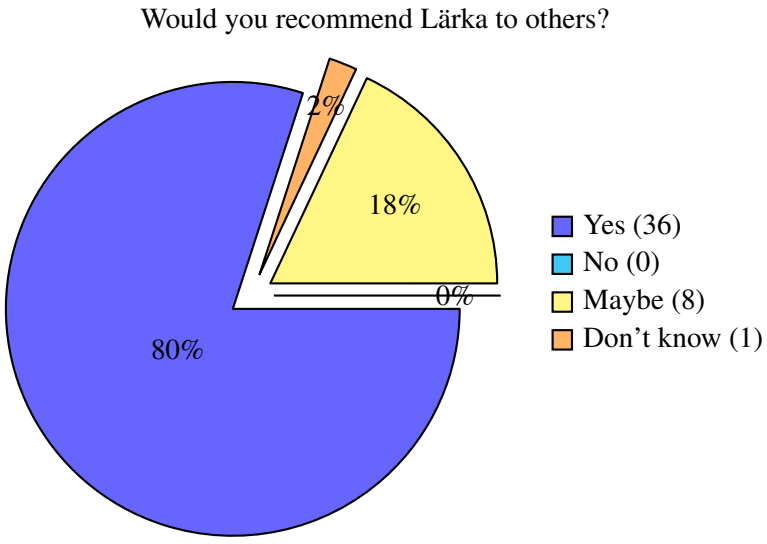


Figure 8: Evaluation results 2

A more recent analysis of the linguistic exercise log data collected through the 2016 version of the platform shows that during the time span from October 2016 to May 2018, there were 2086 sessions. One session is counted as a user using the platform from the moment of opening the page to closing it. As we do not require users to login, we create anonymous session identifiers each time a user opens the page. Thus, multiple sessions can stem from the same user. There were 126 sessions in the period from

(a)

Exercise type	# interactions
Part of Speech 1	28,544
Part of Speech 2	6,717
Semantic roles	553
Syntactic relations 1	8,426
Syntactic relations 2	2,842

(b)

Operating system	# interactions
Android	1,081
Linux	1,093
Mac OS X	11,516
Windows	18,962
iOs	2,102

Table 1: Interaction by exercise type (a) and operating system (b)

October 2016 to December 2016, 1449 sessions during 2017 and 511 sessions from January 2018 to May 2018.

During those 2086 sessions, a total of 47082 interactions were carried out. One interaction counts as an exercise item being completed. Interaction counts do not include self-corrections, mode changes or helps consulted. Table 1 (a) shows the breakdown of the interactions per exercise type. A logging feature that was added later[5] was the logging of whether the page was accessed from a mobile device and which operating system was used to access the page. The logs show that the linguist exercise was accessed 3,184 times (~10%) from a mobile device, as opposed to 31,571 times from a non-mobile device. Table 1 (b) shows the breakdown of interactions by operating system.

Furthermore, we can see that the platform was mainly accessed from Sweden (91%) and Finland (8%), but also from other countries such as the US, Poland, Germany, the Netherlands, Turkey, Estonia, the UK, India, Belgium, Switzerland, Japan, Canada and Russia, together making up the remaining 1%.

5 Lärka as research infrastructure

Lärka is being developed to serve as one of the e-infrastructure components offered to the research community by the Swedish CLARIN B-centre Språkbanken Text at the University of Gothenburg. Specifically Lärka is intended to be used as an infrastructure for research in (Swedish as) L2 acquisition. Currently Lärka offers modules for (1) collection of data from learners through their interaction with the platform, i.e. exercise logs; (2) text-level annotation of learner essays and course book texts; as well as (3) experimentation and visualization of the ongoing research in support of language learning.

With these modules, materials and exercises can be tailored drawing on vast collections of naturally occurring language, in a precise yet flexible as well as replicable way, and students' responses and reactions can be recorded in detail for subsequent quantitative and qualitative analysis. In order to achieve the necessary combination of precision and flexibility, we integrate natural language processing tools and algorithms for corpus example selection, text assessment and automatic exercise generation. These aspects are described in more detail below. A recent direction is "profiling" lexical and grammatical competences that learners of Swedish have, where we experiment with different lexical resources for exercise creation, and in the near future expect to integrate research on grammar profiles.[6]

5.1 Corpus example selection

In Lärka, the automatically generated exercises for language learners rely on *HitEx* (*Hitta Exempel* 'find examples'), a tool for selecting and ranking corpus examples (Pilán et al., 2017). The main purpose of HitEx is to identify sentences from generic corpora which are suitable as exercise items for L2 learners. The suitability of the sentences is determined based on a number of parameters that reflect different linguistic characteristics of the sentences. Through a graphical user interface, it is also possible to conduct a sentence search based on parameters customized by the user. The selection criteria include a wide variety of linguistic aspects such as the desired difficulty level based on CEFR, typicality based on word

[5] That is why the total is lower than 47,082

[6] `https://spraakbanken.gu.se/eng/l2-profiling`

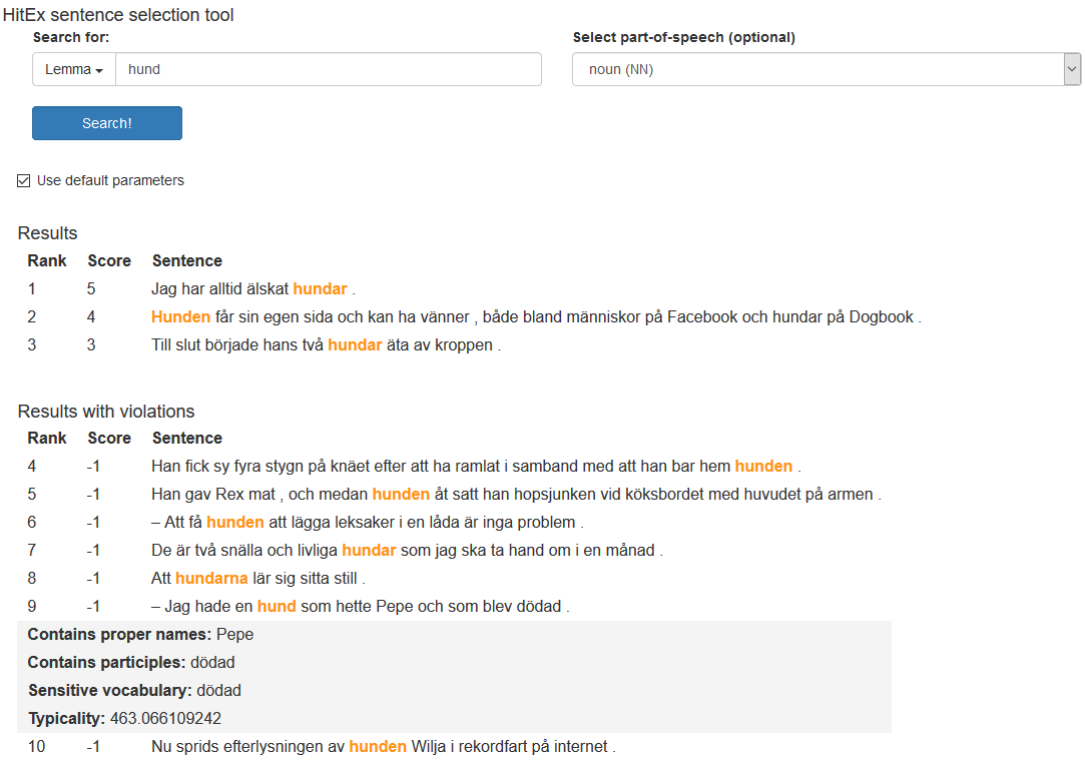


Figure 9: Corpus example selection tool HitEx: Results

co-occurrence measures, as well as the absence of anaphoric expressions and sensitive vocabulary (e.g. profanities), just to name a few. It is also possible to use a set of default parameters for searching. Figure 9 shows the results of HitEx. Sentences which fulfill all the required parameter constraints are shown on top while results that violate one or more constraints are shown under 'Results with violations'. Upon clicking on one of the sentences, more information is shown.

5.2 Text complexity evaluation

Another functionality, *TextEval*, offers an interface to automatically assess Swedish texts for their degree of complexity according to the CEFR. Texts can be either learner productions (e.g. essays) or texts written by experts as reading material for learners. The machine learning based automatic analysis returns an overall CEFR level for the text, as well as a list of linguistic indicators relevant for measuring text complexity, such as the average length of sentences and tokens, LIX score and nominal ratio. In addition, it is possible to add a color-enhanced highlighting for words per CEFR levels which provides users with a straightforward visual feedback about the lexical complexity of a text. Figure 10 shows the analysis of a text with word-level CEFR highlighting. We use the aforementioned lists SVALex and SweLLex to mark up receptive and productive vocabulary respectively. For each CEFR level, a darker and a lighter shade of the same color represents productive and receptive vocabulary respectively at the given level.

5.3 Lexical complexity prediction

Based on the word lists SVALex and SweLLex, which have been transformed so as to map each word to a single CEFR level as described in Alfter et al. (2016), we have built a module capable of predicting the complexity of any Swedish word, not only words occurring in the word lists (Alfter and Volodina, 2018). For each word, we extract both traditional word-based features such as length, number of syllables, number of homonyms and also information about topics, i.e. which topics a word belongs to. For example, the word *fisk* 'fish' would occur in the topics 'Animals' and 'Food'. We then feed a machine

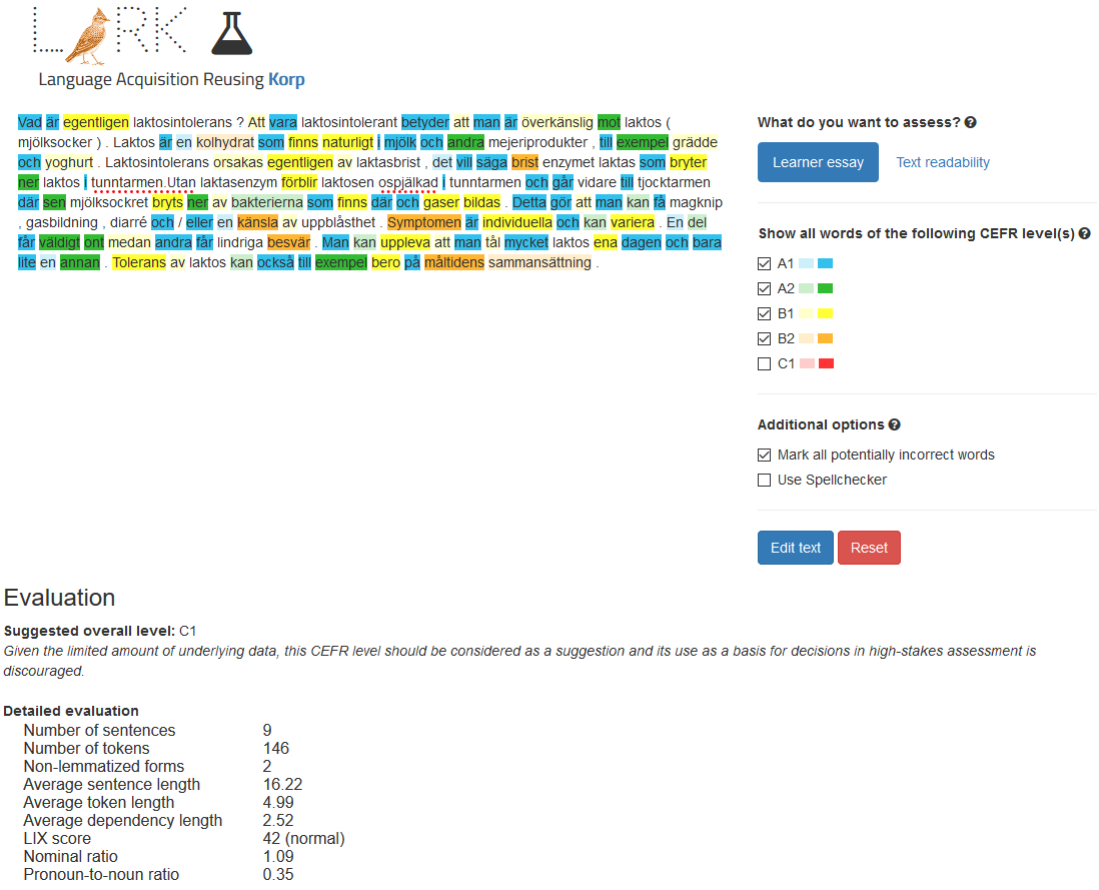


Figure 10: Text complexity evaluation tool TextEval

learning algorithm these feature vectors as well as the predicted mapped single CEFR level of the word and let the algorithm learn how to map from these features to CEFR levels.

An interested user can test a bespoke interface to get predictions about the complexity of a word and its target level (receptive versus productive), as shown in Figure 11. This user interface can be used for getting predictions of any word, not only words present in the word lists. The input word is transformed into a feature vector as described above and then fed into the classifier, which predicts a label. Figure 11 shows the predictions for *hund* 'dog', *vovve* 'doggy' (childish or endearing term for 'dog') and *byracka* 'mutt' (derogatory term for 'dog').

5.4 Annotation editor

Lärka contains an annotation editor that can be used for XML markup of textbooks. The editor provides an intuitive menu that makes adding XML tags easy. The editor keeps track of current settings in order to make adding new elements as easy as possible. It also automatically increments lesson counters and other counters. The editor offers the possibility to download the annotated text as an XML file. The current version of the editor also includes the possibility to save one's progress and continue working on it at a later moment in time without the need to login. The SweLL corpus pilot project (Volodina et al., 2016a) and the COCTAILL corpus project (Volodina et al., 2014a) used a previous version of the annotation editor to achieve consistent XML markup of essays and course books as well as to simplify the annotation process by providing an intuitive and intelligent user interface.

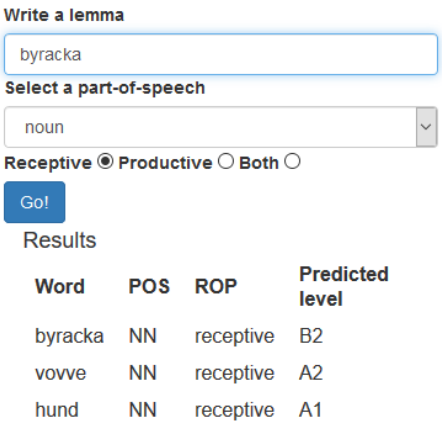


Figure 11: User interface for lexical complexity prediction

5.5 Lexicographic annotation tool

Another annotation tool that has recently been added to Lärka is the Lexicographic Annotation Tool, Legato. This tool can be used to annotate words or word senses on different lexicographic levels. Figure 12 shows the tool in the 'register' annotation mode. Here, the annotator is presented with a SALDO sense (viz. *gammal* 'old'), its part-of-speech (adjective) and the predicted CEFR level (A1). In addition, the tool shows the primary and secondary SALDO descriptors, if available. As different senses of a word can still be ambiguous as to the category to be annotated, we also show an example sentence where the word sense is highlighted, in this case surrounded by two asterisks (**). The example sentences have been selected to be of the same CEFR level as the word sense in question.

The main part of the interface shows the annotation possibilities. In the example shown, different options for register are shown. The annotator can select none, one, or more than one of these possibilities.

Finally, using the buttons at the bottom, annotators can leave the interface to annotate either another lexicographic category or to stop annotating altogether. Items can be skipped if the annotator is unsure about the annotation. In this case, the item will be added to the list of skipped items which can be accessed by clicking the button on top next to the 'Guidelines' button. This opens up a side menu which shows all the skipped items. By clicking on any of these items, the interface returns to the item in question. The interface also offers a search functionality which makes searching through the list of items easy.

In addition, the interface keeps track of different annotators and their progress across different annotation categories. Thus, if an annotator annotates ten items in 'morphology', then returns to the main screen and annotates ten items in 'nominal gender', then returns to morphology, the interface will resume at item number eleven. This also works across sessions. Thus, annotation does not have to be done in one fell swoop but can be done intermittently. The skipped items are also saved per annotator and category. For example, if annotator A skips *gammal* 'old' in 'register' but not in 'morphology', it will turn up for annotator A under 'register' until it is resolved. All data is saved to a data base on the server.

Besides fully manual annotation, the tool also offers a semi-automatic annotation mode where some of the values have been automatically extracted by linking together various resources. In this annotation mode, if values have been found, the annotator's task is to check whether the values are correct and correct them if necessary. If no values have been found, the annotator proceeds as in manual mode.

6 Ongoing work and planned extensions

Besides the activities described in this paper, the addition of new exercise formats and the implementation of a diagnostic placement test are currently under development. In the near future we plan to add a login functionality as well as an infrastructure to log more specific user data. This would enable us to create a

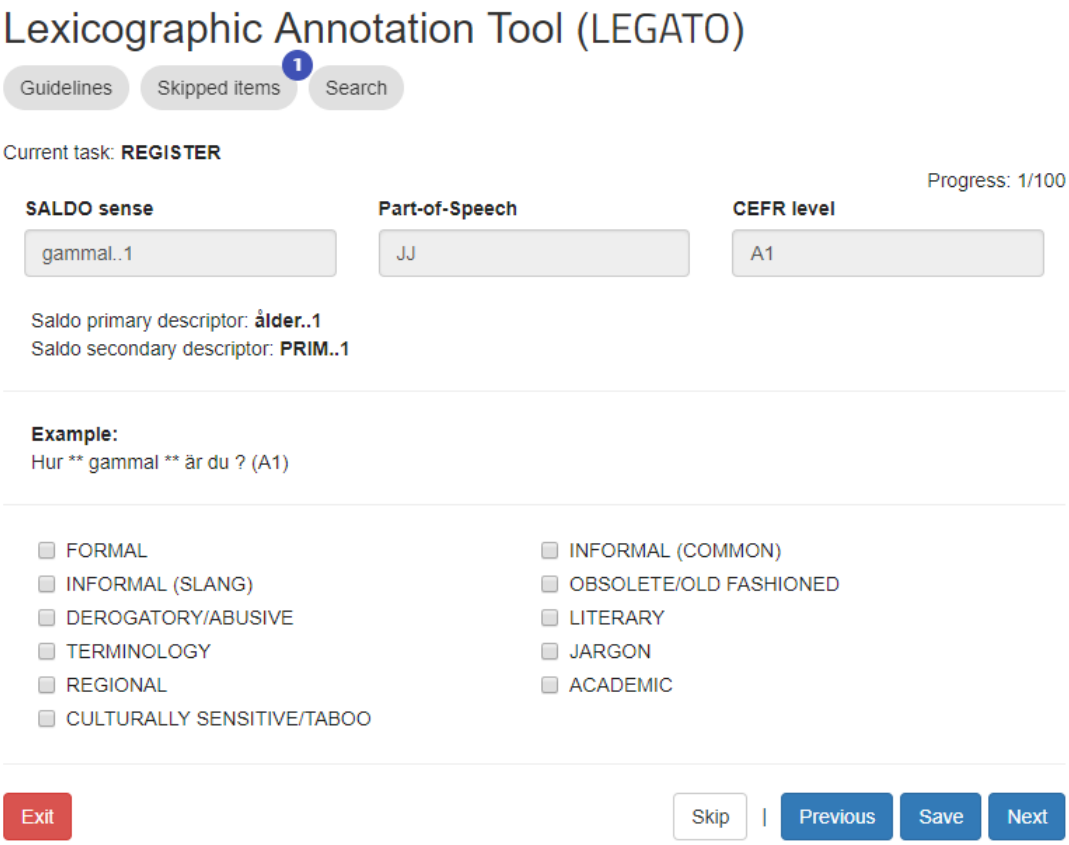


Figure 12: Lexicographic annotation tool Legato

valuable resource for modeling learners (e.g. L1-specific errors, learners' development over time) and to offer adaptive exercises.

References

David Alfter and Elena Volodina. 2018. Towards single word lexical complexity prediction. In *Proceedings of the 13th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 79–88.

David Alfter, Yuri Bizzoni, Anders Agebjörn, Elena Volodina, and Ildikó Pilán. 2016. From distributions to labels: A lexical proficiency analysis using learner corpora. In *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition at SLTC, Umeå, 16th November 2016*, number 130, pages 1–7. Linköping University Electronic Press.

Annika Andersson, Susan Sayehli, and Marianne Gullberg. 2018. Language background affects online word order processing in a second language but not offline. *Bilingualism: Language and Cognition*, pages 1–24.

Lars Borin and Anju Saxena. 2004. Grammar, incorporated. In Peter Juel Henrichsen, editor, *CALL for the Nordic languages*, pages 125–145. Samfundslitteratur, Copenhagen.

Lars Borin, Markus Forsberg, and Johan Roxendal. 2012. Korp-the corpus infrastructure of Spräkbanken. In *PLRroceedings of EC*, pages 47 20124–478.

Lars Borin, Markus Forsberg, and Lennart Lönngren. 2013a. SALDO: a touch of yin to WordNets yang. *Language Resources and Evaluation*, 47(4):1191–1211.

Lars Borin, Markus Forsberg, Leif-Jöran Olsson, Olof Olsson, and Jonatan Uppström. 2013b. The lexical editing system of Karp. In *Proceedings of the eLex 2013 conference*, pages 503–516.

Lars Borin, Nina Tahmasebi, Elena Volodina, Stefan Ekman, Caspar Jordan, Jon Viklund, Beáta Megyesi, Jesper Näsman, Anne Palmér, Mats Wirén, Kristina Nilsson Björkenstam, Gintar Grigonyt, Sofia Gustafson Capková, and Tomasz Kosiski. 2017. Swe-Clarin: Language resources and technology for digital humanities. *Digital Humanities 2016. Extended Papers of the International Symposium on Digital Humanities (DH 2016) Växjö, Sweden, November, 7-8, 2016. Edited by Koraljka Golub, Marcelo Milra*, Vol-2021.

Jill Burstein and John Sabatini. 2016. The language muse activity palette. *Adaptive educational technologies for literacy instruction*, pages 275–280.

Jill Burstein, Nitin Madnani, John Sabatini, Dan McCaffrey, Kietha Biggers, and Kelsey Dreier. 2017. Generating language activities in real-time for English learners using language muse. In *Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale*, pages 213–215. ACM.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Press Syndicate of the University of Cambridge.

Thomas François, Elena Volodina, Ildikó Pilán, and Anaïs Tack. 2016. SVALex: a CEFR-graded lexical resource for Swedish foreign and second language learners. In *LREC*.

Núria Gala, Thomas François, and Cédrick Fairon. 2013. Towards a French lexicon with difficulty measures: NLP helping to bridge the gap between traditional dictionaries and specialized lexicons. *E-lexicography in the 21st century: thinking outside the paper., Tallin, Estonia*.

Núria Gala, Thomas François, Delphine Bernhard, and Cédrick Fairon. 2014. Un modèle pour prédire la complexité lexicale et graduer les mots. In *TALN 2014*, pages 91–102.

Martin Gellerstam. 1999. LEXIN-lexikon för invandrare. *LexicoNordica*, (6).

Elma Kerz and Daniel Wiechmann. 2017. Individual differences in l2 processing of multi-word phrases: Effects of working memory and personality. In *International Conference on Computational and Corpus-Based Phraseology*, pages 306–321. Springer.

Therese Lindström Tiedemann, Elena Volodina, and Håkan Jansson. 2016. Lärka: ett verktyg för träning av språkterminologi och grammatik. *LexicoNordica*, 23:161–181.

Nitin Madnani, Jill Burstein, Norbert Elliot, Beata Beigman Klebanov, Diane Napolitano, Slava Andreyev, and Maxwell Schwartz. 2018. Writing mentor: Self-regulated writing feedback for struggling writers. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 113–117.

Naiara Perez and Montse Cuadros. 2017. Multilingual call framework for automatic language exercise generation from free text. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 49–52.

Ildikó Pilán, Elena Volodina, and Lars Borin. 2017. Candidate sentence selection for language learning exercises: from a comprehensive framework to an empirical evaluation. *TAL*, 57(3/2016):67–91.

Björn Rudzewitz, Ramon Ziai, Kordula De Kuthy, and Detmar Meurers. 2017. Developing a web-based workbook for english supporting the interaction of students and teachers. In *Proceedings of the Joint 6th Workshop on NLP for Computer Assisted Language Learning and 2nd Workshop on NLP for Research on Language Acquisition at NoDaLiDa, Gothenburg, 22nd May 2017*, number 134, pages 36–46, Linköping. Linköping University Electronic Press.

Elena Volodina and Dijana Pijetlovic. 2015. Lark trills for language drills: Text-to-speech technology for language learners. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 107–117.

Elena Volodina, Lars Borin, Hrafn Lofsson, Birna Arnbjörnsdóttir, and Guðmundur Örn Leifsson. 2012. Waste not; want not: Towards a system architecture for ICALL based on NLP component re-use. In *Proceedings of the SLTC 2012 workshop on NLP for CALL; Lund; 25th October; 2012*, number 080, pages 47–58. Linköping University Electronic Press.

Elena Volodina, Ildikó Pilán, Stian Rødven Eide, and Hannes Heidarsson. 2014a. You get what you annotate: a pedagogically annotated corpus of coursebooks for Swedish as a second language. In *Proceedings of the third workshop on NLP for computer-assisted language learning at SLTC 2014, Uppsala University*, number 107. Linköping University Electronic Press.

Elena Volodina, Ildikó Pilán, Lars Borin, and Therese Tiedemann Lindström. 2014b. A flexible language learning platform based on language resources and web services. In *Proceedings of LREC 2014, Reykjavik, Iceland*, pages 3973–3978.

Elena Volodina, Ildikó Pilán, Ingegerd Enström, Lorena Llozhi, Peter Lundkvist, Gunlög Sundberg, and Monica Sandell. 2016a. SweLL on the rise: Swedish learner language corpus for european reference level studies. *arXiv preprint arXiv:1604.06583*.

Elena Volodina, Ildikó Pilán, Lorena Llozhi, Baptiste Degryse, and Thomas François. 2016b. SweLLex: second language learners productive vocabulary. In *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition at SLTC, Umeå, 16th November 2016*, number 130, pages 76–84. Linköping University Electronic Press.

Elena Volodina, Lena Granstedt, Sofia Johansson, Beáta Megyesi, Julia Prentice, Dan Rosén, Carl-Johan Schenström, Gunlög Sundberg, and Mats Wirén. 2018. Annotation of learner corpora: first swell insights. *Proceedings of SLTC 2018*.

Oral History and Linguistic Analysis. A Study in Digital and Contemporary European History

Florentina Armaselu
Luxembourg Centre for Contemporary and Digital History (C^2DH)
University of Luxembourg
florentina.armaselu@uni.lu

Elena Danescu
Luxembourg Centre for Contemporary and Digital History (C^2DH)
University of Luxembourg
elena.danescu@uni.lu

François Klein
Luxembourg Centre for Contemporary and Digital History (C^2DH)
University of Luxembourg
francois.klein@uni.lu

Abstract

The article presents a workflow for combining oral history and language technology, and for evaluating this combination in the context of two use cases in European contemporary history research and teaching. Two experiments have been devised to analyse how interdisciplinary connections between history and linguistics are built and evaluated within a digital framework. The longer-term objective of this type of enquiry is to draw up an "inventory" of strengths and weaknesses and potentially build an online collection of use cases to share reflections and render more transparent the process of applying language technology to research and teaching in different areas of study in the humanities.

1 Introduction

To what extent can the combination of digital linguistic tools and oral history assist research and teaching in contemporary history? How can this combination be evaluated? Is there any added value in using linguistic digital methods and tools in historical research/teaching as compared with traditional means? What are the benefits and limitations of this type of method? The paper will address these questions starting from two experiments based on an oral history collection, XML-TEI annotation and textometric analysis.

In her outline of an oral history "à la française", Descamps (2013: 109-110) talks about a "linguistic age" or a "first age of recorded speech" starting in the 1910s when language scientists began to show an interest in oral sources. With the "invention of oral history, in the 1960s", the use of the spoken word emerged in the historical discipline, subsequently becoming an "indispensable method for contemporary history".[1] Various linguistic aspects have since been considered in the study of spoken corpora. More traditional approaches dealt with this type of data from a number of different perspectives, such as formal and functional narrative analysis of oral versions of personal experiences (Labov and Waletzky, 1967), discursive analysis of the construction of gender identity in life story interviews (Slabakova, 2016), linguistic analysis of metaphor and agency in narrative-biographical interviews (Leonardi, 2018) or close reading by applying discourse analysis and systemic functional linguistics to human rights-related testimonies (Bock, 2007). Digitally oriented research, on the other hand, adopted methods such as topic modelling and sentiment analysis for oral communication data (Choudhury et al., 2018), discourse structure analysis and automatic segmentation of speech corpora transcripts (Zhang and Soergel, 2006), word frequency and co-occurrence computation and qualitative

[1] Fr. "[...] un premier âge de la parole enregistrée, *l'âge linguistique* [...]"; "[...] l'invention de l'histoire orale, dans les années 1960 [...]"; "[...] une « méthode » incontournable de l'histoire contemporaine." (Descamps, 2013: 109).

Florentina Armaselu, Elena Danescu and François Klein 2019. Oral History and Linguistic Analysis. A Study in Digital and Contemporary European History. *Selected papers from the CLARIN Annual Conference 2018*. Linköping Electronic Conference Proceedings 159: 15–27.

analysis for oral history life-course interviews (Hájek and Vann, 2015), and corpus linguistics for dialect speech data or for self-representation in life story interviews (Anderwald and Wagner, 2007; Sealey, 2009).

Since the mid-1970s, the resources and methods of European integration history research have been enhanced with sources from oral history, which are now regularly used alongside both traditional and digital text- and image-based sources (archives, published material, official publications, etc., as well as Web archives and online databases). This "epistemological continuity between written and oral sources" (Bloch, 1999) confirms that oral sources and resources are contributing to the creation and transmission of historical knowledge, while also adding a dimension related to memory and heritage (Ritchie, 2003). Oral history can be seen as a "negotiated history" (Janesick, 2010) or as an "intermediated, influenced history" (Descamps, 2006). In other words, it is "recreated" by the historian in cooperation with the interviewee. It is, therefore, a subdiscipline of the humanities in which critical analysis remains vital. Oral sources are complementary and often prolific, but they should never be viewed in isolation; historians must constantly compare and contextualise them by referring to other sources, especially written sources, which confirm or refute them.

Bridging oral history and linguistics in a digital context has also been the object of dedicated event-oriented initiatives and research, both inside and outside the framework of CLARIN (CLARIN-PLUS OH, 2016; Oral History meets Linguistics, 2015; Georgetown University Round Table on Languages and Linguistics, 2001). Within this context, different tools and perspectives have been adopted, such as language technologies for annotating, exploring and analysing spoken data (Drude, 2016; Van Uytvanck, 2016; Van Hessen, 2016), online platforms for Multimodal Oral Corpus Analysis (Pagenstecher and Pfänder, 2017) or the use of oral histories as "data" for discourse analysts (Schiffrin, 2003).

However, the question of how oral history and linguistics may impact the historian's exploration and interpretation of data seems so far to have been the focus of less research. The theme of digital tool adoption by humanist scholars, and in particular by historians, has already been addressed, either within the scope of tool-building projects and attempts to identify user needs (Gibbs and Owens, 2012; Kemman and Kleppe, 2014) or within the areas of digital tool criticism and digital hermeneutics (Traub and Van Ossenbruggen, 2015; Koolen et al., 2018). Our study is situated in between these approaches: it explores how digital linguistic methods are applied (to answer specific research questions) and perceived by historians (especially as far as added value and innovative potential are concerned). It presents a methodology for preparing and analysing oral history data via tools of corpus linguistics and for observing the "human factor" while dealing with this language technology to accomplish history-related tasks. The proposal aims to contribute to this topic (which in our opinion is of potential interest for the CLARIN community, as it is related to building and evaluating interdisciplinary connections between history, linguistics and digital technologies) and consists of a workflow for: (1) transforming and processing historical spoken data intended for linguistic analysis; (2) evaluating the impact of the use of language technologies in historical research and teaching.

2 Methodology

The growing enthusiasm among the European Union (EU) institutions for oral history on the theme of European integration[2] has led to the systematic use of audiovisual sources for university-based research in this field. Adopting this approach, the Centre virtuel de la connaissance sur l'Europe (CVCE) composed an extensive collection of original historical interviews (more than 160 hours)[3]

[2] Since 1997, the European institutions (Commission, Parliament and Council) have begun gathering a series of oral accounts which have now been compiled into a dedicated collection within the Historical Archives of the European Union. The European Commission was a pioneer in this field, with its "Voices of Europe" programme (1997) (a collection of oral accounts from politicians, diplomats and senior officials who made a significant contribution to the European integration process and its early developments) and "European Commission (1958-1972) – History and memories of an institution" (2002) (a series of oral accounts on the history of the European Commission at the time of the Six, from the creation of the Common Market and Euratom institutions to the eve of the first enlargement). Since 2009, the European Parliament has been building up an oral history collection entitled "Oral history of the European Parliament Presidents" (http://www.europarl.europa.eu/historicalarchives/en/multimedia-gallery/interviews-of-the-presidents.html).

[3] https://www.cvce.eu/histoire-orale. The CVCE is now part of the Luxembourg Centre for Contemporary and Digital History (C²DH) at the University of Luxembourg, https://www.c2dh.uni.lu/.

with key actors and witnesses of the European integration process from Luxembourg and Europe, conducted in French, English, German, Spanish and Portuguese (Klein, 2011-2017).

The present study is based on a selection from this oral history collection, focused particularly on the topic of Economic and Monetary Union (EMU). These interviews represented entirely new sources for the topic under examination and more broadly for the research community as a whole and, given their heritage value, for other sectors of the public. The selection referred to in this paper included 5-10 hours of filmed recordings and transcriptions, in French. The selected transcriptions were converted to a structured format, XML-TEI[4], then imported into the TXM[5] textometry software (Heiden et al., 2010) for linguistic analysis. Two experiments were devised. The first (EUREKA_2017) functioned as a pilot using a smaller corpus and involved a small group of C^{2}DH researchers. The second (MAHEC_2018) was part of a course in Political and Institutional History for Master's students in Contemporary European History at the University of Luxembourg. For each experiment, a set of research questions was prepared, and questionnaires were designed to investigate the role of the language technology in answering these research questions (or in identifying other related research questions).

2.1 Corpus selection and research questions

The "History of European political integration" course, part of the Master's in Contemporary European History, looks at the history of European integration from the early 20th century to the Treaty of Maastricht in 1993 from a political and institutional angle. The learning objectives are not just to provide students with a solid grounding in the political and institutional processes involved in European integration (its origins and development, interconnected structures, mechanisms and players, etc.), including the role played by Luxembourg and its elites, but above all to give them the skills they need to apply critical examination and analysis techniques to the various conceptual and historical perspectives on the building of a united Europe. In terms of methodology, it is hoped that the use of digital primary sources (textual, audio, visual) and methods and tools for digital analysis and visualisation will foster a new historical approach and facilitate access to the complex issues involved in the European integration process.

In light of these goals, we identified the topic of EMU as being of particular interest. EMU not only represents a vital stage in European integration, of which the euro is a tangible result; it is also a valuable object of study in terms of the lessons in economic governance learned following the 2008-2018 economic and financial crisis. Examining the historical processes that gave rise to these events can help shed light on early warning signs pointing to the crisis and avenues for resolution. The corpus that was compiled for the course arose from the "Pierre Werner and Europe" interdisciplinary research project, which was based on a thorough exploration of the Werner family private archives, opened for the first time for research purposes (Danescu, 2013). A series of historical interviews (see Appendix) conducted with key figures from Luxembourg and the international community (more than 55 hours of footage in total) complement the extensive research carried out in these and other archives, offering added value and new resources for the research community.

The corpus developed for the EUREKA and MAHEC experiments is composed of original oral history sources that particularly focus on the plan for the establishment by stages of an economic and monetary union (the Werner Report), the events that subsequently led to EMU, the Luxembourg Compromise, the accession of the United Kingdom, and cooperation between the Benelux countries and the Belgium-Luxembourg Economic Union (BLEU).[6] The number of selected interviewees varied from six (EUREKA) to eight (MAHEC), including figures such as Jean-Claude Juncker, Viviane Reding, Jacques Delors and Étienne Davignon (see Appendix). The selection criteria focused on important milestones in the development of the European Union and the interviews had to be in French for homogeneity purposes. One research question was proposed for the pilot experiment and seven for the second. They were either general queries, e.g. discern the multiple dimensions of the European integration process (EUREKA), or more specialised questions related to the topic of the

[4] http://www.tei-c.org/index.xml.
[5] http://textometrie.ens-lyon.fr/?lang=en.
[6] All these interviews, their transcriptions and translations into English, French and German are published in E. Danescu, The Werner Report of 8 October 1970 in the Light of the Pierre Werner Family Archives (research corpus), Source:. https://www.cvce.eu/project/werner/.

course, e.g. identify the European institutions mentioned in the interviews, their role and interconnections, reconstruct the process of Economic and Monetary Union or determine which of the interviewees is speaking more about Luxembourg's role in European integration, which less, and why (MAHEC).

2.2 Corpus preprocessing

Figure 1 shows the general workflow for preprocessing the corpus before TXM analysis. The filmed recordings were first transcribed[7] into Microsoft Word or Open Office formats. In this project we used the transcriptions as Microsoft Word files that contained markers for identifying the interviewer/respondent and, occasionally, timecodes. As the interviews were structured and included tables of contents and sections, heading styles were added to mark section titles in the documents.

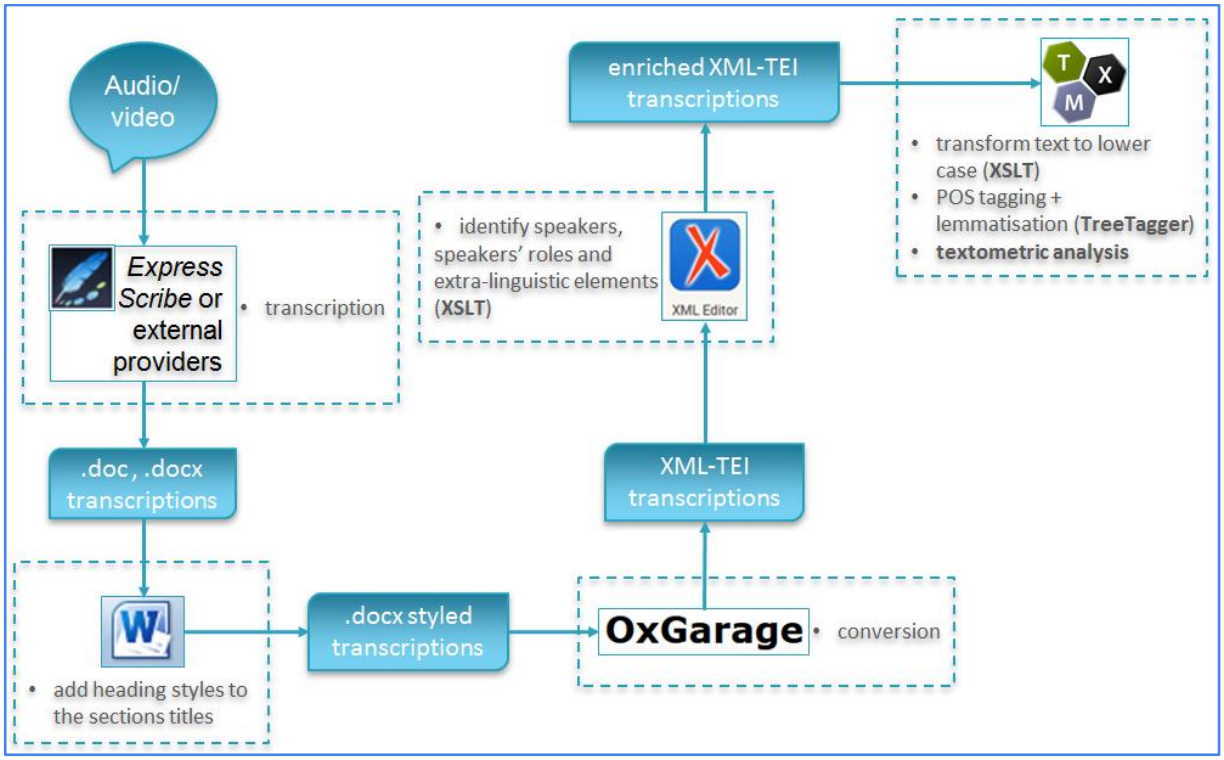


Figure 1. Preprocessing workflow for TXM analysis

The transcriptions were first converted from styled Microsoft Word .docx to a raw XML-TEI[8] version that contained only generic encoding for the metadata in the teiHeader and divisions, paragraphs or highlighting marks for the content (body) area. A series of XSLT[9] stylesheets, specially created for this purpose, were then applied to the converted output,[10] in order to transform it into specific TEI encoding for the transcription of speech. Additional information was inserted into the teiHeader, e.g. speaker roles and speaker list in the participant description to serve as a reference that could be pointed to from the body of the text.

The extract in Figure 2 illustrates how the identity (name) and type of speaker (interviewer/respondent) were encoded in the <particDesc> area from the teiHeader and by using the <u> tag (utterance) and the @who and @corresp attributes in the body of the document. Time points (when present) were encoded by <timeline> and <anchor/> elements, in order to mark the text with respect to time. Extra-linguistic aspects, although rare in the selected data, e.g. <pause>, <kinesic>, marking a pause within the utterance or a gesture, were also considered.

[7] From H264 format, using *Express Scribe* (https://www.nch.com.au/scribe/index.html) or by external providers.
[8] Via the OxGarage online service, http://www.tei-c.org/oxgarage/.
[9] https://www.w3.org/TR/xslt/.
[10] Using oXygen XML Editor, https://www.oxygenxml.com/.

```
<profileDesc>
   <particDesc>
      <p>Speaker roles:
         <list xml:id="speaker_roles">
            <item xml:id="interviewer">Interviewer</item>
            <item xml:id="respondent">Respondent</item>
         </list>
      </p>
      <p>Speaker list:
         <list xml:id="speaker_list">
            <item xml:id="hervé_bribosia">hervé_bribosia</item>
            <item xml:id="wilfried_martens">wilfried_martens</item>
         </list>
      </p>
   </particDesc>
</profileDesc>
```

```
<u who="#hervé_bribosia" corresp="#interviewer"><anchor synch="#t262"/> Et un siège
   unique pour le Parlement européen, on y arrivera un jour ?</u>
<u who="#wilfried_martens" corresp="#respondent"><anchor synch="#t263"/> Ah, c'est le
   Traité. C'est réglé dans le Traité, il faut l'accord de tous. Même le Parlement
   européen ne peut pas l'imposer. C'est un élément du Traité. Et honnêtement, je
```

Figure 2. XML-TEI encoding of speakers and utterances – interview with Wilfried Martens

2.3 TXM analysis

TXM is a piece of textometry software based on a methodology allowing quantitative and qualitative analysis of textual corpora by combining developments in lexicometric and statistical research with corpus technologies (Unicode, XML, TEI, NLP, CQP, R) (TXM Manual; TXM Website).

The corpus in XML-TEI format was imported into TXM, lemmatised[11] and parts of speech were tagged. An XSLT stylesheet was also created and applied during the import to convert the text to lower case. The analysed samples contained a total of 38,687 (EUREKA) and 110,563 (MAHEC) word occurrences. Given the encoding, it was possible to build sub-corpora and partitions corresponding to the name and type of the speaker. Separate sub-corpora were created for interviewer and respondent, respectively, and inside them, partitions for the speakers corresponding to each role, by selecting a structural element (<u>) and an appropriate property (attribute @corresp or @who). Taking into account their potential for contrasting and quantitative/qualitative exploration, the following TXM features were recommended to the participants to be used in their tasks of finding answers to the proposed questions or formulating new research questions: specificities[12] (Lafon, 1980), index, concordances and co-occurrences (TXM Manual).

Figure 3 illustrates specificities, that is a comparative view of the vocabularies of the respondents. The tool allows direct computation of specificities, based on a single property (e.g. *word*, *lemma*, *part of speech*) or more complex processing. For instance, particular queries can be entered via the index using single properties or a combination of properties (e.g. different parts of speech). Lexical tables,[13] specificity scores and diagrams may then be built based on query results. The figure shows the results of computing specificities for the combination *noun* + *adjective*. For the top five European institutions most frequently mentioned in the text, an overuse can be observed for *banque centrale*[14] (first vertical bar for each speaker) in the discourse of Yves Mersch and Jean-Claude Juncker (speakers 8 and 5), and an underuse in the speech of Étienne Davignon (speaker 2), with scores over or under a banality threshold of +/- 2.0 marked by horizontal red lines in the figure.

[11] Via TreeTagger.
[12] "The Specificities command calculates a statistic indicating whether in each part of a partition the occurrences of a word or CQL query appear in abundance (or in decline)." (TXM Manual: 94)
[13] "A Lexical Table assembles together the different lexical units of a partition and displays them in table form." (TXM Manual: 111)
[14] Eng. "Central Bank."

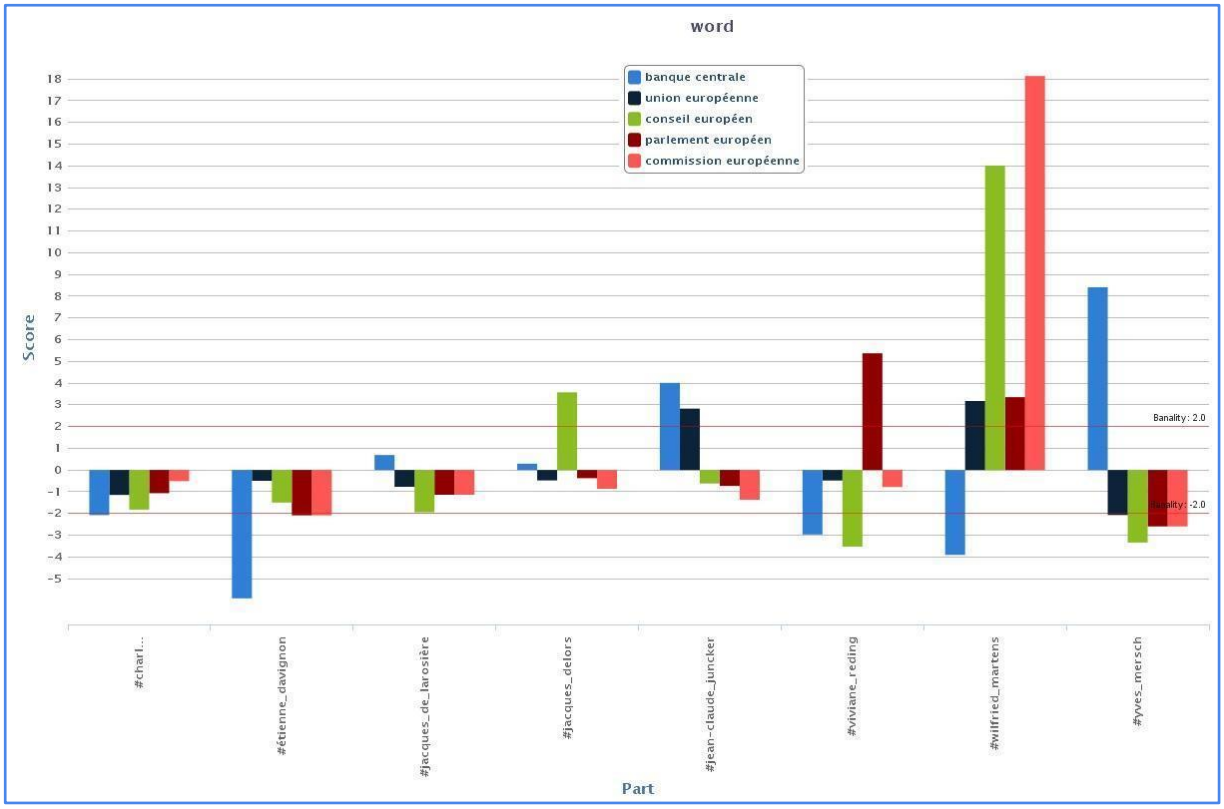


Figure 3. Specificities for European institutions within the respondents' partition (MAHEC_2018)

Other features allowed detection of forms having a tendency to occur together (co-occurrences, e.g. *banque centrale* + *européenne*) or a switch from a synthetic, tabular view to mini-contexts (concordances, e.g. *la banque centrale européenne est en charge de la politique monétaire …*[15]) or document visualisation (Figure 4a, b).

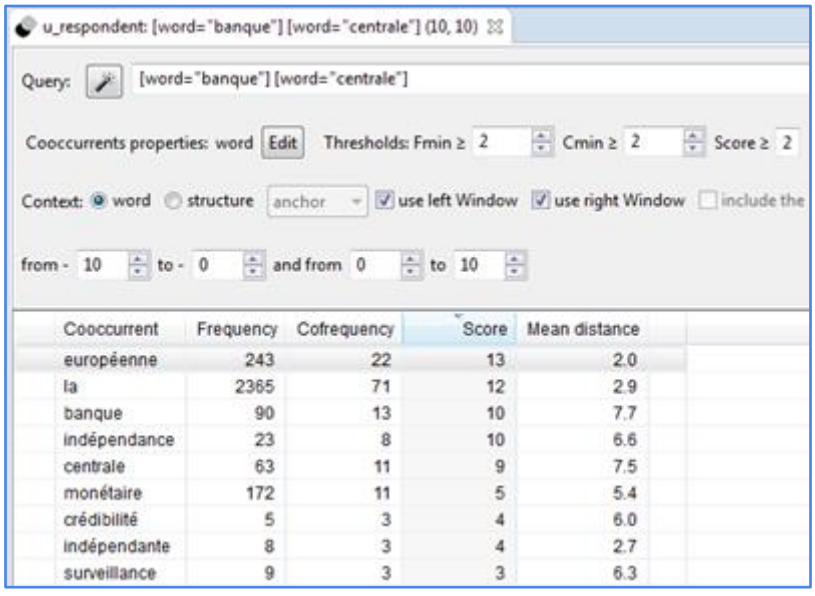


Figure 4a. Co-occurrences (MAHEC_2018)

Our hypothesis was that this type of linguistic analysis, mingling quantitative and qualitative perspectives, may help the participants in their quest for answers to the proposed questions or new questions. For instance, we assumed that different dimensions of European integration (e.g. monetary, economic, political, diplomatic or legal) may be discerned by analysing the specific vocabularies of each of the interviewees as an expression of the particular roles they played in the process (EUREKA). It was supposed that more precise questions may be answered as well. For example, examining the combinations of *pronoun* + *verb* or *noun* + *adjective*, *noun* + *noun*, query by *numerals*, etc. and their specific usage in the respondents' speech may provide insight into nuanced role distinctions such as *actor/witness* in the events

[15] Eng. "the European Central Bank is in charge of monetary policy … ."

discussed, enable identification of important entities (institutions and key figures) and their respective roles or highlight temporal milestones and how concepts have evolved (MAHEC). Given that the degree of familiarity of the participants with the tool was not high and the aim of the experiments was also didactic, suggestions for possible paths of exploration in TXM were made either when assisting with the tasks (EUREKA) or within the assignment instructions themselves (MAHEC). At the same time, the participants were encouraged to look for alternative solutions on their own.

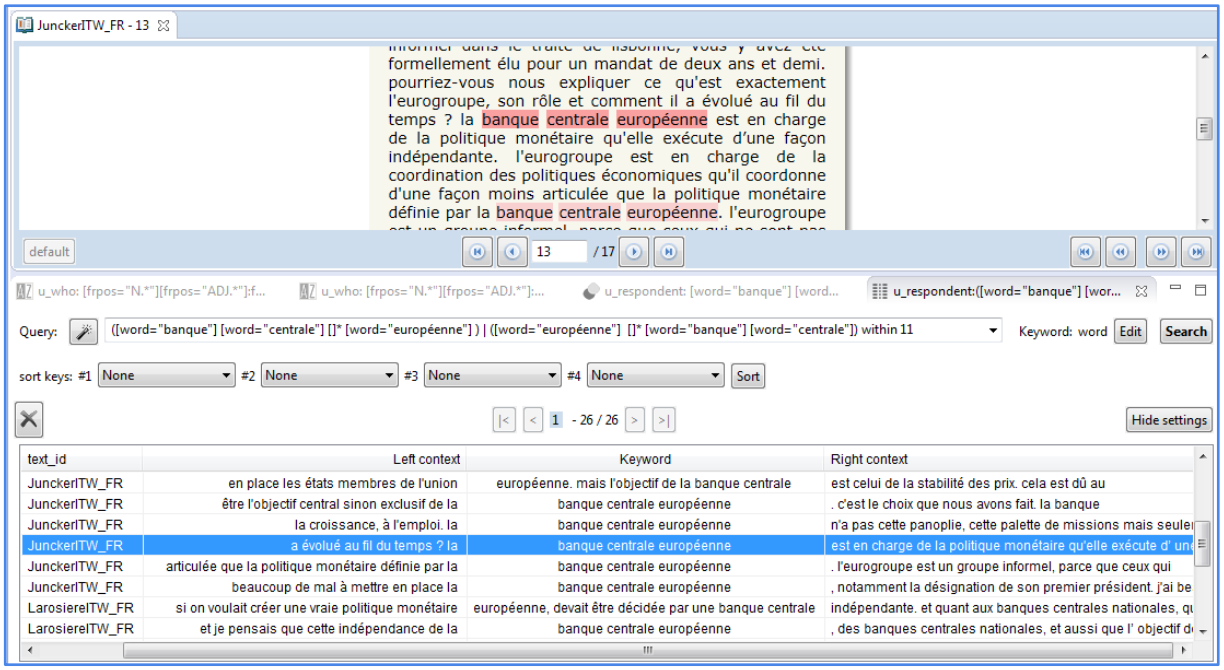


Figure 4b. Concordances and document view (MAHEC_2018)

2.4 Evaluation

The evaluation[16] was intended to confirm/disconfirm the above-mentioned hypothesis and to "measure" the impact of the linguistic technology, its innovative aspects and limitations, when applied to the study of history. Evaluation questionnaires were designed via Google Forms and made available at the end of each test phase or assignment. For anonymisation purposes, identification codes (ID) were assigned to the participants and distributed in sealed envelopes before they answered the questionnaires. The links to the questionnaires were communicated by email (EUREKA) or via the MOODLE page of the course (MAHEC). The language used for the questions/answers was French, as for all the materials (instructions, tutorials, slips indicating ID codes) previously distributed.

Each questionnaire was designed to contain three sections: (1) Participant, including: participant ID code list, age range and gender, main field of expertise, self-evaluation on a scale of 1 to 5 (*Not at all* to *Expert*) in the fields of: *European integration history*, *Multimedia and oral history* (EUREKA) and *Textometry*. Agreement to use the anonymised answers for research/publication was explicitly required by a *Yes/No* question. All the answers from this section were mandatory. (2) Evaluation of the: a) multimedia technology and the oral history collection (EUREKA, first phase); b) textometric analysis (EUREKA, second phase; MAHEC). (3) Evaluation of the proposed experimental scenario.

The overall protocol and questionnaire content were simplified for the second experiment in order to make the students' work more straightforward. However, the general structure, most of the questions and the types of queries were maintained. Sections 2 and 3 above included three types of questions: (1) *Yes/No* questions; (2) Likert-scale queries (with five possible answers from *Don't agree at all* to *Fully agree*, *Very weak* to *Essential* or *Not at all interesting* to *Very interesting*); (3) open questions.[17]

[16] For legibility purposes, English translations of sample questions/answers are provided in footnotes. When the description applies to only one experiment/phase, this is mentioned in brackets; otherwise, the prose/examples apply to both experiments.
[17] Examples: (1) "Did you find answers to the research questions?"; "Would you like to formulate other research questions

3 The experiments

3.1 Description

The pilot experiment EUREKA_2017[18] took place from 11 to 15 and 18 to 22 September 2017 and involved the study of: (1) online filmed interview sequences (5 hours, 6 interviewees) and transcriptions; (2) XML-TEI transcriptions imported into TXM and ready for analysis (sub-corpora for respondents and interviewers and corresponding speaker partitions were provided). The participants were four C[2]DH researchers specialised in *European integration*, *Contemporary history* and *History and political science*. While the profile data showed specialisation in European integration with medium knowledge in multimedia and oral history, the self-evaluation of the textometry skills was placed at the lower end of the scale (Table 1, left).

The second experiment, MAHEC_2018, involved five Master's students in Contemporary European History and took place from 16 April to 14 May 2018. The data sample contained transcriptions of interviews (10 hours, 8 interviewees) in XML-TEI format imported into TXM (with sub-corpora and partitions prepared in advance). Links to access the selected video sequences online were also provided but their analysis was not part of the tasks (however, a student reported having consulted them to learn more about the history of European integration). The students' backgrounds varied from *History* and *Contemporary European history* to *Mediaeval history*, with medium and good knowledge of *European integration history* reported. Compared with the previous experiment, the self-evaluation of the textometric analysis skills covered a larger spectrum (Table 1, right).

EUREKA_2017										
Age range		**Gender**		**Area of expertise**		**Knowledge**				
20 – 34	**1**	F	**3**	*European integration*	**1**	*History of European integration*				
		M	**1**			*None at all*				*Expert*
								1	**3**	
35 – 44	**2**			*Contemporary history*	**2**	*Multimedia + Oral history*				
						None at all				*Expert*
							1	**2**	**1**	
45 - 54	**1**			*History and political science*	**1**	*Textometry*				
						None at all				*Expert*
						3	**1**			

MAHEC_2018										
Age range		**Gender**		**Area of expertise**		**Knowledge**				
18 – 34	**5**	*F*	**1**	*History*	**2**	*History of European integration*				
		M	**4**			*None at all*				*Expert*
								3	**2**	
				Contemporary history	**2**					
				Mediaeval history	**1**	*Textometry*				
						None at all				*Expert*
						1	**1**	**2**	**1**	

Table 1. Profile of the participants in the two experiments

3.2 Discussion of results

Outcomes from the evaluation of the textometric analysis only (EUREKA, second phase; MAHEC) are presented, since they are more closely related to the topic targeted in the article. In general, a slightly higher percentage of positive responses was observed in the first experiment (75%) than in the second (60%) to the *Yes/No* questions asking (1) whether answers and (2) new questions were found or (3) whether there is any added value in applying textometric techniques as compared to direct exploration of the online collection or to a more traditional non-digital approach: (1) 3 positive/4, (2) 2 positive/4, (3) 4 positive/4 (EUREKA, second phase); (1) 4 positive/5, (2) 1 positive/5, (3) 4 positive/5 (MAHEC). This difference might be explained by the fact that the students seemed to be more reticent than the researchers (20/50% positive) in (2) formulating new questions based on the TXM analysis.

For the Likert-type queries, the results of the first experiment (Figure 5, left) indicate moderate value attributed to the (1) role of textometric analysis in finding the answers to questions, (2) the occurrence of a "Eureka" effect as a result of this technology and (3) the evaluation of the proposed scenario. For analytical and comparative purposes, the five values of the scales were transposed to a numerical range (-2 to +2). Average scores were calculated by considering the numeric values and the distribution of responses by number of participants and answer type, e.g. the role of the textometric analysis was scored as -1 by one, 0 by two and +1 by one participant, with an average value of 0. Slightly higher values were observed for the two other questions. In the second experiment (Figure 5,

related to the proposed ones?" (2) "There is a 'Eureka' effect created by the use of this technology in this study." (EUREKA); "How do you view the role played by textometric analysis in finding answers to the questions?"; "How do you view the proposed experimental scenario?" (3) "Please provide a short description of the 'Eureka' effect, or the absence of this effect, observed during the experiment." (EUREKA); "[...] please describe this 'added value' in a few sentences"; "Other reflections on the innovative character of the considered technology and/or its limitations, bias, etc. for the studied case"; "Please list some strong/weak points of this approach" [proposed scenario].

[18] Presented at *Les rendez-vous de l'histoire. Eurêka-inventer, découvrir, innover*, Blois, France, 4-8 October 2017.

right), the average value of the role of textometric analysis in finding the answers was slightly higher (0.4/0) but the experimental scenario got less points (0.4/0.75) than in the first case.

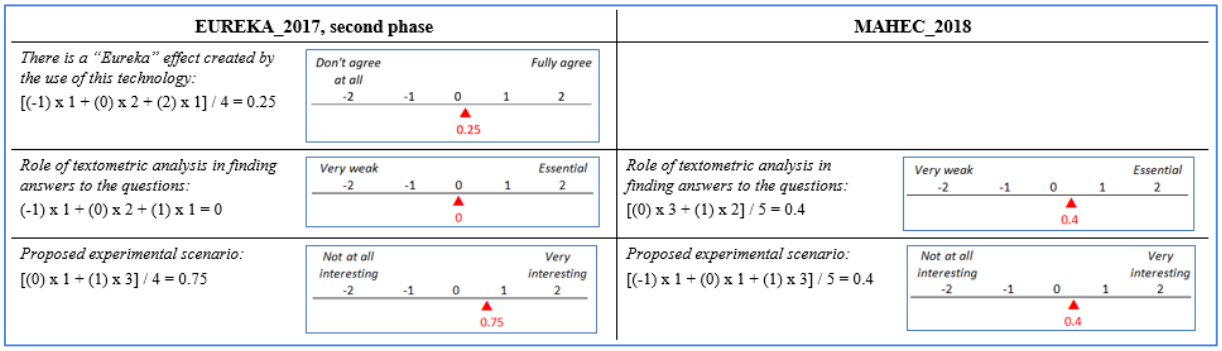


Figure 5. Average Likert-based scores for textometric analysis

More insight into the feedback was provided by the answers to the open questions. In terms of the (1) added value of textometric analysis, the participants in the first experiment mentioned: usefulness for analysing textual corpora by quantitative/statistical techniques allowing observation at both local and more general level, rapid identification of the main themes, and graphical representation of results.[19] The responses to the question (2) asking to describe the "Eureka" effect observed (or not) while using this method of analysis reiterated and enforced some of the above reflections, especially concerning the visual transformation of results and the possibility to highlight and de-contextualise/re-contextualise the linguistic units via a quantitative/qualitative perspective shift. [20] As in a previous quote, considering the second phase (textometric analysis) rather as a "refinement" of the first (online exploration of the videos and transcriptions), another participant noted that no new elements were detected; the only difference was the speed at which different topics could be identified.[21] The nature of the data sample and the usability of the tool were evoked as factors preventing the Eureka effect.[22] It was also observed that textometric analysis alone is not sufficient for research.[23] Other comments provided as (3) additional reflections on the innovative character and limitations of the method reiterated concerns about the impact of the data sample selection/size on the analysis results and highlighted the potential but also drawbacks, difficulties and uncertainties in using the method/interface.[24] Other benefits were mentioned and suggestions for alternatives were provided as (4) comments on the proposed experimental scenario.[25] One of the participants also enquired about the amount and type of preprocessing work necessary for this type of analysis.[26]

[19] Participant's code is provided in square brackets. "Possibility for quantitative and technical statistical analysis to explore a text-based corpus, study of occurrences of a linguistic motif, graphical visualisation of results" [EKA_PIL-P03]; "[...] overview, comprehensive view of the discourse of the interviewees without having to view the videos" [EKA_PIL-P04]; "[...] enables identification of the main themes addressed by the interviewees in a few clicks" [EKA_PIL-P02]; "[...] can be used to refine the results found in the first phase and study the views expressed in the discourse" [EKA_PIL-P01].

[20] "[...] the co-occurrences made it possible to contextualise words or groups of words and to stay close to the text. Textometry therefore combines quantitative and qualitative approaches." [EKA_PIL-P02]; "It highlights 'units', the possibility of visually transforming results through graphs and tables. Extracting elements from their original context but also being able to reintegrate them if needed [...]" [EKA_PIL-P01]

[21] "[...] [it] didn't bring out any new elements as compared with the results of the first phase. However, it enabled the different topics to be identified more quickly [...]" [EKA_PIL-P02]

[22] "The sample studied is not representative enough – it is too consensual for a real Eureka effect. Difficulty in getting to grips with the tool." [EKA-PIL_P03]

[23] "There is a Eureka effect but it should be viewed with caution since using textometric analysis alone is insufficient for research. However, textometric analysis can be a good tool for 'mind mapping'". [EKA-PIL_P04]

[24] "This technology has great potential but more time and a much larger sample are needed in order to fully exploit the potential of the tool." [EKA-PIL_P03]; "[...] The scores are not always effective for analysis and the words are not always representative of the discourse [...] The selection of interviews and excerpts is subjective, which may produce bias in the critical analysis of the research question." [EKA-PIL_P04]; "[…] without prior knowledge in linguistics and discourse analysis, I don't see how I can interpret the 'underuse' of a term […]" [EKA-PIL_P01]; "The interface could be more intuitive and the visualisations and graphics more appealing." [EKA-PIL_P02]

[25] "Textometric analysis can certainly be very useful in examining a large research corpus [...]" [EKA-PIL_P02]; "[...] another possible scenario. Define 2 groups. Group 1 works on the analysis of the interviews using traditional methods [...].

In the second experiment, only four of the five students answered the open questions. The (1) added value elements as compared to more "traditional" analysis methods were similar to those mentioned in the first experiment, e.g. enabling analysis of a large corpus of texts, "fast reading", speed and rigour.[27] As (2) additional reflections on the innovative characteristics and limitations of the studied technology, respondents pointed out the possibility to compare different interviews and the lack of features allowing annotation or modification of the texts.[28] Unlike the first experiment, the facility to pass from quantitative to qualitative view didn't seem to be fully grasped, or perhaps what was meant is that the quantitative aspect is more "tempting", which can lead to overlooking the qualitative facet needed in an enquiry of this nature.[29] As with the first experiment, it was observed that the analysis often served to prove something already known, rather than providing new information.[30] As (3) strong points of the experimental scenario, respondents noted the queries based on combined properties and the suitability of textometric analysis for assisting interpretation.[31] (4) Weak points mentioned were the size of the text/results window and the heterogeneity of the questions asked to interviewees.[32]

4 Conclusion and future work

Given time and resource constraints, the experiments had certain limitations. The number of participants was small and their background and familiarity with the proposed topic were not very diverse, since, as specialists or students in the field, the subject of European integration was relatively well known to them. The data samples, although selected to cover a given theme and percentage from the total collection of interviews, were not very large and did not involve a high number/variety of interviewees. The time allocated to TXM training prior to the experiments was limited (no training but a tutorial and assistance for EUREKA, 90 minutes of training and a tutorial and assistance for MAHEC). Taking into account these limitations, it can be hard to draw out generalisations, though various observations can be made.

Although the speed of processing and visualising linguistic features in large numbers of texts was mainly seen as a plus point, and attributes such as "innovative", "audacious" and "avant-garde" were used in the comments, the results showed a certain degree of reservation as to the innovative added value of the analysis tool. This was expressed both by a lower percentage of proposed new research questions and by explicit statements casting doubt on the new information gained as a result of the method. While this type of response can be partly explained by the above-mentioned limitations – which were also referred to by the participants through concerns raised about the data sample and the need for better knowledge of the tool – it can also indicate, to a certain degree, a specific approach to digital tools. That is, they are seen more as a means for proving hypotheses or known information than as "serendipitous" instruments for envisaging new paths of enquiry. However, this is an aspect that needs to be further examined in future experiments.

On the other hand, the results demonstrated awareness, from both the researchers and the students, of the different aspects involved in applying language technology to answering/identifying questions, such as data, methods, interface and general context of use. These aspects were repeatedly evoked in

Group 2 works on the interviews using the textometric tool [...] Comparison of the results [...]" [EKA-PIL_P03]

[26] "What about the manual effort needed to prepare a large corpus for textometric analysis?" [EKA-PIL_P02]

[27] "Textometric analysis enables the study of a large corpus of texts and saves a lot of time for historians. The analysis of the vocabulary used is greatly facilitated in particular." [TXM-HO_P01]; "Possibility of analysing several documents instead of reading them one by one." [TXM-HO_P02]; "Speed, rigorous analysis." [TXM-HO_P06]; "More efficient for 'fast reading' […]" [TXM-HO_P10].

[28] "[...] it is possible to compare the results for the documents, but this requires the interviews to be transcribed so that they can be read using the tool." [TXM-HO_P02]; "What is missing is a function to mark or modify the text [...]" [TXM-HO_P10]

[29] "Another problem is distancing from quality; with the tool it is very appealing to take a large number of documents for analysis [...]" [TXM-HO_P02]

[30] "An issue in textometric analysis is whether there is a real gain of new information. In most cases, textometric analysis proved the position and the known role of a person, but did not really contribute any new information." [TXM-HO_P01]

[31] "I liked the functionalities grouping certain queries, e.g. personal pronouns, nouns, adjectives, verbs, etc. [...]" [TXM-HO_P02]; "The approach is audacious and avant-garde in the field of history. It makes us reflect on different ways of reading sources, as well as on the logic that connects words in a text. [...]" [TXM-HO_P10]

[32] "The window displaying the text and analysis results should be larger [...]." [TXM-HO_P10]; "I would have liked interviews on a specific theme for all the respondents, [...] to compare the answers and see the result [...]." [TXM-HO_P02].

answers referring to the selection of interviews and the questions proposed to the interviewees and in observations about the qualitative/quantitative enquiry allowed by the tool, the more or less useful or easy-to-understand features provided by the interface and the experimental scenario itself. We would argue that this type of reflection is important not only for the "development of tools to be compatible with specific research methods of scholars" (Kemman and Kleppe, 2014) or for building "reflective tools and methods" (Koolen et al. 2018), but also to shed light on the process of research and teaching via digital tools. In this regard, we agree with Traub and Van Ossenbruggen's (2015) suggestion "to collect use cases and to compare evaluations of different tools" but with the intent of going beyond the creation of "checklists and guidelines for both, tool builders and users". Collecting use cases in this way also represents a means of sharing experiences and rendering the research process more transparent, thus improving understanding of emerging shifts in humanities practices brought about by digital technologies.

To sum up, the project combined original sources of oral history and digital linguistic analysis, and evaluated the use of language technology via two use cases of history research and teaching. Two experiments were devised. Although the data samples and the groups of participants were small and not very diverse, and additional experiments are needed for generalisations to be made, the results provided an insight into how researchers and students apply this type of tool and reflect on its use. We would argue that the creation (potentially within the framework of CLARIN) of an interdisciplinary collaborative platform containing an online collection of use cases, evaluation data and workflow descriptions from different areas of study will encourage the pooling of experience and practices with a view to stimulating debate, creativity and the exchange of ideas within the academic community. By sharing reflections and drawing up an "inventory" of strengths and weaknesses, it will thereby facilitate understanding of current and emerging practices in applying language technology to research and teaching in the humanities.

Acknowledgements

We would like to thank our colleagues and students for participating in the experiments, and Sarah Cooper, from the Faculty of Language and Literature, Humanities, Arts and Education at the University of Luxembourg, for English proofreading.

References

[Anderwald and Wagner 2007] Lieselotte Anderwald and Susanne Wagner. 2007. "FRED — The Freiburg English Dialect Corpus: Applying Corpus-Linguistic Research Tools to the Analysis of Dialect Data". In *Creating and Digitizing Language Corpora*. 10.1057/9780230223936_3.

[Bloch, 1999]. Marc Bloch 1999. "Réflexions d'un historien sur les fausses nouvelles de la guerre". Allia, Paris (Re-publication of the article of the same name, originally published in 1921 in the *Revue de synthèse historique*, T.33).

[Bock 2007] Zannie Bock. 2007. *A Discourse Analysis of Selected Truth and Reconciliation Commission Testimonies: Appraisal and Genre*. PhD Thesis, University of the Western Cape, Republic of South Africa, November 2007.

[Choudhury et al. 2018] Prithwiraj (Raj) Choudhury, Natalie A. Carlson, Dan Wang and Tarun Khanna. 2018. "Machine Learning Approaches to Facial and Text Analysis: An Application to CEO Oral Communication". Working Paper 18 – 064, Harvard Business School.

[CLARIN 2016] CLARIN. 2016. CLARIN-PLUS OH workshop: *Exploring Spoken Word Data in Oral History Archives*. University of Oxford, United Kingdom. https://www.clarin.eu/event/2016/clarin-plus-workshop-exploring-spoken-word-data-oral-history-archives. (Accessed March 31, 2019)

[Danescu 2013] Elena Danescu 2013. *A rereading of the Werner Report of 8 October 1970 in the light of the Pierre Werner family archives*. Research corpus. Source: https://www.cvce.eu/en/project/werner/introduction. (Accessed March 31, 2019)

[Dedman 2009] Martin Dedman 2009. *The Origins & Development of the European Union: 1945–2008*. Routledge, London (Second edition).

[Descamps 2013] Florence Descamps. 2013. "Histoire orale et perspectives. Les évolutions de la pratique de l'histoire orale en France". In F. d'Almeida and D. Maréchal (Ed.), *L'histoire orale en questions*, p. 105–138. INA, Paris.

[Drude 2016] Sebastian Drude. 2016. "ELAN as a tool for oral history". CLARIN-PLUS OH workshop.

[EU Consilium 2018]. Council of the European Union 2018. *Blue guide to the Archives of Member States' Foreign Ministries and European Union institutions*. https://www.consilium.europa.eu/media/29595/blueguide_pdf_201404.pdf. (Accessed March 31, 2019)

[Georgetown University 2001] Georgetown University. 2001. *Georgetown University Round Table on Languages and Linguistics (GURT)*, Washington, DC, USA.

[Freiburg Institute for Advanced Studies 2015] Freiburg Institute for Advanced Studies. 2015. Conference *Oral History meets Linguistics*, Freiburg, Germany. https://www.frias.uni-freiburg.de/en/events/frias-conferences/conference-oral-history-and-linguistics. (Accessed March 31, 2019)

[Gibbs and Owens 2012] Fred Gibbs and Trevor Owens. 2012. "Building Better Digital Humanities Tools: Toward broader audiences and user-centered designs". In *Digital Humanities Quarterly*, 2012, Volume 6, Number 2. http://www.digitalhumanities.org/dhq/vol/6/2/000136/000136.html. (Accessed March 31, 2019)

[HAEU 2018]. *Historical Archives of the European Union 2018*. http://www.eui.eu/Research/HistoricalArchivesOfEU/Index.aspx. (Accessed March 31, 2019)

[Hájek and Vann 2015] Martin Hájek and Barbara H. Vann. 2015. "Gendered Biographies: The Czech State-socialist Gender Order in Oral History Interviews". *Sociologický ústav AV ČR*, v.v.i., Praha.

[Heiden et al. 2010] Serge Heiden, Jean-Philippe Magué and Bénédicte Pincemin. 2010. "TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement". In Sergio Bolasco, Isabella Chiari, Luca Giuliano (Ed.), *Proc. of 10th International Conference on the Statistical Analysis of Textual Data* - JADT 2010, Vol. 2, p. 1021–1032. Edizioni Universitarie di Lettere Economia Diritto, Roma, Italy. https://halshs.archives-ouvertes.fr/halshs-00549779/fr/. (Accessed March 31, 2019)

[Hix and Høyland 2011] Simon Hix and Bjørn Høyland 2011. *The Political System of the European Union*, Basingstoke, Palgrave MacMillan (Third edition).

[Janesick 2010] Valerie J. Janesick. 2010. *Oral history for the qualitative researcher. Choreographing the story*. Guilford Press, New York.

[Kemman and Kleppe 2014] Max Kemman and Martijn Kleppe. 2014. "User Required? On the Value of User Research in the Digital Humanities". In CLARIN 2014 Selected Papers; Linköping Electronic *Conference Proceedings # 116*. http://www.ep.liu.se/ecp/116/006/ecp15116006.pdf. (Accessed March 31, 2019)

[Klein 2010-2017]. François Klein 2010–2017. *Oral history of European integration collection*: Source: https://www.cvce.eu/en/oral-history/presentation. (Accessed March 31, 2019)

[Koolen et al. 2018] Marijn Koolen, Jasmijn van Gorp, Jacco van Ossenbruggen. 2018. "A Hands-on Approach to Digital Tool Criticism. Tools for (self-)Reflection". Conference on *Digital Hermeneutics in History: Theory and Practice*, University of Luxembourg, 25 October 2018. https://docs.google.com/presentation/d/1om6BK4xNNJ0-_hKKwOYArdHrHAn-VRPp0mH6dRo8ViI/edit#slide=id.p. (Accessed March 31, 2019)

[Labov and Waletzky 1967] William Labov and Joshua Waletzky. 1967. "Narrative analysis: Oral versions of personal experience". In J. Helm (Ed.), *Essays on the verbal and visual arts*. Seattle, WA: University of Washington Press. pp. 12–44.

[Lafon 1980] Pierre Lafon. 1980. "Sur la variabilité de la fréquence des formes dans un corpus". *Mots*, no. 1, p 127–165. http://www.persee.fr/doc/mots_0243-6450_1980_num_1_1_1008. (Accessed March 31, 2019)

[Leonardi 2018] Simona Leonardi. 2018. "Metaphors in the Life Story of A German- Jewish Immigrant to Palestine/Israel. How Metaphorical Constructions and Remembering Process Interweave". *Remembrance and Research*, ILOHA, no. 2, January 2018, pp. 51–68.

[Pagenstecher and Pfänder 2017] Cord Pagenstecher and Stefan Pfänder. 2017. "Hidden Dialogues: Towards an Interactional Understanding of Oral History in Interviews". In *Oral History Meets Linguistics*, edited by Erich Kasten, Katja Roller, and Joshua Wilbur, pp. 185–207. Fürstenberg/Havel: Kulturstiftung Sibirien, Electronic Edition. http://www.siberian-studies.org/publications/PDF/orhili_pagenstecher_pfaender.pdf. (Accessed March 31, 2019)

[Passerini 2006] Luisa Passerini 2006. *Memory and Utopia: The Primacy of Inter-Subjectivity*. Equinox Publishing: London.

[Portelli 2009] Alessandro Portelli 2009 "What Makes Oral History Different". In Giudice L.D. (Ed.), *Oral History, Oral Culture, and Italian Americans. Italian and Italian American Studies*. Palgrave Macmillan, New York, pp.21–30.

[Radelli and Featherstone 2003] Claudio Radelli and Kein Fetherstone 2003 (Ed.) *The Politics of Europeanization*. Oxford University Press, Oxford.

[Ritchie 2003]. Donald A. Ritchie 2003. *Doing Oral History*. Oxford University Press, New York.

[Schiffrin 2003] Deborah Schiffrin. 2003. "Linguistics and History: Oral History as Discourse". Georgetown University Round Table on Languages and Linguistics (GURT) 2001: *Linguistics, Language, and the Real World: Discourse and Beyond*, Deborah Tannen and James Alatis (Ed.), pp. 84–113, Georgetown University Press, Washington, D.C. http://faculty.georgetown.edu/schiffrd/index_files/Linguistics_and_oral_history.pdf. (Accessed March 31, 2019)

[Sealey 2009] Alison Sealey. 2009. "Probabilities and surprises: A realist approach to identifying linguistic and social patterns, with reference to an oral history corpus". In *Applied Linguistics*: 1–21, Oxford University Press, doi:10.1093/applin/amp023.

[Slabakova 2016] Radka Slabakova. 2016. "The Meaning of His Life Was Work: The Construction of Identities in the Oral Narratives of Older Czech Men". *Gender Studies*. 15. 10.1515/genst-2017-0008.

[Traub and Van Ossenbruggen 2015] Myriam C. Traub and Jacco van Ossenbruggen. 2015. "Workshop on Tool Criticism in the Digital Humanities". CWI *Techreport*, 1 July 2015, https://pdfs.semanticscholar.org/d337/ce558c2fd1d8be793786c9cfc3fab6512dea.pdf. (Accessed March 31, 2019)

[TXM Manual]. 2018. *TXM User Manual*, Version 0.7 ALPHA, February 2018. http://textometrie.ens-lyon.fr/files/documentation/TXM%20Manual%200.7.pdf. (Accessed March 31, 2019)

[Van Hessen 2016] Arjan van Hessen. 2016. "Increasing the Impact of Oral History Data with Human Language Technologies, How CLARIN is already helping researchers". CLARIN-PLUS OH workshop.

[Van Uytvanck 2016] Dieter van Uytvanck. 2016. "CLARIN Data, Services and Tools: What language technologies are available that might help process, analyse and explore oral history collections?" CLARIN-PLUS OH workshop.

[Zhang and Soergel 2006] Pengyi Zhang and Dagobert Soergel. 2006. "Knowledge-Based Approaches to the Segmentation of Oral History Interviews". MALACH Technical Report, University of Maryland. College of Information Studies, May 2006.

Appendix. List of interviewees in the "Pierre Werner and Europe" project

The figures that have been interviewed thus far in connection with the "Pierre Werner and Europe" research project are as follows (in alphabetical order): Michel Camdessus, Honorary Governor of the Banque de France, Managing Director of the IMF (1987–2000); Luc Frieden, Luxembourg Finance Minister (2009–2013); Albert Hansen, Secretary-General of the Luxembourg Government (1979–1998); Edmond Israel (1924–2011), Luxembourg banker, President of the Board of Directors of Cedel International (1970–1999); Jean-Claude Juncker, President of the European Commission since 2014, Prime Minister of Luxembourg (1995–2013), first permanent President of the Eurogroup (2005–2013); Helmut Kohl (1930–2017), Chancellor of the FRG (1982–1998); Philippe Maystadt (1948–2017), Belgian Finance Minister (1988–1998), President of the European Investment Bank (2000–2011); Yves Mersch, Member of the Executive Board of the European Central Bank (since 2012), President of the Banque centrale du Luxembourg (1998–2012); Guy de Muyser, Marshal of the Grand Ducal Court (1971–1981); Charles-Ferdinand Nothomb, President of the Belgian Chamber of Representatives (1979–1980, 1988–1995), Honorary President of the Pierre Werner European Circle; Viviane Reding, Member of the European Commission (1999–2010), Vice-President of the European Commission with responsibility for Justice, Fundamental Rights and Citizenship (2010–2014), Member of the European Parliament (since 2014); Lex Roth, Director of the Information and Press Service of the Luxembourg Government (1988–1993); Charles Ruppert, Chairman of the Luxembourg Bankers' Association (1992–1995), Chairman of the Pierre Werner Foundation; Fabrizio Saccomanni, Vice President of the European Bank for Reconstruction and Development (EBRD) (2003–2006), Italian Minister for Economic Affairs and Finance (2013–2014); Jacques Santer, Prime Minister of Luxembourg (1984–1995), President of the European Commission (1995–1999); Bernard Snoy et d'Oppuers, International President of the European League for Economic Cooperation, President of Robert Triffin International; Gaston Thorn (1928–2007), Prime Minister of Luxembourg (1974–1979), President of the European Commission (1981–1985); Hans Tietmeyer (1930–2017), President of the Deutsche Bundesbank (1993–1999), Member of the Werner Committee (1970); Niels Thygesen, Member of the Delors Committee (1988–1989), Chairman of the European Fiscal Board (since 2016), Professor at the Institute for New Economic Thinking; Sir Brian Unwin, President of the European Investment Bank (1993–1999), Governor of the European Bank for Reconstruction and Development (1993–1999), Chairman of the Supervisory Board of the European Investment Fund (1994–1999); Henri Werner, son of Pierre Werner; Marie-Anne Werner, daughter of Pierre Werner. Other accounts emerged as a result of the project "Accounts by Luxembourg Ambassadors" (Jean-Jacques Kasel, Adrien Meisch and Jean Mischo), and interviews were also conducted with Étienne Davignon, Member of the European Commission (1977–1981) and Vice-President of the European Commission (1981–1985); Jacques de Larosière, Assistant Director (1967–1974) then Director of the French Treasury (1974–1978), Managing Director of the IMF (1978–1987); Jacques Delors, President of the European Commission (1985–1995); Mark Eyskens, Belgian Finance Minister (1980–1981) and Prime Minister (1981); and Wilfried Martens, Prime Minister of Belgium (1979–1981/1981–1992).

Towards a protocol for the curation and dissemination of vulnerable people archives

Silvia Calamai
University of Siena
Italy
silvia.calamai@unisi.it

Chiara Kolletzek
Lawyer and Record Manager,
Bologna, Italy
chiara.kolletzek@live.it

Aleksei Kelli
University of Tartu
Estonia
aleksei.kelli@ut.ee

Abstract

This paper aims at introducing possible guidelines in defining a protocol for the curation and dissemination of speech archives, which appear to have – *de jure* – the highest restrictions on their curation and dissemination. This case study has been undertaken because of the discovery of the Anna Maria Bruzzone archive, containing the voices of people with mental disabilities recorded in 1977 in an Italian psychiatric hospital.

1 Introduction

This paper presents a coherent reflection on the possibility of defining a protocol for the curation and dissemination of speech archives which appear to have – *de jure* – the highest restrictions on their curation and dissemination since they contain the voices of insane people. This case study has been undertaken because of the discovery of the Anna Maria Bruzzone archive[1]. Bruzzone's interviews were recorded long before the Italian Data Protection Code (IDPC) was issued (2004), so that the informants were not explicitly asked to give their authorization for the use and dissemination of the recordings, although during the interviews the recording device was always kept visible.
The archives are covered with several rights (for further discussion, see Kelli et al. 2015). Firstly, speech itself could be protected as copyrighted work. Secondly, individuals who speak could have performer's rights. Thirdly, the person who created the archive has database rights. Lastly, interviewees' personal data have to be protected from unauthorized use and dissemination. Due to the focus of this article, the analysis is limited to personal data protection.
In the paper, some legal issues affecting the use and re-use of the archive are presented and discussed. The model envisaged aims at finding a balance between the rights of the recorded people (and their heirs) such as privacy, and the right of information and the protection of memory. The focus is on the General Data Protection Regulation (GDPR) which is applicable in all EU member states from 25 May 2018. National laws may specify its application, especially the provisions concerning specific areas of personal data processing (e.g., research: see Kelli et al. 2018).

[1] In the near future, the Archive will be part of the CLARIN Infrastructure and metadata description will be pursued according to COALA. COALA generates corpus and session CMDIs according to the media-corpus-profile and the media-session-profile for the Component Registry, by converting five CVS tables to the CMDI format. A mobility grant under the H2020 project CLARIN-PLUS allowed the first author to prepare a feasibility study on this topic (Bayerisches Archiv für Sprachsignale c/o Institut für Phonetik, Universität München; 4-7 December 2017).
The paper has also benefited from the CLARIN workshop „Hacking the GDPR to Conduct Research with Language Resources in Digital Humanities and Social Sciences" (Vilnius, 7 December 2018), which aimed at bringing together legal experts and researchers from the Digital Humanities and Social Sciences disciplines working with Language Resources in order to exchange views and explore ways of creating and using LRs under the GDPR regime.

Silvia Calamai, Chiara Kolletzek and Aleksei Kelli 2019. Towards a protocol for the curation and dissemination of vulnerable people archives. *Selected papers from the CLARIN Annual Conference 2018*. Linköping Electronic Conference Proceedings 159: 28–38.

The paper is organised as follows: in § 2 Bruzzone's speech archive is described, in § 3 the topic of personal data and special categories of personal data is addressed, while in § 4 and in the Conclusion the possibilities of finding a balance between research, dissemination and protection of privacy are discussed. Finally, the English translation of the informed content is provided in the Appendix.

2 The speech archive of Anna Maria Bruzzone

Anna Maria Bruzzone's book *Ci chiamavano matti. Voci da un ospedale psichiatrico* (Einaudi, Torino 1979) contains the testimonies of thirty-seven patients in the Arezzo psychiatric hospital collected in 1977 (see Fig. 1).

Fig. 1 The book and the original recording device used for the fieldwork.

The book – out of print – testifies the patients' miserable lives inside and outside the hospital and sheds light on the atrocity of their everyday condition by letting them speak for themselves. The author wrote it after a two-month stay in Arezzo, when she spent almost every day in the hospital, attending the general meetings and participating in the lives of the inpatients, in a continuous dialogue of which only a part is collected in the published interviews. The oral recordings on which the book is based were believed to be lost forever. After a long and strenuous search we have been able to locate the original tapes (see Fig. 2), which were donated to the Department of Educational Sciences, Human Sciences and Intercultural Communication of the University of Siena (UNISI) – Arezzo.

Fig. 2 The original tapes of Anna Maria Bruzzone's archive.

This discovery is extremely important because the digitisation and cataloguing of this archive would produce the first digital oral archive related to an Italian psychiatric hospital – which was located in the same buildings as the UNISI Department, where the Historical Archive of the Arezzo psychiatric hospital is also housed.

Reading a testimony and listening to it from the voice of the interviewee are not the same thing and Bruzzone herself was well aware of this (Bruzzone 1979: 22). Furthermore, the published texts are not the exact transcriptions of the original testimonies. In fact, after producing the first, complete transcriptions, Bruzzone had to edit them to make them suitable for publishing. In addition to editing out the speeches so that the interviewees' voices could flow without interruptions, she had to make other cuts and adjustments in order to make the text clearer or more readable, and she even had to give up on publishing some of the testimonies because otherwise, the book would have been too long (see Fig. 3). As she admits, this task was a hard, painful one to her (Bruzzone 1979: 25). Therefore, having the original tapes at our disposal is of fundamental importance, as it allows us to re-connect the published testimonies to the original ones.

The archive consists of 36 tapes accompanied by the handwritten and the typewritten transcriptions of all the interviews. In addition to the complete transcriptions, different versions show all the work of editing made by A.M. Bruzzone so that the interviews could be prepared for publishing (see Fig. 3-4).

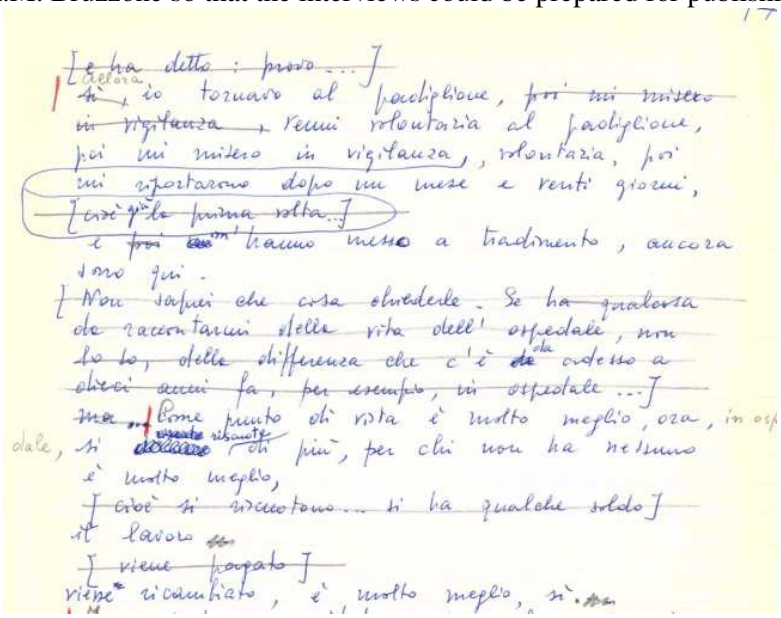


Fig. 3 The handwritten transcription.

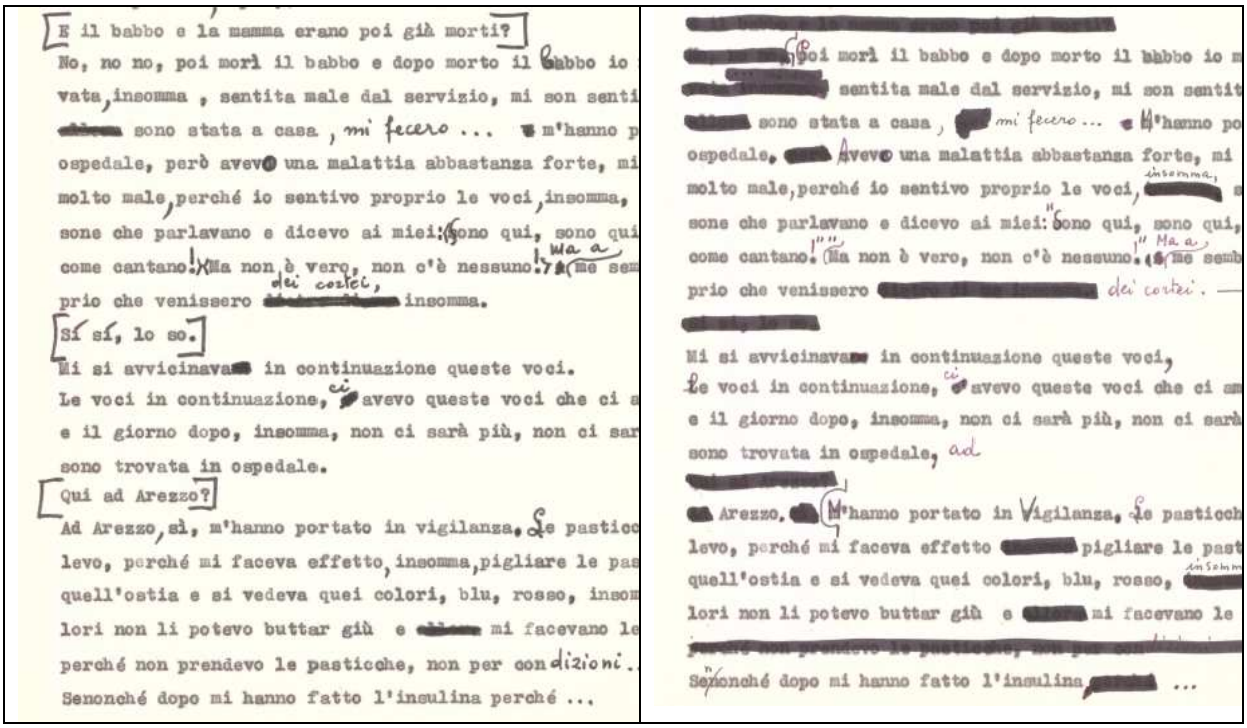


Fig. 4 The typewritten transcription (two versions of the same passage in the interview).

This opens up the possibility to understand, document and examine the changes undergone in an interview from the moment it was recorded on tape to its publication in the book, through the comparative study of all the available documents: the original audio recording, the first, handwritten transcription, the typewritten transcription, the edited version and, finally, the one published in the book. Moreover, it is now possible to associate the oral life stories with the medical diagnosis of every single inpatient (preserved in the Historical Archive of the Arezzo psychiatric hospital), since the real names and not the pseudonyms were been found in the box of every single tape.

3 Personal data and special categories of personal data

The curation and dissemination of archives of vulnerable people are subject to the regulation of personal data. The GDPR defines personal data as "any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person". Article 29 Working Party[2] (WP29) explains that "it is not necessary for the information to be considered as personal data that it is contained in a structured database or file. Also information contained in free text in an electronic document may qualify as personal data" (2007: 8).
The critical issue here is how to interpret the concept of 'identifiable'. The absolute and relative approaches described in the literature are displayed in Fig. 5 (from Spindler, Schmechel 2016).

[2] According to the Data Protection Directive, the Working Party on the Protection of Individuals with regard to the Processing of Personal Data (WP29) is composed of a representative of the supervisory authority or authorities designated by each Member State and of a representative of the authority or authorities established for the Community institutions and bodies, and of a representative of the Commission. The GDPR replaces the Data Protection Directive.

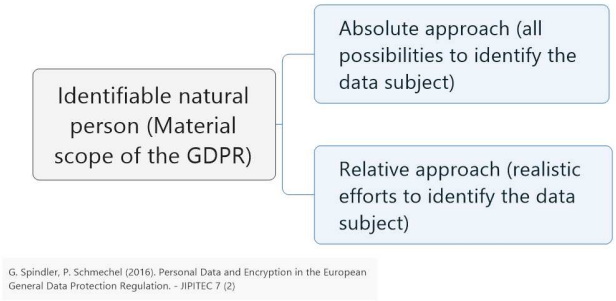


Fig. 5 Absolute and relative approach in the identification of the data subject.

Some authors have emphasized the context-dependency of identifiability (Oostveen 2016: 306). In the analysed case, the individuals are identifiable and no further analysis is required.
The situation concerning speech archives becomes even more complicated for several reasons. Firstly, the human voice is considered biometric data (see González-Rodríguez et al. 2008; Jain et al. 2004). Biometric data is defined as "personal data resulting from specific technical processing relating to the physical, physiological or behavioural characteristics of a natural person, which allow or confirm the unique identification of that natural person" (GDPR art. 4). Secondly, the archive under consideration concerns health data[3]. Biometric and health data both belong to the special categories of data (sensitive or delicate data). According to the GDPR, special categories of personal data are "data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation" (Art. 9). Interviews with psychiatric patients relate to special categories of personal data; they took place inside the psychiatric hospital, and they explicitly and directly identify the subjects as 'patients' or – more often – as 'crazy' (with a label that also has serious consequences for the inpatient's family). Thus, for the data subject, the interviews highlight data about health, but also about sex life, racial/ethnic origin, religious, philosophical or other beliefs. All these data must be processed with special protection measures.
Another relevant question is whether curation and dissemination of archives is processing personal data. The GDPR conceptualizes the processing in a very broad manner so that it catches almost all data-related activities. According to the GDPR, processing *inter alia* covers collection, structuring, storage, adaptation retrieval, use, dissemination, erasure or destruction. It means that the curation and dissemination of the archive constitute the processing of personal data. The potential options for using speech archives are analysed in the next section.

4 The challenge: how to strike a fair balance between research, dissemination, and protection of privacy

The primary challenge for historical and linguistic research on past speech archives is represented by finding a balance between two socially relevant interests: the protection of personal data (the right to privacy) and the transmission of knowledge and freedom of research. Privacy and data protection do not exist in isolation. On the one hand, the Charter of Fundamental Rights of the European Union (Charter) protects private and family life and personal data (Art. 7-8). On the other hand, freedom of expression, information and science are also protected (Art. 11, 13). Even the GDPR itself expresses the following principle: "[t]he processing of personal data should be designed to serve mankind. The right to the protection of personal data is not an absolute right; it must be considered in relation to its function in society and be balanced against other fundamental rights, in accordance with the principle

[3] Early EU case law determined that even "[r]eference to the fact that an individual has injured her foot and is on half-time on medical grounds constitutes personal data concerning health" Case C-101/01). The analysed case has more intensive impact on the data subject's rights.

of proportionality" (Recital 4). Although the principles above can be used as guidelines, there is a need to search for possible solutions. The following routes are briefly described: duration of the data subject's rights, anonymisation, consent and research exemption.

The GDPR does not apply to the personal data of deceased persons (Recital 27). It means that EU member states can regulate the issue. WP 29 has correctly pointed out that data on the dead can relate to the living and be protected personal data (2007: 22). Therefore, this option is not a solution to the problem.

The GDPR also does not apply to anonymous data, which means the natural person is not identifiable (Recital 26). It is explained in the literature that "there is a strong incentive to anonymise data. Through anonymisation the data are placed outside the scope of data protection; by making data non-identifiable, the controller is relieved of the burden of compliance with data protection's rules and limitations" (Oostveen 2016: 307). WP29 in its opinion on the anonymisation techniques emphasizes "the potential value of anonymisation in particular as a strategy to reap the benefits of 'open data' for individuals and society at large whilst mitigating the risks for the individuals concerned. However, case studies and research publications have shown how difficult it is to create a truly anonymous dataset whilst retaining as much of the underlying information as required for the task" (2014: 3).[4] WP29 describes the problem very well. Anonymisation of data without destroying its informational value is almost impossible. In fact, anonymisation may not correspond to the needs and scope of historical research, which may be interested in, among others, the analysis of the relational structures of individuals. In short, historians are interested in "names and faces".

There are two potential legal grounds to use the archives (GDPR Art. 6): 1) consent; 2) processing in the public interest.

It is explained in the GDPR that it is not always possible to fully identify the purpose of processing for research purposes. Therefore, data subjects may give their consent to certain areas of scientific research (Recital 33). The GDPR defines consent as: "any freely given, specific, informed and unambiguous indication of the data subject's wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her" (Art. 4 (11)). There are several additional consent requirements which can be summarised as follows (GDPR Art. 7):

1) the controller has to demonstrate that the data subject has consented to processing;
2) consent is presented in a manner which is clearly distinguishable from the other matters;
3) consent is in an intelligible and easily accessible form, using clear and plain language;
4) the data subject has the right to withdraw his or her consent at any time:
    4.1) the withdrawal of consent does not affect the lawfulness of prior processing;
    4.2) the data subject is informed of the right to withdraw;
    4.3) it is as easy to withdraw as to give consent.

Processing of special categories of personal data requires explicit consent (GDPR Art. 9 (2)a).

The consent requirements are visualised in the following graph:

[4] It is also necessary to bear in mind that anonymisation itself is a further processing of personal data which must meet the GDPR requirements (WP29 2014: 3).

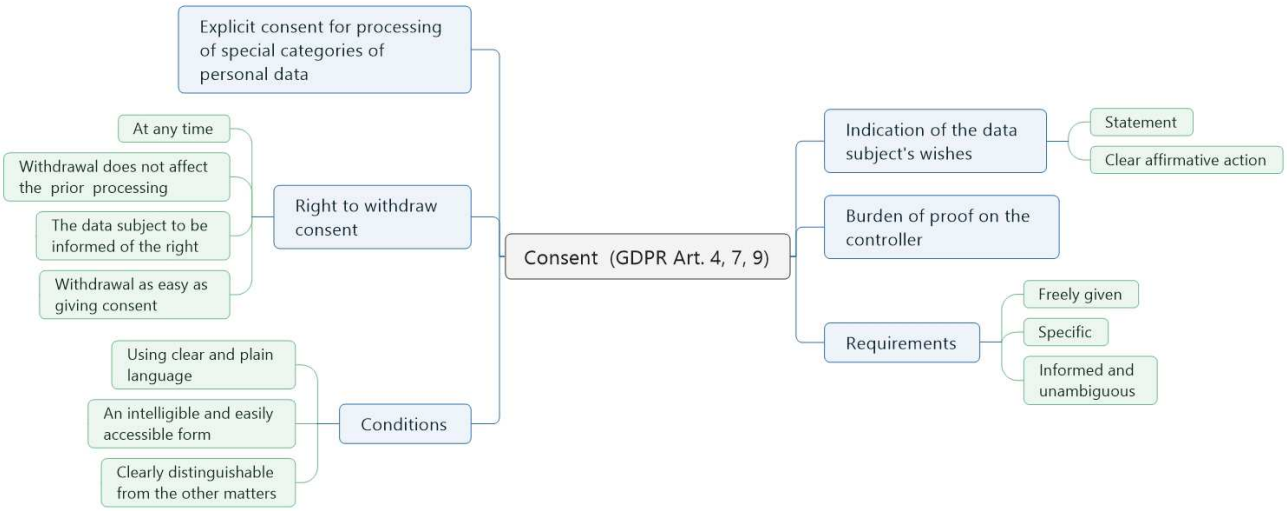


Fig. 6 Consent requirements in the GDPR regime.

The last option is to process personal data on the grounds of research exemption. GDPR prohibits the processing of special categories of personal data unless special grounds exist (Art. 9 (1)). Processing of special categories of personal data is allowed if it is necessary for scientific or historical research purposes. It must be proportionate to the aim pursued, respect the essence of the right to data protection and provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject (GDPR art. 9 (2) j) (For further discussion on processing personal data without consent see, Kelli, et al. 2018).

Given this general framework, we can return to the case study under investigation, since the re-use of Bruzzone's interviews for historical purposes could be a prime example of establishing a "legal chain" for personal data processing, which can be summarized as follows. Firstly, throughout the inspection of medical records – if any – preserved in the Psychiatric Hospital Archive, the research group[5] identified the real names of the patients and matched them with the pseudonyms, as attested to the volume (Bruzzone 1979). Secondly, the research group tried to go back to the interviewees, even contacting all the network of the people – physicians, nurses, social workers, ordinary citizens – involved in the recent history of the Psychiatric hospital.
This reconstruction helped to obtain detailed and clear informed consents (for further discussion on the consent see WP29 2017; GDPR art. 7, 9 (2) a), describing the aims, the scope and the positive spill-over effects of the dissemination of such an oral archive (see Appendix, with the translated consent form). If the consent form is obtained, the oral archive could be finally enjoyed by the research communities and the entire civil society.
At present two former patients were discovered. One of the two – who is now living an ordinary life – signed the consent form, gave additional interviews to the research group, and also actively collaborated with the project. The other one is now in a nursing home and has a legal guardian who agreed to sign the informed consent. Reasonably, it might be concluded that all the other voices belong to deceased persons. A plan of communication and dissemination was therefore set up in order to reach possible relatives and right holders, and two events were organised in late 2018 in order to disseminate the research project. One was included in Bright 2018 The European Researchers' Night at University of Siena (in Arezzo), while the other took place in an extra-academic context, in order to reach a different public and to involve different networks (see Fig. 7 and 8 respectively).

[5] Carried out by a linguist (the first author), a philosopher of science (Marica Setaro), an oral historian (Caterina Pesce), and an archivist (Lucilla Gigli).

Fig. 7 The European Researchers' Night at University of Siena

Fig. 8 The poster of the dissemination event ("Shared voices") in a meeting place in Arezzo

There was much media coverage of the events. The regional television news interviewed Roberto (the former patient) who listened live to his 1977 voice (see Fig. 9)[6].

Fig. 9 The boardcast service ("Voices from the madhouse").

5 Conclusions

The analysed archive is subject to the GDPR since it contains special categories of personal data. The curation and dissemination of the archives is the processing of personal data which requires legal grounds and other measures assuring GDPR compliance. Personal data relating to the deceased data subject could still be protected due to its links to living individuals. One option would be data anonymisation which is not always an option for historical research. An additional option is to acquire explicit informed consent. If it is not possible to obtain consent, the research exemption might be applicable. The use of research exception requires the introduction of safeguards protecting the data subject's rights (e.g. pseudonymization, limited access, and so forth).

[6] The broadcasting service can be retrieved at the following url: https://www.rainews.it/tgr/toscana/video/2018/09/tos-arezzo-manicomio-bruzzone-88ee6948-a3ab-4135-a034-fdec303b90cd.html?wt_mc=2.www.fb.tgrtoscana_ContentItem-88ee6948-a3ab-4135-a034-fdec303b90cd.&wt (01.04.2019)

References

[Case C-101/01] Case C-101/01. Criminal proceedings against Bodil Lindqvist. 6 November 2003. Available at http://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1521039149443&uri=CELEX:62001CJ0101 (14.4.2018);

[Charter] Charter of Fundamental Rights of the European Union. 2012/C 326/02. OJ C 326, 26.10.2012, p. 391–407 (BG, ES, CS, DA, DE, ET, EL, EN, FR, IT, LV, LT, HU, MT, NL, PL, PT, RO, SK, SL, FI, SV). Available at http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:12012P/TXT (14.4.2018);

[Data Protection Directive] Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. Official Journal L 281, 23/11/1995 p. 0031 – 0050. Available at http://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:31995L0046&qid=1522340616101&from=EN (29.3.2018);

[GDPR] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). OJ L 119, 4.5.2016, p. 1-88. Available at http://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1515793631105&uri=CELEX:32016R0679 (29.3.2018);

[González-Rodríguez et. al. 2008] Joaquín González-Rodríguez, Doroteo Torre Toledano, Javier Ortega-García (2008). Voice Biometrics. In Handbook of Biometrics edited by Anil K. Jain, Patrick Flynn, Arun A. Ross. Springer;

[IPDPC] Italian Personal Data Protection Code. Legislative Decree 30.06.2003 No. 196. English version available at: http://194.242.234.211/documents/10160/2012405/Personal+Data+Protection+Code+-+Legislat.+Decree+no.196+of+30+June+2003.pdf (11.4.2018);

[Jain et. al. 2004] Anil K. Jain, Arun Ross, Salil Prabhakar (2004). An Introduction to Biometric Recognition. - IEEE Transactions on Circuits and Systems for Video Technology 14(1). Available at https://www.cse.msu.edu/~rossarun/BiometricsTextBook/Papers/Introduction/JainRossPrabhakar_BiometricIntro_CSVT04.pdf (31.3.2018);

[Kelli et al. 2015] Aleksei Kelli, Kadri Vider, Krister Lindén (2015). The Regulatory and Contractual Framework as an Integral Part of the CLARIN Infrastructure. 123: Selected Papers from the CLARIN Annual Conference 2015, October 14–16, 2015, Wroclaw, Poland. Ed. Koenraad De Smedt. Linköping University Electronic Press, Linköpings universitet, 13−24. Available at https://www.ep.liu.se/ecp/article.asp?issue=123&article=002 (20.8.2018);

[Kelli et al. 2018] Kelli, Aleksei; Lindén, Krister; Vider, Kadri; Kamocki, Pawel; Birštonas, Ramūnas; Calamai, Silvia; Kolletzek, Chiara; Labropoulou, Penny; Gavriilidou, Maria (2018). Processing personal data without the consent of the data subject for the development and use of language resources. CLARIN Annual Conference 2018 Proceedings: CLARIN Annual Conference 2018, 8-10 October 2018 Pisa, Italy. Ed. Inguna Skadin, Maria Eskevich. CLARIN, 43−48. Available at https://office.clarin.eu/v/CE-2018-1292-CLARIN2018_ConferenceProceedings.pdf (28.1.2019);

[Oostveen 2016] Manon Oostveen. Identifiability and the applicability of data protection to big data. International Data Privacy Law, 2016, Vol. 6, No. 4, 299- 309;

[Spindler, Schmechel 2016] G. Spindler, P. Schmechel (2016). Personal Data and Encryption in the European General Data Protection Regulation. - JIPITEC 7 (2), 163-177. Available at https://www.jipitec.eu/issues/jipitec-7-2-2016/4440/spindler_schmechel_gdpr_encryption_jipitec_7_2_2016_163.pdf (14.4.2018);

[WP29 2017] WP29. Guidelines on Consent under Regulation 2016/679. Adopted on 28 November 2017 [adopted, but still to be finalized]. Available at http://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=615239 (2.4.2018);

[WP29 2014] WP29. Opinion 05/2014 on Anonymisation Techniques. Adopted on 10 April 2014. Available at http://collections.internetmemory.org/haeu/20171122154227/http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf (20.8.2018).

[WP29 2007] WP29. Opinion 4/2007 on the concept of personal data. Adopted on 20th June. Available at https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2007/wp136_en.pdf (26.3.2019).

Appendix. The English translation of the informed consent form

To the relatives of the former patients of the Neuro-psychiatric hospital in Arezzo

Dear Sir/Dear Madam,

We are a research group at Siena University and we are writing to you because we have succeded in finding the compact cassettes collected in the summer of 1977 containing the interviews of the patients in the Neuro-psychiatric hospital carried out by Anna Maria Bruzzone (and partially published in the volume *Ci chiamavano matti*, Einaudi 1979). These cassettes contain invaluable and unique information and we hope that in the near future it will be possible to listen to these precious and touching life stories, and thus to give voice to voiceless people.

These cassettes were digitized and were located in the Arezzo archive. In order to make them accessible for to scientific research it is necessary to provide an informed consent from the relatives of the patients. Obviously, the single interview will be disseminated anonymously or under the pseudonym used by Bruzzone herself, with all due caution. We do want first to honor your wishes, but at the same time we also want to fulfil our research duties, i.e. *to ensure that such information is not lost and that a relevant piece of history – non only for the city of Arezzo – could pass to future generations*.

Therefore we need your consent in order to work – as researchers – on the retrieved speech archive, which is now bound by the „Soprintendenza Archivistica e Bibliografica della Toscana". Needless to say, all our activities are without interest for financial gain and take place at University (as part of teaching, research, and public engagment). If you agree with the spirit that underpins our initiative, we kindly ask you to sign the attached informed consent. We also would like to offer you a digital copy of the interview containing the voice of your relative.

We are at your disposal for additional information.

Contact information:

Silvia Calamai, coordinator of the Scientific Committee of the Historical Archive of the neuro-psychiatric hospital silvia.calamai@unisi.it 0575926439

Lucilla Gigli, archivist at the Historical Archive of neuro-psychiatric hospital lucilla.gigli@unisi.it 0575926264 - 0575926292

Authorisation to use the interview

I undersigned _____

Born _____ on _____ and resident in

_____

First-degree relative of_____

(or)

Second-degree relative of _____

Authorise

The research group led lead by Silvia Calamai, associate professor at Siena University, Department of Education, Humanities and Intercultural Communication

1- to the re-use of my relative's interview for teaching and research activities:

[yes] [no]

2- to the consulting of the interview by third parties under the following condition: *(choose only one option)*

[ ] free consulting

[ ] consulting after _____ years from the present date

[ ] consulting subordinate to my approval

I relinquish all rights on a royalty-free basis for

- Teaching and research activities [yes] [no]

- Publishing of the partial or full interview [yes] [no]

Any other use shall be subject to my additional authorisation.

My details are as follows:

Postal address_____

E-mail address_____

Phone number_____

All personal data will be used and processed by Siena University solely for the present purposes, and their protection and confidentiality will be ensured in accordance with General Data Protection Regulation (art. 5, art 15ff).

Place and date Signature

_____ _____

Corpus-driven conversational agents: tools and resources for multimodal dialogue systems development

Maria Di Maro
Department of Humanities
University of Naples 'Federico II', Italy
maria.dimaro2@unina.it

Abstract

In this paper, we describe how tools made available through CLARIN can be applied for research purposes in the development of corpus-driven conversational agents. The starting point will be the description of a standard architecture for multimodal dialogue systems. For some of its parts, specific available tools will be briefly described, according to their suitability to mutimodal dialogue systems development.

1 Introduction

The present paper gives an overview on tools and resources available within the CLARIN infrastructure, which can be exploited in the development of conversational agents, especially as far as language and dialogue modelling are concerned. Spoken dialogue systems are nowadays in the spotlight in different commercial, academic and industrial sectors: it will suffice to consider the success and popularity of tools like Amazon Alexa and Google Home [López et al., 2017], or of the widespread in-car dialogue systems [Becker et al., 2006, Kousidis et al., 2014]. Conversational Agents are computer systems capable of conversing with humans. These dialogue systems are one of the most currently researched field in Artificial Intelligence, since the ability to communicate ones understanding by means of language is one possible way to manifest intelligence. In the Macmillan Dictionary[1], *intelligence* is defined as the ability to understand and think about things, and to gain and use knowledge. In this definition, one concept draws particular attention: 'knowledge'. Building the knowledge base for such systems is the first step to give them intelligence. For this particular goal, the use of some tools facilitates the job of interaction designers, such as linguists. At the two extremes of the learning continuum, we find on the one hand deterministic rules given to the system to interpret some particular signals and react to them appropriately [McGlashan et al., 1992], whereas on the other hand we have end-to-end dialogue systems which do not make any distinction in the abilities the system should perform at different levels, but it is provided with data from which tendencies are statistically extracted [Ritter et al., 2010, Vinyals and Le, 2015, Serban et al., 2016, Bordes et al., 2016]. In the middle, we have the possibility to train different modules with the application of different strategies and tools. Overall, the corpus-driven approach is becoming more and more important to infer knowledge and communicative strategies in the field of spoken language understanding and generation for applying different statistic and machine learning algorithms [Serban et al., 2018]. This means that appropriate collection of data, in combination with specific tools, are required to model one's own system.

In this work, we will concentrate on multimodal dialogue systems, which not only make use of spoken language, but which also use other communication channels to understand and express intents [Lucignano et al., 2013]. For this reason, the knowledge to be constructed will comprise different linguistic and paralinguistic levels. The standard architecture for a multimodal dialogue system consists of different modules, which serves one another to build the interaction (Figure 1). The input elaborated by the user is first processed by a module, which takes the audio produced by the user and transform it in a string to

[1]Macmillan Dictionary Online: https://www.macmillandictionary.com/ [last consultation on the 24th January 2019]

Maria Di Maro 2019. Corpus-driven conversational agents: tools and resources for multimodal dialogue systems development. *Selected papers from the CLARIN Annual Conference 2018*. Linköping Electronic Conference Proceedings 159: 39–45.

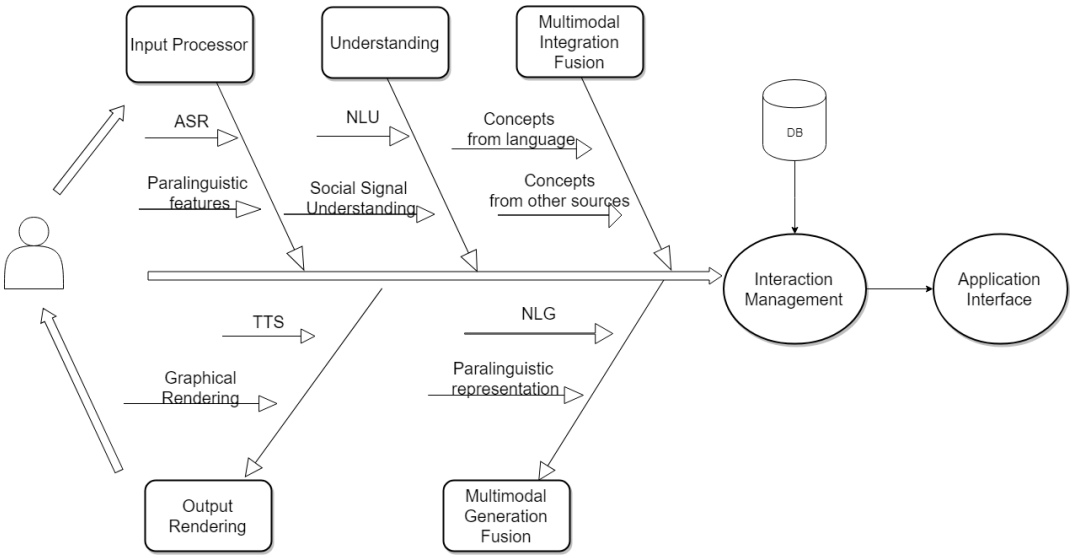


Figure 1: Multimodal Dialogue System Architecture

be further analysed. Parallel to that, gestures, facial expression, prosody and other paralinguistic features arising from the interaction are captured by sensors. The classification and consequent understanding of the meaning of the linguistic and paralinguistic inputs are processed in the second module. The meaning associated to the received signals are fused together to recognise a single intent. The decision concerning the flow of the interaction are taken in the Interaction Management module, which is connected to a knowledge base including the information concerning the accomplishment of specific intents. When all the decisions are statistically or deterministically taken, the linguistic and paralinguistic intent representations are generated. In the last module, tools are used to synthesise the voice with peculiar prosodic characteristics, according to the intent, correlating it to other paralinguistic aspects, such as gestures, facial expressions, and posture. In the next sessions, we will focus on available tools, which can be usefully exploited in the development of some of the above-described modules.

2 CLARIN Tools for Training and Modelling Purposes

For the development of such systems, different approaches, data and tools can be used. For instance, as far as corpus-driven dialogue systems are concerned, there is a vast amount of data documenting human dialogues. Furthermore, annotation standards, annotation platforms, or tools for extracting different kinds of signals can also be exploited in the dialogue development framework. In this particular report, we are going to focus on applications which can be specifically used in the design of a dialogue system for the Italian language. Specifically, we present some tools which are being used for the linguistic and paralinguistic development of a conversational agent, to be framed as part of an ongoing national project, namely CHROME (*Cultural Heritage Resources Orienting Multimodal Experiences*), whose aim is to define a methodology of collecting, analysing and modelling multimodal data in designing virtual agents serving in museums [Cutugno et al., 2018]. The linguistic part is, therefore, mainly concerned with building the interaction with the resulting virtual gatekeeper, which will guide museum visitors in the exploration of cultural contents. In more details, starting from an empirical study of conversational phenomena, especially in cultural heritage domains, common ways of expressing requests and inquiries by visitors, and strategies of communicating cultural contents by guides will be collected and analysed, along with semantic, syntactic and paralinguistic language-dependent strategies. For these purposes, in the next sections, we are going to describe the use of some sources made available via the CLARIN infrastructure, especially as far as input processor, dialogue modelling and multimodal alignment are concerned.

2.1 Input Processor

By input processor, we mean here the pre-processing of speech data, on the basis of which the recognition of specific signals from the audio is modelled and defined. In fact, speech corpora can be used to extract prosodic profiles connected to communicative strategies, in order to train the system to consequently recognise them or use them in specific situations. For this purpose, the web service WebMAUS[2] can be used to fulfil specific phonetic requirements. The Munich AUtomatic Segmentation (MAUS) system [Schiel, 1999, Kisler et al., 2017] is a multilingual tool used to transcribe audio inputs and align transcription to the spectrogram, returning as a result a TextGrid file[3]. Beside the graphic transcription, which can be provided or can be left to the integrated ASR (Automatic Speech Recogniser), the tool also provides the phonetic one in SAMPA for each word and each phone, as in Figure 2. It also provides related services, such as TTS (Text-to-Speech), syllabification, and chunking. By using the resulting files, particular phonetic features, which can be associated to the semantics of linguistic intents, can be extracted, such as intonation, pitch and intensity. For the manual or automatic extraction, the Praat program [Boersma and Weenink, 2002] can be adopted. Furthermore, the obtained data can also be used to outline sociolinguistic profiling of speakers by extracting pieces of information connected to the openness of vowels and other articulative peculiarities, as in [Di Maro et al., 2018]. In the next section, this aspect will be highlighted with regard to the use of annotated spoken corpora with regional variaties, such as CLIPS.

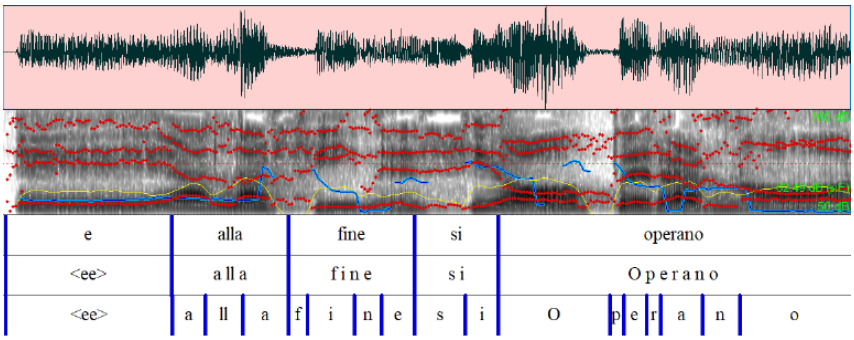


Figure 2: Resulting TextGrid file of a MAUS forced Alignment in Praat

2.2 Dialogue Modelling

Dialogue Modelling refers to the design of the dialogic exchange as far as intents definition and output mapping are concerned. Strictly connected to dialogue modelling is the definition of the communicative strategies arising in conversation, among which we can mention the turn-taking organisation [Sacks et al., 1978]. For the semantic and pragmatic design of dialogues, different sources can be exploited. Among various techniques, the use of SRGS (Speech Recognition Grammar Specification)[4] [Hunt and McGlashan, 2004] is mostly preferred to assure the categorisations of possible intents in a target-oriented dialogue system, with means of the description of each possible structure that can be uttered to express a particular concept. The use of grammars is especially suitable for commercial systems, whose domain can be deterministically better defined, avoiding relying on error-prone machine learning algorithms. These grammars can be automatically extended, as far as lexical variability and inflectional morphology is concerned [Di Maro et al., 2017], making use of semantic networks such as ItalWordNet [Roventini et al., 2000] and POS-tagging tools like Tree-Tagger [Schmid et al., 2007].

The language model to be used for conversational purposes can be enriched with pragmatic information. For this purpose, the Dialogue Act Mark-up Language (DiAML) could be used. Not only is it suitable to annotate the type of intent performed, but it is also effective to specify further information: i)

[2]https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface

[3]A TextGrid file is a text file used for labelling segments of an audio file. It is used in Praat to show the labels lined up to the audio segments.

[4]Speech Recognition Grammar Specification Version 1.0: https://www.w3.org/TR/speech-grammar/

whether the user intent was merely dependent on the action motivating the dialogue itself; ii) whether it was a feedback to the previous turn (auto- and allo-feedback); iii) if it was signalling the turn-giving or turn-taking action; iv) opening, closing or structuring the conversation; v) in case of social obligations adjacency pairs [Bunt et al., 2010]. The specification of the performed act is indeed useful to improve the disambiguation and thus the understanding. For instance, knowing when a museum visitor is giving a feedback on something previously uttered by the guide or asking for more information or clarifications on the same concept is important to assure an appropriate reaction by the virtual agent.

Besides the rule-based approaches, which can make use of grammars, we can use corpora for the statistical extraction of knowledge. Data analysis can be both corpus-based and corpus-driven: on the one hand a given corpus can help to confirm or refute a pre-existing theoretical construct (corpus-based), on the other hand a corpus can be used to generalise rules (corpus-driven). For modelling conversational interactions, spoken corpora are useful to capture all the domain-dependent semantic aspects and the pragmatic characteristics arising from dialogues. Therefore, a corpus-driven approach is preferably adopted. To achieve such aims, the construction of tools like SPOKES is truly interesting. SPOKES - currently available in Polish and English - is an online service for conversational corpus data search and exploration [Pezik, 2015]. By exploring this corpus, information concerning the strategies used in conversation can be extracted to be modelled in a ones own language model. As a result of the research project here described, an Italian version of SPOKES is also desirable. Providing pragmatic annotation in such tools is also an advisable goal to better be applied in the development of conversational agents. As far as the current availability of spoken corpora for Italian, some of them are summarised in Table 1.

Corpus	**Annotation**
AN.ANA.S._MT [5]	syntactic information
Corpus AVIP-API [6]	orthographic transcription
CLIPS[7]	segmental information
EXMARaLDA Demo Corpus [8]	suprasegmental information, accentuation/stress marking
SpIt-MDb[9]	acustic, phonetic, phonological, and lexical information

Table 1: Italian Spoken Corpora

In particular, CLIPS [Savy, 2009] contains dialogues from speakers coming from 15 different Italian cities. This could be useful to train a system to recognise the geographical origin of the speaker for profiling purposes. Among the others, we mention AN.ANA.S 4 [Voghera and Cutugno, 2009] which contains syntactic annotations and whose information could be used for training the system to recognise syntactic structures and disambiguate semantic usages.

In a multimodal perspective, speech and gestures corpora are a further asset in the exploitation of data for training dialogue systems. In particular, deictic information or ellipsis can be recovered by the listener via the the interpretation of gestures. An explanatory example is drawn by the SaGa Corpus [Lücking et al., 2010]. The SaGA corpus consists of 280 minutes of video material containing 4961 iconic/deictic gestures, approximately 1000 discourse gestures and 39,435 words. The annotation comprises gesture segmentation and classification (iconics, deictics, beats), gestural representation techniques (e.g., draw-

[5] AN.ANA.S._MT Corpus. Archived at the University of Salerno. Published in 2010. http://www.parlaritaliano.it/index.php/it/corpora-di-parlato/716-corpus-ananas-multilingue-ananasmt

[6] Corpus AVIP-API. Archivio del Parlato Italiano. Archived at the University of Salerno. Published in 2003. http://www.parlaritaliano.it/api/

[7] CLIPS Corpus. Archived at the University of Naples 'Federico II'. Published in 2005. http://www.clips.unina.it/it/

[8] EXMARaLDA Demo Corpus 1.0. Archived in Hamburger Zentrum für Sprachkorpora. Publication date 2007-11-08. http://hdl.handle.net/11022/0000-0000-4F70-A.

[9] SpIt-MDb Corpus (Spoken Italian - Multilevel Database). Archived at the University of Salerno. Published in 2006. http://www.parlaritaliano.it/index.php/it/corpora-di-parlato/644-spit-mdb-spoken-italian-multilevel-database

ing, placing), morphological gesture features (e.g., hand shape, hand position, palm orientation, movement features), transcription of spoken words and dialogue context information, based on DAMSL dialogue acts, information focus, and thematization [Lücking et al., 2010].

The use of multimodal corpora is also particularly interesting when considering that identical utterances can take on different meanings according to not only the intonational and prosodic structure of the message being conveyed but also according to gestures or facial expressions we use while uttering it. The collection of multimodal corpora is therefore configurable as a necessity. For the Italian language, there are not a lot of data sources, besides language learning (L2) collections, such as the TAITO-project. Nevertheless, a multimodal and multi-party corpus for the Italian language, specifically applied in the cultural heritage domain, has been collected for the CHROME project [Origlia et al., 2018, Cutugno et al., 2018].

2.3 Multimodal Alignment

The module responsible for the fusion of different channels of intents communication - spoken language and paralinguistic features, specifically gestures and prosodic profiles - can rely on data synchronised with a tool like ELAN, before being learned through probabilistic rules or machine learning algorithms. ELAN is a tool designed to annotate audio and video files [Wittenburg et al., 2006]. In ELANs tiers, TextGrids, which are for instance obtained with WebMAUS, can be imported and overlapped to the other pragmatic and paralinguistic information. The fusion of the different annotation levels can be used to process both the understanding and the generation processes. For instance, this tool is being used within the CHROME project to specifically model the way the gatekeeper would communicate cultural contents (Figure 3). After having recorded authentic tour guides, video and audio files have been synchronised in ELAN, where expert annotators marked linguistic and paralinguistic phenomena [Origlia et al., 2018]. In addition to that, postures, gestures and facial expressions of listeners are annotated to capture their aptitude towards the content being conveyed. Fusing different channels of communication together in the modelling phase will result in a virtual tourist guide able to communicate as naturally as human ones, capable of adapting their communicative strategies to the type of interlocutor. In addition to ELAN, pragmatic phenomena can also be manually annotated using tools such as EXMARaLDA [Schmidt and Wörner, 2009], a system for the computer-assisted creation and analysis of spoken language corpora.

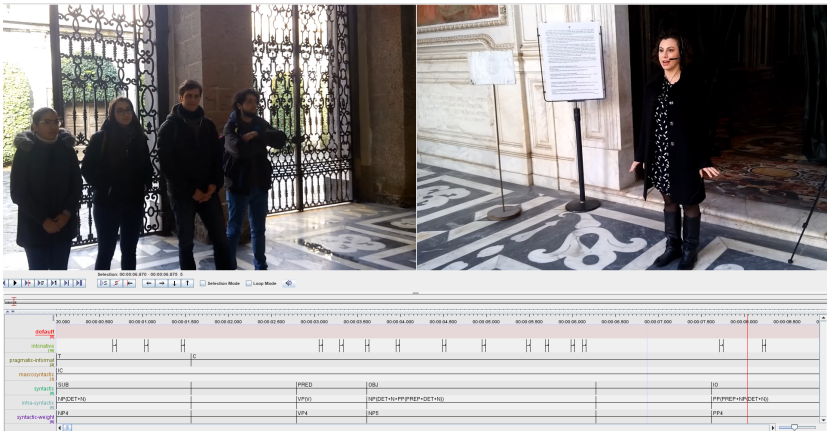


Figure 3: Example of the multimodal annotation of the CHROME corpus via ELAN

3 Conclusion

In this paper, a brief overview of CLARIN tools to be applied in the development of multimodal conversational agents has been presented. This framework will be further developed as a PhD research project, which is part of the Italian National Project CHROME. Specifically, the main aim of the research will be to build a conversational agent for cultural heritage, capable of interpreting multimodal communicative feedback in order to present cultural contents which are adapted to the interpreted mental state and pref-

erences of the human interlocutor. The development of other conversational annotated data to be made available for similar researches is a desirable part of the presented research.

References

Tilman Becker, Nate Blaylock, Ciprian Gerstenberger, Ivana Kruijff-Korbayová, Andreas Korthauer, Manfred Pinkal, Michael Pitz, Peter Poller, and Jan Schehl. 2006. Natural and intuitive multimodal dialogue for in-car applications: The sammie system. *Frontiers in Artificial Intelligence and Applications*, 141:612.

Paul Boersma and David J. M. Weenink. 2002. Praat, a system for doing phonetics by computer. *Glot international*, 5.

Antoine Bordes, Y-Lan Boureau, and Jason Weston. 2016. Learning end-to-end goal-oriented dialog. *arXiv preprint arXiv:1605.07683*.

Harry Bunt, Jan Alexandersson, Jean Carletta, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Kiyong Lee, Volha Petukhova, Andrei Popescu-Belis, Laurent Romary, et al. 2010. Towards an iso standard for dialogue act annotation. In *Seventh conference on International Language Resources and Evaluation (LREC'10)*.

Francesco Cutugno, Felice DellOrletta, Isabella Poggi, Renata Savy, and Antonio Sorgente. 2018. The chrome manifesto: integrating multimodal data into cultural heritage resources. *Proceedings of the Fifth Italian Conference on Computational Linguistics, CLiC-it 2018*.

Maria Di Maro, Marco Valentino, Anna Riccio, and Antonio Origlia. 2017. Graph databases for designing high-performance speech recognition grammars. In *IWCS 201712th International Conference on Computational SemanticsShort papers*.

Maria Di Maro, Sara Falcone, and Francesco Cutugno. 2018. Prosodic analysis in human-machine interaction. *Studi AISV*, 1:to appear.

Andrew Hunt and Scott McGlashan. 2004. Speech recognition grammar specification version 1.0. *W3C Recommendation, March*.

Thomas Kisler, Uwe Reichel, and Florian Schiel. 2017. Multilingual processing of speech via web services. *Computer Speech & Language*, 45:326–347.

Spyros Kousidis, Casey Kennington, Timo Baumann, Hendrik Buschmeier, Stefan Kopp, and David Schlangen. 2014. A multimodal in-car dialogue system that tracks the driver's attention. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 26–33. ACM.

Gustavo López, Luis Quesada, and Luis A Guerrero. 2017. Alexa vs. siri vs. cortana vs. google assistant: a comparison of speech-based natural user interfaces. In *International Conference on Applied Human Factors and Ergonomics*, pages 241–250. Springer.

Lorenzo Lucignano, Francesco Cutugno, Silvia Rossi, and Alberto Finzi. 2013. A dialogue system for multimodal human-robot interaction. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 197–204. ACM.

Andy Lücking, Kirsten Bergmann, Florian Hahn, Stefan Kopp, and Hannes Rieser. 2010. The bielefeld speech and gesture alignment corpus (saga). In *LREC 2010 workshop: Multimodal corpora–advances in capturing, coding and analyzing multimodality*.

Scott McGlashan, Norman Fraser, Nigel Gilbert, Eric Bilange, Paul Heisterkamp, and Nick Youd. 1992. Dialogue management for telephone information systems. In *Proceedings of the third conference on Applied natural language processing*, pages 245–246. Association for Computational Linguistics.

Antonio Origlia, Renata Savy, Isabella Poggi, Francesco Cutugno, Iolanda Alfano, Francesca D'Errico, Laura Vincze, and Violetta Cataldo. 2018. An audiovisual corpus of guided tours in cultural sites: Data collection protocols in the chrome project. In *Proceedings of the 2018 AVI-CH Workshop on Advanced Visual Interfaces for Cultural Heritage*, volume 2091.

Piotr Pezik. 2015. Spokes-a search and exploration service for conversational corpus data. In *Selected Papers from the CLARIN 2014 Conference, October 24-25, 2014, Soesterberg, The Netherlands*, pages 99–109. Linköping University Electronic Press.

Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised modeling of twitter conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 172–180. Association for Computational Linguistics.

Adriana Roventini, Antonietta Alonge, Nicoletta Calzolari, Bernardo Magnini, and Francesca Bertagna. 2000. Italwordnet: a large semantic database for italian. In *LREC*.

Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1978. A simplest systematics for the organization of turn taking for conversation. In *Studies in the organization of conversational interaction*, pages 7–55. Elsevier.

Renata Savy. 2009. Clips: diatopic, diamesic and diaphasic variations of spoken italian. In *Proceedings of the Corpus Linguistics Conference 2009 (CL2009),*, page 213.

Florian Schiel. 1999. Automatic phonetic transcription of non-prompted speech. *Proc. of the ICPhS*, pages 607–610.

H Schmid, M Baroni, E Zanchetta, and A Stein. 2007. The enriched treetagger system. In *proceedings of the EVALITA 2007 workshop*.

Thomas Schmidt and Kai Wörner. 2009. Exmaralda–creating, analysing and sharing spoken language corpora for pragmatic research. *Pragmatics. Quarterly Publication of the International Pragmatics Association (IPrA)*, 19(4):565–582.

Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*, volume 16, pages 3776–3784.

Iulian Vlad Serban, Ryan Lowe, Peter Henderson, Laurent Charlin, and Joelle Pineau. 2018. A survey of available corpora for building data-driven dialogue systems: The journal version. *Dialogue & Discourse*, 9(1):1–49.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Miriam Voghera and Francesco Cutugno. 2009. An. ana. s.: aligning text to temporal syntagmatic progression in treebanks. In *Proceedings of the 5th Corpus Linguistics Conference, Liverpool*, pages 20–23.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. Elan: a professional framework for multimodality research. In *5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 1556–1559.

Looking for hidden speech archives in Italian institutions

Vincenzo Galatà
Institute of Cognitive Sciences and Technologies, National Research Council, Italy and University of Siena, Italy
vincenzo.galata@pd.istc.cnr.it

Silvia Calamai
University of Siena, Italy
silvia.calamai@unisi.it

Abstract

The paper presents the aims and the main results of an on-line survey concerning speech archives collected in the fields of Social Sciences and Humanities among Italian scholars. A huge amount of speech archives is preserved among researchers: most of the resources are not accessible and legal issues are generally not addressed in detail. The great majority of the respondents would agree to storing their archives in national repositories, if any.

1 Introduction

Very few surveys describe the amount and the size of speech archives in Italy. To our knowledge, only Barrera et al. (1993) and Benedetti (2002) map the existing audio archives. The first survey was made under the aegis of the Ministry of Cultural Heritage and listed only the public archives (Barrera et al. 1993). Benedetti (2002) listed also private archives, especially in the field of music. Sergio (2016) presented the photo and audiovisual archives that were digitised (or were in the process of being digitised) by public and private institutions in Italy. Other surveys were limited to a single area, such as AAVV (1999), devoted to Piedmont region, and Andreini & Clemente (2007) and Cappelli & Rioda (2009) which restricted the survey to the Tuscany. Partial inventories can be found scattered around the internet, especially within the context of the "Istituti Italiani per la Resistenza". It has to be noted that the great majority of the inventories focused on music and oral history archives and completely neglected the huge amount of material collected by linguists during their fieldwork. At the European level, an overview on the Oral History collections was made accessible and maintained by CLARIN ERIC[1]. At present, the overview contains about 260 collections scattered in 17 European countries (with great disparities between EU countries in terms of coverage and details). As for Italy, 86 collections are listed (data were collected in 2016 by the second author together with the Italian Association for Oral History).

The present paper aims at providing an updated map of Italian speech archives generated by field researches within and outside the academia, especially in the areas of linguistics and oral history, but also in other areas that we included while the survey was already running. Most of the archives we discovered are inaccessible and can be labelled as audio 'legacy data': that is, data stored in obsolete audio media by individual researchers outside of archival sites such as libraries or data centres. For this purpose, we set up an online survey in order to:

i) draft a survey of institutional archives, that is a survey of the existing speech archives deposited in (and by) institutions and associations;
ii) draft a survey of the existing speech archives owned by single researchers;
iii) provide an extensive analysis of the existing practices of collection, preservation and reuse in order to give a detailed description of the state of conservation and accessibility, the access policies, costs and sustainability.

The survey also made it possible to verify how the knowledge of the CLARIN infrastructure is widespread among Italian research communities. A bottom-up approach, involving the main Italian scientific associations, allowed us to reach as many researchers as possible and to bring a hidden, inaccessible, endangered treasure to light.

The paper is conceived as follows: §2 presents the structure and the content of the questionnaire prepared to run the survey; §3 reports on the sample that answered to the survey together with the main

[1] http://oralhistory.eu/collections/clarin-eric.

Vincenzo Galatà and Silvia Calamai 2019. Looking for hidden speech archives in Italian institutions. *Selected papers from the CLARIN Annual Conference 2018.* Linköping Electronic Conference Proceedings 159: 46–55.

results; §4 addresses some concluding remarks and underlines the urgent need to find an Italian repository to host these materials.

2 The questionnaire

The survey was jointly designed by both authors and was administered in Italian through an online questionnaire (implemented via Google Forms). The following Italian scholarly organisations were involved in the dissemination of the survey by means of their respective mailing list: *CLARIN-It*, *Associazione Italiana di Scienze della Voce* (AISV), *Associazione Italiana di Storia Orale* (AISO), *Società di Linguistica Italiana* (SLI), *Associazione di Storia della Lingua Italiana* (ASLI), *Associazione Italiana di Linguistica Applicata* (AITLA), *Società Italiana per la Museografia e i Beni Demoetnoantropologici* (SIMBDEA), *Associazione Italiana di Sociologia* (AIS), *Società Italiana di Antropologia Culturale* (SIAC), *Società Italiana di Antropologia Applicata* (SIAA), *Associazione Nazionale Professionale Italiana di Antropologia* (ANPIA). Other formal and informal networks were targeted (e.g. *Analisi dell'Interazione e della Mediazione* network, AIM) and also individual researchers – both in Italy and abroad – were personally contacted by email. We can presumably assume that several hundred scholars were reached by the survey.

The questions included in the survey were mostly yes-no and multiple response type (for three questions, for which it was impossible to predict or suggest any option, open answers were allowed). The questions were as generic and as inclusive as possible in order to be answered by all of the respondents and thus avoiding to focus on very specific scientific domains. In most of the cases, besides every yes-no or multiple response question, an "*Other, please specify*" field was provided in order to account for responses not foreseen by the authors. The choice of including such an open-ended response option had the disadvantage of increasing the amount of post-processing needed at the time of results reporting for each question (the responses resulting in highly scattered distributions), but at the same time this allowed the authors to account for multiple domains and issues not previously considered.

The survey was administered in Italian and was structured according to four distinct sections:

1) the first section was mainly informative and preceded the questionnaire itself by providing a brief presentation of the aims and the scope of the survey, as well as general information on the treatment of the recorded responses to the questionnaire;
2) the second section (displayed in the Appendix) contained the actual questionnaire consisting of 19 questions. The first question (Q.0) gave the participants the possibility to opt-out from the survey (thus registering their participation) or eventually to contribute to the survey, without necessarily completing the survey, by jumping to the third section of the survey (see point 3). The core questionnaire, consisting of 17 questions (Q.1 to Q.17), was devised in order to obtain a rough description and quantification of audio-visual resources (also with respect to accessibility and legal issues). One last question (Q.18) asked the respondents if they were aware of the existence of the CLARIN European infrastructure. A translated version of this section is provided in Appendix at the end of the paper;
3) the third section allowed all the respondents to contribute to the survey dissemination by suggesting the authors further potential contacts they considered worth to be contacted;
4) the last section of the questionnaire asked the respondents for some personal information (contact, academic position and affiliation).

For the aim of the current paper, in the next paragraph we report the results from selected questions of the survey, leaving the rest and more elaborate analyses to an extended version on the same topic. The questions on which we focus here are intended to:

1) uncover the scientific domains with the highest amount of hidden spoken resources;
2) identify what sort of resources we are coping with;
3) understand if digitised data (such as transcriptions, annotations etc.) are eventually available for these resources and in what format they are stored;
4) establish if the mentioned resources are accessible and who is in charge of their maintenance;
5) take stock of the ethical issues related to the creation of the resources under scrutiny;
6) assay how much the knowledge of the CLARIN European infrastructure is widespread in the different scientific domains.

3 Main results

The results we report on in this section refer to the responses gathered from the survey at the time of writing[2] with reference to selected questions as mentioned in the previous paragraph. So far, 151 respondents took part in the survey: 131 participants completed the survey, 17 opted-out and 3 only suggested other contacts. If we consider the affiliation of the respondents specified in the last section of the questionnaire, we reached 86 individuals declaring an affiliation in Italy and other 8 with affiliation either in Switzerland, Spain, UK, Norway, Belgium or Ireland (36 did not declare any affiliation).

Since for most of the questions the participants were allowed to select multiple responses and eventually to specify further responses on an extra field, if not otherwise stated we will always refer to the number and percentage of cases mentioned by our respondents.

3.1 Spoken resources and their scientific domains

One of the first questions of the survey (Q.1) asked the respondents to mention all of the scientific domains to which their resources belong to. Besides some possible options provided by the authors, the respondents had the possibility to report other domains in an optional field. The responses to this question have been grouped to form a sort of top list of domains as in Figure 1. The most mentioned domains in decreasing order are: *Oral History* (n = 53), *Phonetics & Phonology* (n = 35), *Dialectology* (n = 33), *Anthropology* (n = 31), *Ethnomusicology* (n = 8), *Language Acquisition* (n = 7), *Sociolinguistics* (n = 6), *Applied Linguistics* (n = 6), *Sociology* (n = 5), *Conversation Analysis* (n = 4), *Speech Technology* (n = 2).[3]

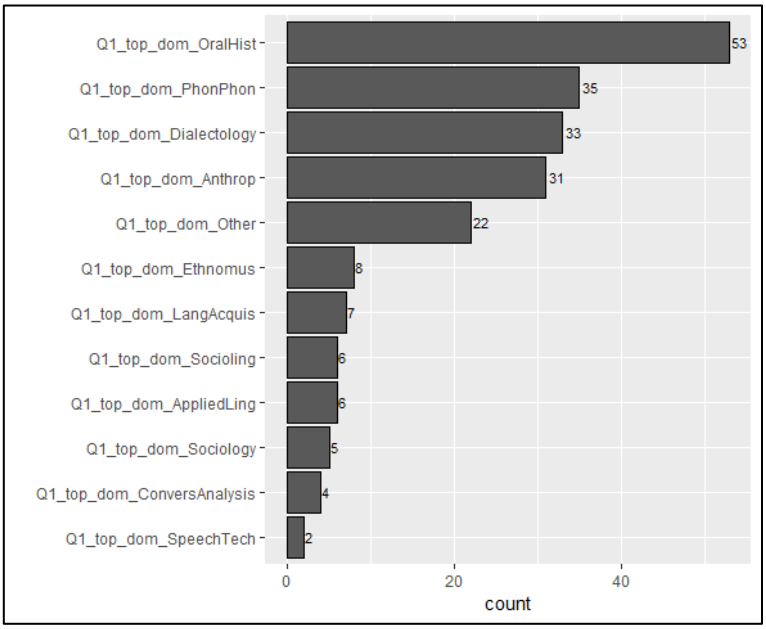


Figure 1. Scientific domains under which the respondents placed their resources.

After grouping the responses to this question into macro-areas[4], our initial intuition (e.g. that the huge amount of material collected by linguists during their fieldwork is neglected) stands out (see Figure 2). The majority of the participants we were able to reach indicated *Linguistics* (40.7%), *Oral History*

[2] The survey (available at https://goo.gl/8uHYK1) started on February 20th, 2018. Despite our initial intentions, the survey is still open and will be kept open until end 2019. This will allow the authors to continue the survey by reaching more respondents and eventually to include areas we might have neglected so far.

[3] The other *Linguistic* subfields mentioned in sparse order (e.g. prosody, syntax) and other hapax domains have been categorized as *Other* (n = 22).

[4] Due to the possibility the respondents had to fill in the "*Other, please specify*" option when indicating the scientific domains under which they considered their resources, the results on the disciplines were unavoidably scattered. To this end, following the *Linguistics* subfields grouping in the OLAC project (http://www.language-archives.org/REC/field.html), we recoded the responses to further reduce the sparseness of the data.

(30.8%) and *Anthropology* (18.0%) as core domains for their resources, with a minor portion of them indicating *Ethnomusicology* (4.7%), *Sociology* (2.9%), *Speech Technology* (1.2%), *Other* (1.7%).

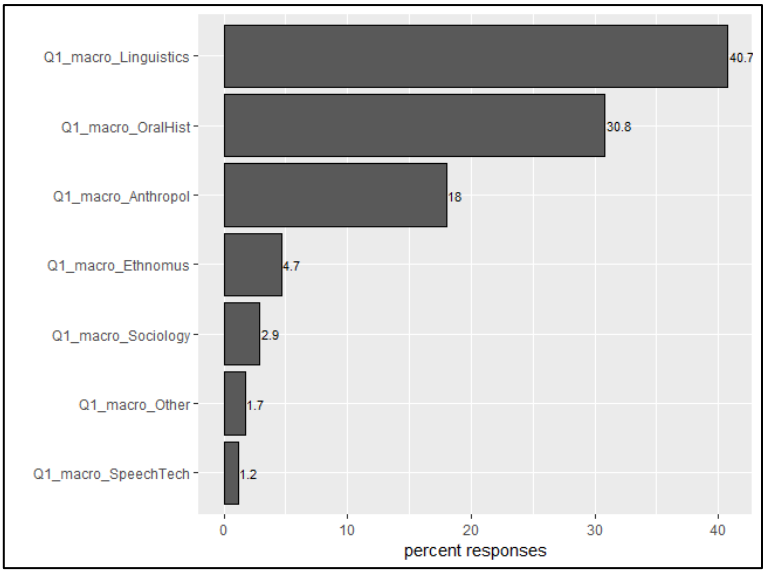


Figure 2. Macro-areas under which the respondents placed their resources.

3.2 Type of resources involved

When collecting speech in the different domains, the spoken productions can be recorded as uni-modal signals (e.g. Audio only) or as bi-modal signals (e.g. Audiovideo). This consideration led the authors to include this distinction in the survey (Q.2, see Figure 3). As few as 13% of the respondents selected Audiovideo only, while 40.5% of them declared having both Audio and Audiovideo resources. Those who declared having Audio only resources represent 46.6% of the cases.

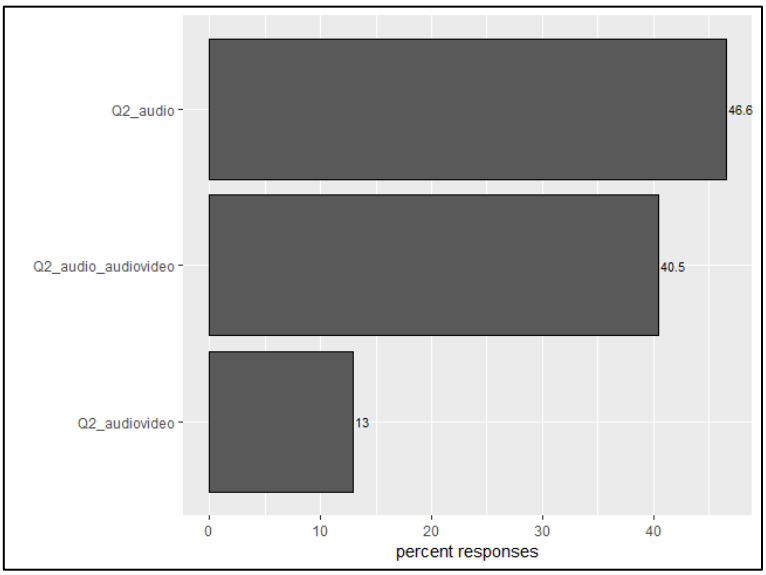


Figure 3. Type of resources owned by the respondents.

In the survey we further asked the respondents to indicate of what type of resources they were in possession of (Q.7), meaning here if these were of digital or analogue nature. As much as 70.5% of the resources were mentioned to be of digital nature (e.g. *.wav, *.Mp3, *.eaf, *.TextGrid, *.txt etc.), 26.7% of analogue nature (tapes, forms etc.), one respondent selected DAT (Digital Audio Tape) and other four did not answer the question. Additionally, for those who mentioned *Digital* in their answer, we asked to clarify the nature of those resources (Q.8). The options (and percent of responses in brackets) were:

a) born as original digital resources (54%);

b) the product of digitised analogue resources of which the respondents still own the originals (20%);
c) the product of digitised analogue resources of which they do no more own the originals (9%);
d) digitised "copies" of files owned by others (12%);
e) no answer (5%).

As our survey reveals, more than half the resources (54%) consists of original digital resources. It is undoubtedly a sign of the time: in the past, interviews "tended to be recorded on professional quality and somewhat bulky open-reel tape recorders" or, more recently, on "prosumer grade audio cassette recorders" (Cieri 2011: 31). Such venerable devices are now replaced by modern digital recording equipment which is nowadays within the reach of everyone.

Going more into detail, with Q.9 we also asked the respondents to tell us on what type of media the resources were stored: we gave them a list of predefined options with the possibility to select *everything that applies* and an additional "*Other*" option to specify other media types not listed. The responses are obviously scattered, but what is somehow surprising - if one looks at Figure 4 - is that only 28 of those who took part in the survey have their resources safely stored (*Server with back-up*).

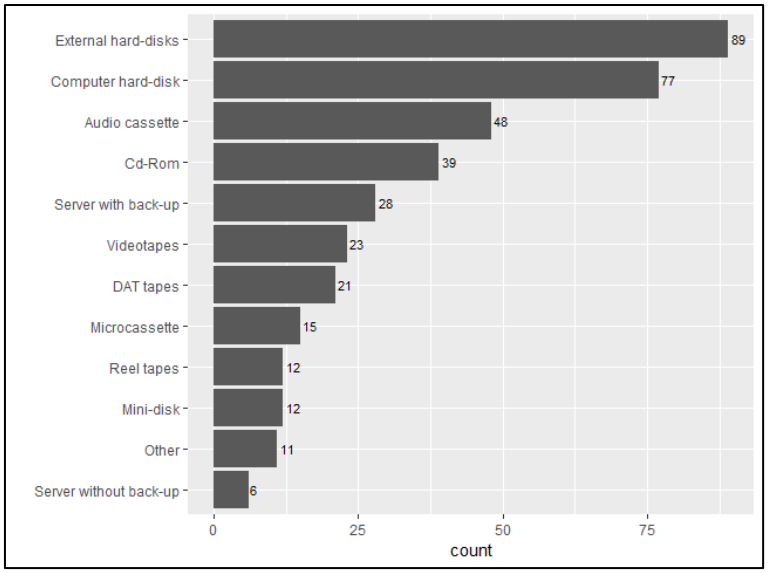


Figure 4. Media types on which the resources are stored.

The asked distinctions above are anything but trivial as they have direct consequences both on the quality and on the size (and eventually format) of the resources. In this respect, when it comes to quantifying the owned spoken resources (mainly in terms of hours) only half respondents (57%, 75 out of 131) are able to tell how much data they have in their archives (Q.6). Some of the respondents quantified their resources either in terms of number of interviews or number of files or number of tapes etc.; others were more precise and indicated an estimate in number of hours. Taking into account the responses gathered so far from the 75 respondents who were able to quantify their resources, and if we dare to do a very brutal conservative estimate of the amount of audio resources of those who were able to quantify it in terms of hours, the amount is impressing: more than 12 thousand hours of recorded material emerge from our survey, with more than 10 thousand hours mentioned (either exclusively or additionally) in the *Linguistics* domain.

The data above just provide a rough estimate giving us an idea of how much data is somehow remaining hidden to us. In addition to what was shown so far, we also asked to specify the language of their resources (Q.5). It might be of interest to know that 80.9% of the respondents answered having resources in *Italian language*, 42.6% listed *dialect varieties of the Italian peninsula* and 23.7% declared resources in *other languages*.

3.3 Type and format of additional data available for the targeted resources

Another question (Q.3) asked whether additional textual data in digital format (e.g. transcriptions, annotations etc.) are available besides the speech resources and, if there are, what type of format these data have (Q.4).

The great majority of the respondents (80.9%) declared additional data in digital format. As can be seen from Figure 5, considering all the mentioned types of files, the most common ones listed are *.doc (26.9%)[5], *.txt (24%), PRAAT's *.TextGrid (15.6%), ELAN's *.eaf files (7.2%), *.pdf (4.8%), Transcriber's *.tag files (1.2%) and of other sparse formats (4.8%).

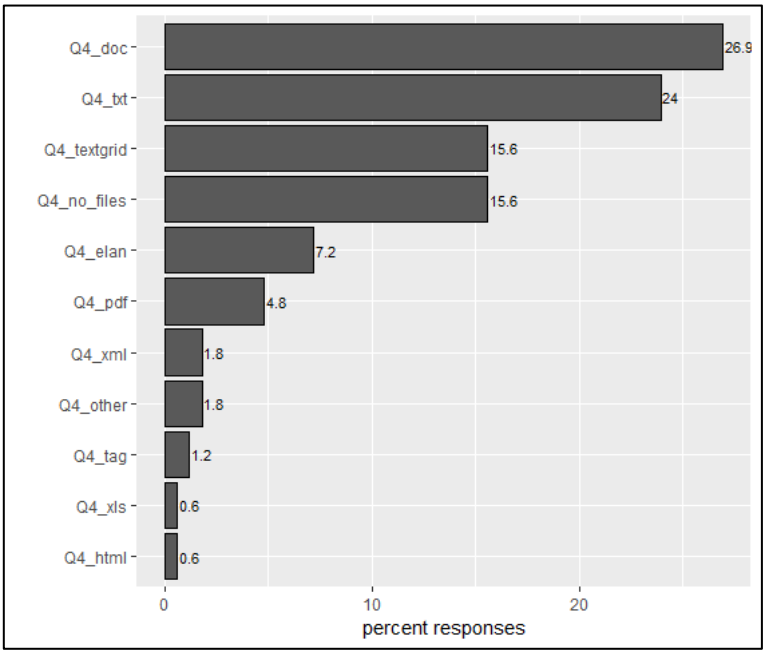


Figure 5. Declared additional textual data in digital format.

As far as the format of additional data available besides the speech resources is concerned, we categorized the above responses into two distinct groups: binary and non-binary files. This categorization was considered important also in order to verify if the information stored in those files is easily accessible and thus void of any restriction. Binary files are commonly application specific (e.g. proprietary) files. Due to the obsolescence of many applications, the use of binary files (as opposed to non-binary files, which allow unrestricted access and interoperability) has serious side-effects related to accessibility issues on the long term. Among the file formats, the respondents listed both *binary* (34%, including *.doc, *.pdf etc.) and *non-binary* formats (51%, including *.txt, PRAAT's *.TextGrid, ELAN's *.eaf, Transcriber's *.tag etc.); for 26 respondents (15%) no additional files are available.

3.4 Accessibility and maintenance issues

Accessibility of the data was addressed with Q.10 and it is not surprising to discover that almost half of the resources listed in our survey (49.6%) is barely accessible. Only 9.2% of the resources is accessible and available, 4.6% is partially accessible, 35.1% is available upon request, 1.5% is available upon request and only for selected parts. Moreover, we also wanted to know how one might have access to the resources they declared to be accessible and so we asked if specific access policies were available (Q.11). Only 9.2% of these resources is freely accessible online (with no authentication); 7.6% is accessible online via authentication; 29% is accessible onsite (i.e. where the resources are physically stored).

However, the two questions mentioned above open up to another important question: who is in charge of their long-term maintenance and preservation? For this reason we asked Q.12. Not surprisingly, the answer receiving the highest number of responses was *nobody* (43%), followed by *reference Institutes*[6] (17%), *reference Universities* (16%), the *owners/individuals* themselves (15%). Most surprisingly is the very small number of individuals (n = 5, i.e. 3.5%) who mentioned an *external repository* (NA's = 5.6%).

[5] For sake of economy, we grouped under the *.doc extension a series of other extensions such as *.docx, *.odt and *.rtf as well as all the generic responses (such as "word") referred to the popular word processor.
[6] Under this label we grouped institutions such as associations, foundations, libraries and their archives.

We would also like to stress the fact that the necessity of a national repository is of the highest urgency if we consider that most of those owning speech resources in our survey (about 47%) fall within the category which we defined as *casual workers* (e.g. workers without a permanent position or a permanent affiliation to an institution). Only 37% of the remaining respondents declared a *permanent* position and affiliation (for example to universities or other public institutions), while 9.2% did not provide any information (the remaining 7.6%, which we were not able to ascribe to any of the two categories, has been categorized as a generic "*other*" category).

3.5 Ethics and legal issues concerning oral resources

One further information emerging from our survey (Q.14) relates to ethics and legal issues, which are addressed by the respondents only in 46% of the cases.

Even if undoubtedly most part of the resources have been collected at a time when privacy and data protection issues were not addressed as nowadays, the effects on the accessibility and reusability of such resources are unavoidable. Not all the researchers are probably aware of new elements introduced by the recent General Data Protection Regulation (GPDR, EU n.2016/679), although certain scientific associations are providing information to their members, in order to support them in such a challenging issue (e.g. Italian Association of Oral History, http://aisoitalia.org/buone-pratiche-di-storia-orale-alcune-importanti-novita/). At the same time, the authority responsible for privacy in Italy has been organising several information meetings with Italian universities and public research bodies in order to raise awareness among the different research communities and university administrative staff on the changes introduced by the GDPR and their impact on research and dissemination activities.[7]

3.6 The CLARIN European infrastructure in our survey's scientific community

An unexpected result emerging from our survey at Q.18 is that only 31% of the respondents declared to have knowledge of the CLARIN infrastructure. This low percentage, however, should not discourage and diminish the activities carried out so far within the CLARIN infrastructure, on the contrary. There is indeed a large pool of resources owners (e.g. 64%, more than half of our respondents) who would agree to storing their archives and their speech resources in national repositories (Q.16). This manifestation of interest should give CLARIN's mission more strength and actuality.

4 Conclusion

In the past, researchers usually considered their speech data valuable only for the immediate purposes of their research. Nowadays, we are facing a change in attitude, since it is clear that legacy data document previous states of languages and linguistic changes from different points of view, and allow to work on historical questions about languages. Moreover, speech archives perfectly fit into the international debate concerning the use and reuse of past research data. Several scholars pointed out many important advantages of re-analysis: from sustainability to the maximization of the results. At the beginning of a novel research project, the re-analysis of past archives can be invaluable in providing a first orientation on the topics to be investigated, and therefore making the pilot stage of the research both more effective and swifter. By making previous research data available to re-analysis by others, it is possible to multiply the research outcomes through the publications of further interested scholars.

Nevertheless, the outcome of our survey shows a rather delicate picture: rather limited accessibility of the resources, ethical and legal issues only partially addressed, scant knowledge of the CLARIN infrastructure. In order to start filling the gap, the topic of the annual conference of the Italian Speech Sciences Association (AISV, *Associazione Italiana Scienze della Voce*) held in February 2019, was devoted to speech archives. The conference also saw the participation of the Executive Director of CLARIN who gave a keynote lecture exactly on *Spoken Word Archives as Societal and Cultural Data* (https://www.aisv.it/aisv2019/en/program).

[7] See for example https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/8318508 and https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/7977380 [date accessed: 26.01.2019].

Acknowledgements

The authors thank the anonymous reviewers for valuable comments on an earlier version of this paper, the organisations for their collaboration and all respondents for their participation in the survey.

References

Andreini A., Clemente P. (eds) 2007. *I custodi delle voci. Archivi orali in Toscana: primo censimento*, Firenze: Regione Toscana.

Barrera G. et al. 1993. *Fonti orali. Censimento degli istituti di conservazione*, Min. Beni Culturali e Ambientali.

Benedetti A. 2002. *Gli archivi sonori: fonoteche, nastroteche e biblioteche musicali in Italia*, Genova.

AA.VV 1999. *Archivi sonori. Atti dei seminari di Vercelli (22 gennaio 1993), Bologna (22-23 settembre 1994), Milano (7 marzo 1995)*, Roma, Min. Beni e le Attività Culturali-Ufficio centrale per i Beni archivistici, 1999.

Cappelli F., Rioda A. 2009. Archivi sonori in Toscana: un'indagine, Musica/Tecnologia, 3: 9-69.

Cieri C. 2011. Making a field recording. In *Sociophonetics. A student's guide*, M. Di Paolo, M. Yaeger-Dror (eds.), 24-35. Abingdon: Routledge.

Sergio G. (ed) 2016. *Atlante degli archivi fotografici e audiovisivi italiani digitalizzati*, Venezia: Fond. di Venezia-Marsilio.

Appendix

Questionnaire used in the survey (English translation)

Questionnaire - Section 2
All answers are optional. However, please try to answer as completely as possible, trying not to leave any questions.

0. Do you own any oral data / multimedia resources?
   Select only one.
   ○Yes
   ○ No (stop filling out the form and scroll down to the last question to terminate the survey)
   ○ No, but I know someone who owns some (scroll down and go to section 3)

1. Under which disciplinary area does your collection of oral data / multimedia resources fall?
   Select all that applies.
   □ Phonetics/phonology
   □ Dialectology
   □ Oral history
   □ Anthropology
   □ Sociology
   □ Psychology
   □ Applied linguistics
   □ Sociolinguistics
   □ Ethnomusicology
   □ Other, please specify: _____

2. Of what kind of resources is it about?
   Select all that applies.
   □ Audio
   □ Audiovideo

3. For those resources, do you own also textual files in digital format such as for example transcriptions, annotations etc.?
   Select only one.
   ○ Yes
   ○ No

4. If you answered “Yes” to the previous question, what format do these textual files have?
   Select all that applies.
   □ *.TextGrid (PRAAT)
   □ *.eaf (ELAN)
   □ *.tag (TranscriberAG)
   □ *.txt
   □ *.doc
   □ *.pdf
   □ Other, please specify: _____

5. In what language are these resources?
   Select all that applies.
   □ Italian
   □ Dialect varieties of the Italian peninsula (specify which in “Other, please specify “)
   □ Other languages (specify which in “Other, please specify”)
   □ Other, please specify: _____

6. Are you able to provide one or more estimates on the amount of multimedia resources you own?
   Please, if your answer is “Yes”, use the “Other” field and summarily describe the amount of such data as best you can. In reporting any numerical values, please also indicate the units of measurement to which you refer, for example in terms of gigabytes, hours, minutes, etc.
   Select all that applies.
   □ I am not able to quantify
   □ Yes
   □ Other, please specify: _____

7. The resources are in the following format:
   Select all that applies.
   □ Digital (*.wav, *.Mp3, *.eaf, *.TextGrid, *.txt etc.)
   □ Analogue
   □ DAT
   □ Other, please specify: _____

8. If you selected also “digital” in the previous question, can you please specify whether these resources are:
   Select all that applies.
   □ Born as original digital resources
   □ The result of digitised analogue resources of which I still own the originals
   □ The result of digitised analogue resources of which I no more own the originals
   □ Digital “copies” of resources owned by others
   □ Other, please specify: _____

9. Can you tell us on what type of media they are stored?
   Select all that applies.
   □ Audiotapes Microcassettes
   □ DAT tapes
   □ Video tapes
   □ Reel tapes
   □ Mini-disk
   □ Cd-Rom
   □ External hard-disk
   □ Computer hard-disk
   □ Server with back-up
   □ Server without back-up
   □ Other, please specify: _____

10. Are the data and the multimedia resources freely accessible?
Select only one.
□ Yes
□ Yes, but only upon request
□ No
□ Yes, but only partially
□ Other, please specify: _____

11. Are there specific access rules?
Select only one.
□ None (no access rules are available)
□ Online (with authentication)
□ Online (free access without authentication)
□ On-site
□ Other, please specify: _____

12. Who is in charge of the long-term maintenance and preservation of these resources?
Select all that applies.
□ Nobody
□ Reference University
□ External repository (specify which one in "Other")
□ Reference Institute
□ Owner
□ Other, please specify: _____

13. Where and how are your multimedia resources stored?
Please, provide as much detail as possible.
_____

14. Are the ethical and legal aspects of the data collection regulated (e.g., intellectual property, potential data reusability)?
Select only one.
○ Yes
○ No

15. If "Yes", how? Would you be willing to give us an example of a consent form normally used in your laboratory or in your research group (we will not divulge any form you will provide us)?
You can copy and paste the text of the form into the field below.
_____

16. Would you be interested in depositing also somewhere else your data?
Select only one.
○ Yes
○ No

17. If you answered "yes" to the previous question, what would you base your choice upon in deciding to deposit your resources on a potential repository?
For example, the presence of a graphical user interface, the presence of a dedicated structure offering support if needed, a free deposit service, the possibility to index the resources, etc.
_____

18.Do you know the CLARIN European infrastructure?
Select only one.
○ Yes
○ No

Human-human, human-machine communication: on the HuComTech multimodal corpus

L. Hunyadi
Department of General and Applied Linguistics, University of Debrecen, Debrecen, Hungary
hunyadi@undieb.hu

T. Váradi
MTA Institute of Linguistics, Research Group on Language Techonology Budapest, Hungary
varadi.tamas@nytud.mta.hu

Gy. Kovács
MTA SzTE Research Group on Artificial Intelligence, Szeged, Hungary, Embedded Internet Systems Lab, Luleå University of Technology, Luleå, Sweden
gykovacs@inf.u-szeged.hu

I. Szekrényes
Institute of Philosophy, University of Debrecen, Hungary
szekrenyes.istvan@arts.unideb.hu

H. Kiss
Department of General and Applied Linguistics, University of Debrecen, Debrecen, Hungary
kiss.hermina@arts.unideb.hu

K. Takács
Department of Phonetics, Eötvös Loránd University, Budapest, Hungary
karolin3813@gmail.com

Abstract

The present paper describes HuComTech, a multimodal corpus featuring over 50 hours of video taped interviews with 112 informants. The interviews were carried out in a lab equipped with multiple cameras and microphones able to record posture, hand gestures, facial expressions, gaze etc. as well as the acoustic and linguistic features of what was said. As a result of large-scale manual and semi-automatic annotation, the HuComTech corpus offers a rich dataset on 47 annotation levels. The paper presents the objectives, the workflow, the annotation work, focusing on two aspects in particular i.e. time alignment made with the Leipzig tool WEBMaus and the automatic detection of intonation contours developed by the HuComTech team. Early exploitation of the corpus included analysis of hidden patterns with the use of sophisticated multivariate analysis of temporal relations within the data points. The HuComTech corpus is one of the flagship language resources available through the HunCLARIN repository.

1 Introduction

In the age of the ubiquitous smart phones and other smart devices, robots and personal assistants, the issue of human-machine communication has acquired a new relevance and urgency. However, before communication with machine systems can become anything approaching the naturalness and robustness that humans expect, we must first understand human-human communication in its complexity. In order to rise to this challenge, we must break with the word-centric tradition of the study of communication and we must capture human-human communication in all the richness of the settings that it normally takes place. The foremost requirement for such an enterprise is richly annotated data, which is truly in

Laszlo Hunyadi, Tamás Váradi, István Szekrényes, György Kovács, Hermina Kiss and Karolina Takács 2019. Human-human, human-machine communication: on the HuComTech multimodal corpus. *Selected papers from the CLARIN Annual Conference 2018*. Linköping Electronic Conference Proceedings 159: 56–65.

short supply given the extremely labour intensive nature of the manifold annotation required. The ambition of the HuComTech project, which goes back to 2009, is to provide a rich language resource that can equally fuel application development as well as digital humanities research.

The HuComTech corpus is the first corpus of Hungarian dialogues that, based on multiple layers of annotation offers the so far most comprehensive information about general and individual properties of verbal and nonverbal communication. It aims at contributing to the discovery of patterns of behaviour characteristic of different settings, and at implementing these patterns in human-machine communication.

The paper will be structured as follows. In section 2 we will describe the data (the informants, the settings of the interviews, the size and main characteristics of the data set etc.) and will discuss the annotation principles and will provide a brief overview of the various levels of annotation. Section 3 discusses two automatic methods used in the annotation: forced alignment at the word level using the WEBMaus tool available through Clarin-DE as well as the automatic identification of intonation contours. Section 4 will preview some tentative exploration of the data, describing an approach that is designed to reveal hidden patterns in this complex data set through a sophisticated statistical analysis of the temporal distance between data points.

2 Description of the data and its annotation

2.1 General description of the corpus

The data for the HuComTech corpus was collected in face-to-face interviews that were conducted in a lab. The informants were university student volunteers. During the interviews informants were asked to read out 15 sentences, and were engaged in both formal and informal conversations, including a simulated job interview. The corpus consists of 112 interviews running to 50 hours of video recording containing about 450 000 tokens. Both the verbal and non-verbal aspects of the communication between field worker and informants were recorded through suitably positioned video cameras and external microphones.

The corpus offers a huge amount of time aligned data for the study of verbal and non-verbal behaviour by giving the chance to identify temporal patterns of behaviour both within and across subjects. The native format is .eaf to be used in ELAN (Wittenburg et al 2006), but a format for Theme (Magnusson, 2000), a statistical tool specifically designed for the discovery of hidden patterns of behaviour is also available for a more advanced approach of data analysis.

Through a database the data of the corpus will be made completely available for linguists, communication specialists, psychologists, language technologists.

A non-final version of the HuComTech corpus is already available online and it can be explored using *Trova* and *Annex (*Beck & Russel 2006) tools developed by the Max Planck Institute for Psycholinguistics within the framework of *The Language Archive* project. From there one can also download media and annotation files for academic research purposes.

2.2 The annotation protocol and the annotation scheme

The annotation followed the independent tagging for each of the more than 30 levels. It means that each level was annotated without any information about tags entered on another level. Each level of each file was annotated by two annotators independently, and a third annotator made possible corrections. The annotators formed groups in which they regularly discussed emerging issues, too. It assured a satisfactory inter-annotator agreement.

The annotation, comprised of about 1.5 million pieces of data ranges from the description of nonverbal, physical characteristics of 112 speakers (gaze, head-, hand-, body movements) to the pragmatic, functional description of these characteristics (such as turn management, cooperation, emotions etc.) The annotation of verbal behaviour includes the phonetics of speech (speech melody, intensity, tempo), morphology and syntax. The more than 450000 running words are time aligned enabling the association of the text with non-verbal features even on the word level.

A special feature of the annotation is that, whenever applicable, it was done both multimodally (using signals both from audio and video channels) and unimodally (using signals from either channel). Of course we subscribe to the view that both the production and the perception of a communicative event is inherently multimodal, yet the rationale for separating the two modalities was that the analysis and the generation of such an event by a machine agent needs to set the parameters of each of the modalities separately. Apart from this technical implementational perspective, we believe that by annotating some communicative/pragmatic functions both multimodally (using information from both video and the audio channel) and unimodally (relying on information from either the video or the audio) may pinpoint the primary source of the given function as either a single modality or a complex of several ones.

Accordingly, the annotation layers are organized into the following six annotation schemes in terms of the modalities involved: audio, morpho-syntactic, video, unimodal pragmatic, multimodal pragmatic and prosodic annotation.

The *audio annotation* is based on the audio signal using intonation phrases (head and subordination clauses) as segmentation units (Pápay et al 2011). The annotation covered verbal and non-verbal acoustic signals and included the following elements: transcription, fluency, intonation phrases, iteration, embeddings, emotions, turn management and discourse structure. The annotation was done manually using the Praat tool (Boersma &Weenink, 2016), validation was semi-automatic involving Praat scripts.

The *morpho-syntactic* annotation was done both manually and automatically, covering different aspects. Automatic annotation included tokenization, part of speech tagging and parsing (both constituent and dependency structure) The HMM-based toolkit *magyarlánc* (Zsibrita et al, 2013) developed at Szeged University was used for the automatic morpho-syntactic annotation. In addition, syntax is also annotated manually both for broader linguistic and for specific non-linguistic (especially psychology and communication) purposes (focusing on broader hierarchical relations and the identification of missing elements).

Video annotation included the following annotation elements: facial expression, gaze, eyebrows, head shift, hand shape, touch motion, posture, deixis, emblem, emotions. Annotation was done manually and, where possible, automatically using Qannot tool (Pápay et al, 2011) specially developed for the purpose.

Unimodal pragmatic annotation used a modified (single-modal) version of conversational analysis as its theoretical model and with the Qannot tool manually annotated the following elements: turn management, attention, agreement, deixis and information structure.

Multimodal pragmatic annotation used a modified (multimodal) version of Speech Act Theory and using both verbal and visual signals covered the following annotation elements: communicative acts, supporting acts, thematic control, information structure. The annotation was done manually with the Qannot tool.

Prosodic annotation (see Section 3 below) was prepared automatically using the Praat tool and covered the following elements: pitch, intensity, pauses and speech rate.

As the above detailed description of the annotation schemes reflects, a large part of the annotation was done manually. This was inevitable given the fact that the identification of

perceived emotions as well as a large number of communicative as well as pragmatic functions require interpretation, which are currently beyond the scope of automatic recognition, therefore they have to be determined and annotated manually.

2.3 Automatic annotation of prosody

In this section we describe a method developed for the automatic annotation of intonation, which, however, can be used not just for the HuComTech corpus, and therefore, we feel, deserves discussion in some detail. Our method does not follow the syllable-size units of Merten's Prosogram tool (Mertens, 2004) but an event can integrate a sequence of syllables in larger trends of modulation, which are classified in terms of dynamic, speaker-dependent thresholds (instead of *glissando)*. The algorithm was implemented as a Praat script. It requires no training material, only a two-level annotation of speaker change is assumed.

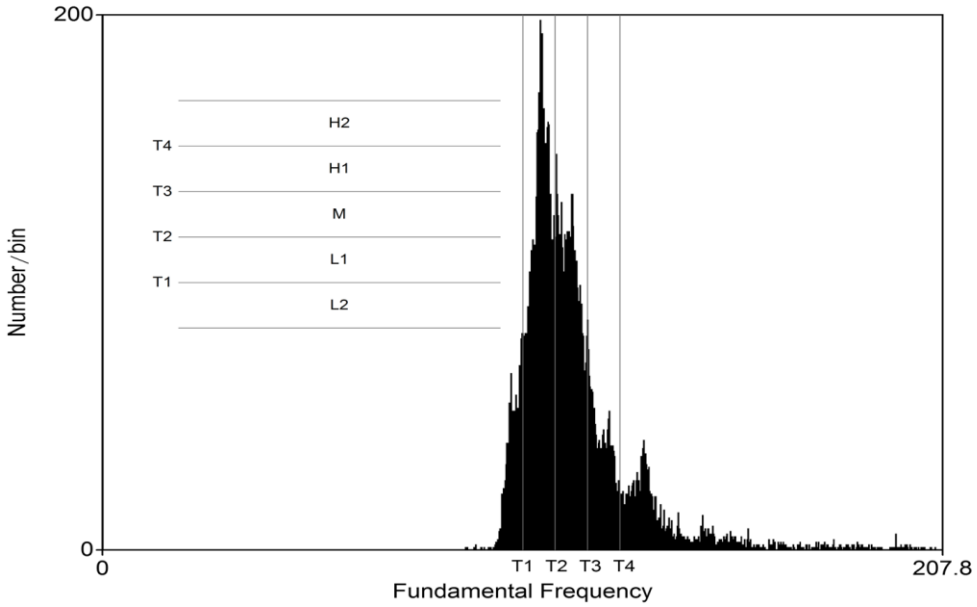


Figure 1: Calculating individual pitch ranges of the speaker based on the F0 distribution: L2 < L1 < M < H1 < H2.

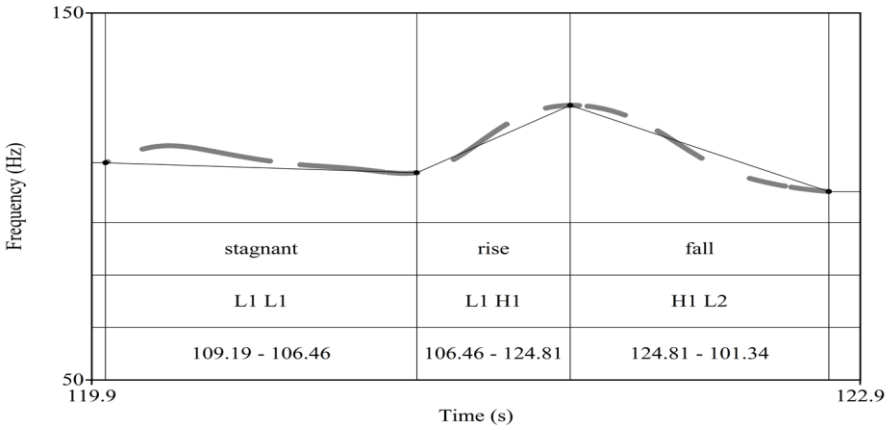


Figure 2: An output sample from the HuComTech Corpus

The output of the algorithm (Szekrényes 2015) contains larger, smoothed and stylized movements of the original data (F0 and intensity values) where the values indicate the contour (descending, falling, rising etc.), the absolute and relative vertical position of every single prosodic event through their starting and ending points. The "relative position" means that we located breaking points of intonation trends in the individual vocal range of the speaker which was divided into five levels based on the distribution of F0 values (see Figure 1).

The resulting labels representing modulations and positions of the prosodic structure can be considered as an automatically generated but perceptually verifiable music sheet of communication based on the raw F0 and intensity data. In Figure 2, one can see an output sample from the HuComTech corpus.

3 Exploring the corpus

We report two preliminary explorations of the HuComTech corpus. Experiments have been conducted with a view to modelling turn management through machine learning using neural networks. Second, through the use of a sophisticated statistical analysis tool we sought to explore hidden patterns within the complex multimodal data sets on the basis of temporal distance between them.

3.1 Modelling turn management: automatic detection of turn taking

The HuComTech corpus provides detailed data on turn management. For each discourse unit it contains manual annotation to indicate topic initiation, topic elaboration, topic change (and the absence of these categories, we will refer to as "no contribution"). Such comprehensive annotation invites experimentation for machine learning to automatically model turn management. Indeed, it is very important for a machine agent to be able to establish if the human interlocutor is keeping to the topic at hand or when they are veering away from it either by opening a completely different topic or slightly altering the course of the conversation. Conversely, it is also important for the machine agent to know when the human interlocutor is not doing any of the above (i.e. not contributing to the topic). Given that – depending on the situation (i.e. the machine agent speaking or not) – this can mean that the human is merely providing backchannel feedback (thus the agent does not have to relinquish the speaking turn) or that the human has finished its turn and the agent can take the speaking turn instead without the risk of interrupting or speaking over the human interlocutor.

Earlier studies on topic structure discovery relied mostly on lexical information (Holz and Teresniak, 2010), prosody (Zellers and Post, 2009), or a combination of the two (Shriberg et al., 2000). The HuComTech corpus, on the other hand, contains a more extensive annotation, facilitating the use of much wider sources of information as cues. This, among others, include such cues as gaze, facial expression, hand gestures, head movements, and so on. Despite the abundance of available information, however, the task is still challenging, and the experiments so far represent only tentative initial steps. One particular difficulty of topic classification is the class imbalance inherent to the task. In a conversation, one naturally spends more time with either not contributing to the topic (by providing only backchannel feedback, or not speaking at all) or elaborating the current topic than with changing the topic (either only slightly altering the course of the conversation, or completely veering away from the current topic).

The effect of this imbalance is twofold, as it can affect both the training and the evaluation of our models. For evaluation purposes the most common metric applied in classification tasks (particularly in multi-class classification) is the accuracy:

$$Accuracy = \frac{\sum_{i \in Classes} \; No.ofcorrectlyidentifiedinstances\text{in}Classi}{\sum_{i \in Classes} \; No.ofallinstances\text{in}Classi}$$

It is easy to see, however, how class imbalance can introduce a bias into this metric, favouring models that perform well on the majority class. In the HuComTech database, for example, the two majority classes (topic elaboration and no contribution) make up 82% of all instances to be classified, which means that a model correctly classifying these instances (but none of the instances from the two minority classes - topic change and topic initiation) can attain an accuracy score of 82%. An accuracy score of 82% may seem like a reasonable performance, but a model that cannot identify the change in topic at all is clearly ill-equipped for the task of topic unit classification. For this reason we suggested the use of a different metric for evaluating topic unit classification models, namely the Unweighted Average Recall (UAR):

$$UAR = \sum_{i \in Classes} \frac{No.ofcorrectlyidentifiedinstances\text{in}Classi}{No.ofallinstances\text{in}Classi \cdot No.ofclasses}$$

The benefit of using UAR (for further information, see Rosenberg, 2012) is that it assigns the same importance to each class, regardless of the cardinality of the classes. This means that a model performing well on the majority classes but bad on the minority classes would receive the same score as a model performing well on the minority classes, but bad on the majority classes.

Another problem class imbalance can cause is a bias towards the majority classes in the model trained. By training a neural network using several examples from certain classes, and relatively few examples of others, we may inadvertently train the network to disregard the minority classes. One technique that can be used to avoid this is that of downsampling, where we use the same amount of samples from each class during the training process. This, however, means that we disregard a large portion of our labeled samples. Another possibility would be to collect more samples from the minority classes. This, however, is both time-consuming, and costly, rendering this option infeasible in most applications. But it is also possible to "fool" the model by using the samples from the minority classes several times during the process of training. One technique that enables us to do so is that of probabilistic sampling (for further information, see Tóth and Kocsor, 2005), where a parameter is used to control the uniformity of class distribution during the training process. When value of the parameter is set to 0, the number of samples used from each class is equal to the cardinality of that class; when the value of the parameter is set to 1, however, the same number of samples are used of each class.

Kovács et al. (2016) built a topic unit classifier with the use of Deep Rectifier Neural Nets (Glorot et al, 2011), applying the technique of probabilistic sampling. We demonstrated in several experiments that this method attains a convincingly better performance than a support vector machine or a deep rectifier neural net by itself. For further information see (Kovács et al. 2016). In our tentative experiments we have found that the same holds true for other neural networks architectures – such as Long Short-Term Memory Unit (LSTM - Hochreiter and Schmidhuber, 1997) networks, and Gated Recurrent Unit (Cho et al., 2014) networks – as well. Our preliminary results show that the application of probabilistic sampling significantly increases the UAR scores attained in both of these models as well.

Given the rich annotation available for the dialogues in the HuComTech corpus, another promising direction of inquiry is to examine the rate of contribution different types of features had towards the identification of the correct topical unit label. For this we used five of the six categories of annotation described in Section 2.2 (morpho-syntactic annotation,

video annotation, unimodal annotation, multimodal annotation, prosodic annotation). First, we examined the performance attainable with Deep Neural Networks when using features from only one annotation category. We found that by using the features from multimodal annotation only (with the exception of the topic unit labels, which were used as targets), an UAR score can be attained on the task of topic unit classification that is competitive with those scores we attain when using all features. What is more, in most cases we got a better UAR score by using only the multimodal features than that we got by using all available features. In the next stage we employed a classifier combination method on the models trained on individual feature categories. Here, we took the weighted average of the posterior probability estimates provided by the five different models. We found using the proper combination of our five models, we can further improve the classification performance. What is more, we also found that we can attain the same performance using only two categories, that is multimodal and morpho-syntactic annotation. For further information, see (Kovács et al. 2017)

3.2 T-pattern analysis to discover hidden patterns of behaviour

Undoubtedly, the HuComTech corpus contains a bewildering number and complexity of annotation data. The possibility to use this rich database to explore possible interdependencies between data points recorded at numerous levels of annotation is an exciting prospect as well as a serious challenge.

The difficulty lies not simply in the number of data points to consider but rather, it is of a theoretical nature. The capturing of a given communicative function cannot usually be done by describing the temporal alignment of a number of predefined modalities and their exact linear sequences, since for the expression of most of the functions a given list of participating modalities includes optionalities for individual variation, and sequences are not necessarily based on strict adjacency relations. As a result, traditional statistical methods (including time series analysis) are practically not capable of capturing the behavioural patterns leading to functional interpretation.

We apply an approach of discovery on multivariate analysis of temporal relationships between any annotation elements within a given time window. T-pattern analysis (Magnusson, 2000) was developed for the discovery of hidden patterns of behaviour in social interactions using the software tool *Theme*. *Theme* is a unique software environment that is intended to override the usual challenges of behavioural research, namely, patterns are composed of events which do not necessarily follow one another in an immediate sequence, and also, these events may be optional in many cases. Accordingly, when searching for patterns for a given communicative function, this function needs to be identified even if its constituents are not adjacent or, in certain cases, an event stereotypical for the given function is not even present at all. The authors of this paper had the chance to participate in the further development of this software by exposing it to the very large and annotationally complex HuComTech corpus.

The T-Pattern analysis offers a framework to meet these serious challenges by simulating the cognitive process of human pattern recognition. The result is a set of patterns as possible expressions of a given function with their exact statistical significance. Moreover, it also suggests which of the constituting elements (events) of a given pattern can predict or retrodict the given function as a whole.

Without exceeding the limits of this paper let us have a few examples of results from Hunyadi (2017) showing how *Theme* can capture the above challenges and dynamic of multimodal communications, based on the HuComTech corpus:

Example 1: f055 (formal dialogue, female subject), ID975:

example for a pattern composed of events from syntax and gaze direction

(([1 (end of incomplete clause, end of coordination)][2(end of gaze forward, start of blink)])([3 (end of blink, start of blink)][4 (gaze down, end of blink)]))

Annotated as:
(([1 (miss,e,yes co,e,yes)][2(v_gaze,e,forwards v_gaze,b,blink)])([3 (v_gaze,e,blink v_gaze,b,blink)][4 (v_gaze,b,down v_gaze,e,blink)]))

Text:
1: *és úgy gondolom [and I think]*,T=A_speaker_text,B=41188,E=42643
2.3.1.0.0.0.4,6,9.,T=S_clauses,B=41188,E=42643

2: forwards,T=V_gazeClass,B=42655,E=43455
blink,T=V_gazeClass,B=43455,E=43855

3: blink,T=V_gazeClass,B=43455,E=43855
blink,T=V_gazeClass,B=44255,E=44655

4: down,T=V_gazeClass,B=44655,E=45055
blink,T=V_gazeClass,B=45055,E=45455

Example 2: f055 (formal dialogue, female subject), ID 508:

pattern of multimodal behaviour

([1 (speaker, end of new information, spekaer, end of topic elaboration)]([2 (agent, beginning of directive, agent, beginning of new information)][3 (agent, end of directive, agent, end of new information)]))

Annotated as:
([1 (mp_spinf,e,new mp_sptopic,e,t_elab)]([2 (mp_agcommact,b,directive mp_aginf,b,new)][3 (mp_agcommact,e,directive mp_aginf,e,new)]))

Preceeded by:

{b} %m *rendben* [all right].,T=S_text,B=23719,E=25109
mivel pályakezdő vagyok {since I am starting my career],,T=S_text,B=25109,E=26429
nem volt még előző {p} *munkahelyem* [I have not had a previous workplace],,T=S_text,B=26429,E=28569
%s *a tanulmányaim eléggé* %s %o %s *jól sikerültek* [I was fairly successful with my studies].,T=S_text,B=28569,E=33748
tehát úgy érdemjegyileg [as for marks] %o %s *mindenben* [and everything] %s {l} *meg vagyok elégedve vele* [I am satisfied with it].,T=S_text,B=34207,E=41188
és úgy gondolom [and I think],,T=S_text,B=41188,E=42643
hogy [that] e— %o %*ezt tudnám kamatoztatni a munkámban is* [I could benefit from it in my work].,T=S_text,B=42643,E=46574)

1: new,T=MP_speaker_Information,B=23695,E=45775
topic_elaboration,T=MP_speaker_Topic,B=23695,E=45775

2: directive,T=MP_agent_CommunicativeAct,B=46720,E=48320
new,T=MP_agent_Information,B=46720,E=48320

és milyen szakot végzett?[and what did you study?],T=A_agent_text,B=46574,E=48228

3: directive,T=MP_agent_CommunicativeAct,B=46720,E=48320
new,T=MP_agent_Information,B=46720,E=48320

Example 3: f060 (formal dialogue beszélgetés, female subject), ID 2030

a pattern of multimodal pragmatics and posture

((([1 agent, beginning of directive] [2 agent, topic initiation])][3 speaker, communicative act, multimodal: none)](([4 speaker, beginning of leaning right speaker, end of leaning right])[5 (speaker, end of leaning right speaker, beginning of leaning right)]))

Annotated as:
((([1 mp_agcommact,b,directive] [2 mp_agtopic,e,t_init])][3 mp_spcommact,b,none)](([4 v_post,b,right,lean v_post,e,right,lean])[5 (v_post,e,left,lean v_post,b,right,lean)]))

1: directive,T=MP_agent_CommunicativeAct,B=52800,E=55360
(the referenced text: {p} {b} %o {p} {b} **mért jelentkezett a felhívásra?* [why did you respond to the call?], T=A_agent_text, B=51358, E=55239)

2: topic_initiation,T=MP_agent_Topic,B=52800,E=53760
(part of the above the referenced text: **mért* [why])

3: none,T=MP_speaker_CommunicativeAct,B=55046,E=56326
(nonverbal backchannel)

4: lean-right,T=V_postureClass,B=57206,E=57606
(the referenced text: *mert* [because] %o *szeretnék munkába lépni* [I want to get the job],, T=S_text, B=56679, E=58614)

5: lean-left,T=V_postureClass,B=57606,E=58006
lean-right,T=V_postureClass,B=58006,E=58806

4 Conclusion

In this short article, we provided a brief overview of the multimodal HuComTech corpus. It is offered as a richly annotated language resource that can serve a number of purposes ranging from supporting application development in the area of human-machine to empirical based research leading to a better understanding of the complex interplay of numerous factors involved in human-human multimodal communication. The corpus is available through the HunCLARIN repository and is made public with the expectation that it will generate further research into multimodal communication.

References

[Boersma &Weenink, 2016] Boersma, D., Paul & Weenink. 2016. Praat : doing phonetics by computer [computer program]. version 6.0.22. http://www.praat.org/. (retrieved 15 November 2016)

[Beck & Russel, 2006] Berck, P. and Russel, A. 2006. ANNEX – a web-based Framework for Exploiting Annotated Media Resources. In: *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*. Genoa: European Language Resources Association, 2006.

[Cho et al., 2014] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio. On the properties of neural machine translation: Encoder-decoder approaches. arXiv preprint arXiv:1409.1259, 2014

[Glorot et al] Glorot, X., Bordes, A., Bengio, Y. 2011. Deep Sparse Rectifier Neural Networks. In: Gordon, G. J., Dunson, D., B. Dudík, M. (eds): *AISTATS JMLR Proceedings 15*. JMLR.org. 315-323.

[Hochreiter and Schmidhuber, 1997] S. Hochreiter and J. Schmidhuber. 1997. Long Short-Term Memory. Neural Comput. 9, 8 (November 1997), 1735-1780.

[Holz and Teresniak, 2010] Holz and Teresnai, 2010: F. Holz and S. Teresniak, "Towards automatic detection and tracking of topic change", in Proc. CICLing, 2010, pp. 327-339

[Hunyadi 2017] Hunyadi, L. 2017. A multimodális kommunikáció grammatikája felé: szekvenciális események rekurzív hierarchikus struktúrája. In: Bánréti, Z. (ed.) Általános Nyelvészeti Tanulmányok XXIX (2017), pp. 155-182.

[Kovacs et al] Kovács, G., Grósz, T., Váradi, T. 2016. Topical unit classification using deep neural nets and probabilistic sampling. In: *Proc. CogInfoCom*, (pp. 199–204)

[Kovács et al. 2017] Kovács, Gy., Váradi, T. 2017. A különböző modalitások hozzájárulásának vizsgálata a témairányítás eseteinek osztályozásához a HuComTech korpuszon, in: Vincze, Veronika (ed.) XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2017) Szeged, Szegedi Tudományegyetem Informatikai Tanszékcsoport, (2017) pp. 193-204. , 12 p.

[Magnusson, 2000] Magnusson, M. S. 2000. Discovering hidden time patterns in behavior: T-patterns and their detection behaviour research methods. *Behavior Research Methods, Instruments, & Computers*, 32:93–110.

[Mertens, 2004] Mertens, P. 2004. The prosogram : Semi-automatic transcription of prosody based on a tonal perception model. In *Proceedings of speech prosody*.

[Pápay et al, 2011] Pápay, K., Szeghalmy, S., and Szekrényes, I. 2011. Hucomtech multimodal corpus annotation. Argumentum 7:330–347.

[Rosenberg, 2012] A. Rosenberg, "Classifying skewed data: Importance weighting to optimize average recall" in Proc. Interspeech, 2012, pp. 2242-2245

[Shriberg et al., 2000] E. Shriberg, A. Stolcke. D. Hakkani-Tür, G. Tür, "Prosody-based automatic segmentation of speech into sentences and topics", Speech Commun. Vol 32, no. 1-2, pp 127-154, 2000

[Szekrényes 2014] Szekrényes, I. 2014. Annotation and interpretation of prosodic data in the hucomtech corpus for multimodal user interfaces. *Journal on Multimodal User Interfaces* 8:(2):143–150.

[Tóth and Kocsor, 2005] L. Tóth and A. Kocsor, "Training HMM/ANN" hybrid speech recognizers by probabilistic sampling

Wittenburg et al 2006] Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. 2006. Elan : a professional framework for multimodality research. In *Proceedings of LREC 2006* (pp. 213–269)

[Zellers and Post, 2009] M. Zellers, B. Post, "Fundamental frequency and other prosodic cues to topic structure", in Workshop on the Discourse-Prosody Interface, 2009. Pp. 377-386

[Zsibrita et al] Zsibrita, János; Vincze, Veronika; Farkas, Richárd 2013: magyarlanc: A Toolkit for Morphological and Dependency Parsing of Hungarian. In: *Proceedings of RANLP 2013*, pp. 763-771.

New exceptions for Text and Data Mining and their possible impact on the CLARIN infrastructure

Pawel Kamocki
ELDA / IDS Mannheim
pawel.kamocki@gmail.com

Erik Ketzan
Birkbeck, University of London
eketza01@mail.bbk.ac.uk

Julia Wildgans
IDS Mannheim / Universität Mannheim
j.wildgans@googlemail.com

Andreas Witt
IDS Mannheim / Universität Mannheim / Universität Heidelberg
witt@ids-mannheim.de

Abstract

The proposed paper discusses new exceptions for Text and Data Mining that have recently been adopted in some EU Member States, and probably will soon be adopted also at the EU level. These exceptions are of great significance for language scientists, as they exempt those who compile corpora from the obligation to obtain authorisation from rightholders. However, corpora compiled on the basis of such exceptions cannot be freely shared, which in a long run may have serious consequences for Open Science and the functioning of research infrastructures such as CLARIN ERIC.

1. Overview of the current system of statutory exceptions in European copyright

Copyright grants authors exclusive rights in relation to their works[1]. In principle, every reproduction[2] or communication to the public[3] of copyright-protected material requires authorisation from the rightholder[4]. Obviously, if applied strictly this could have a chilling effect on freedom of expression, art and research; this is particularly true in the digital environment, where every use of a work necessitates a reproduction (in the device's memory), while copying and worldwide sharing is cheap and instantaneous. In order to strike balance between the interests of rightholders and those of the public, legislators introduce statutory exceptions and limitations to exempt certain unauthorised uses from liability (exceptions) or to limit the scope of the rightholders' monopoly (limitations).

In the European Union, national legislators are not entirely free to adopt exceptions and limitations. Rather, the Directive 2001/29/EC of 22 May 2001 on the harmonisation of certain aspects of copyright and

[1] A work can be defined as an original creation in the literary (including computer programmes), artistic or scientific domain. The threshold of originality (*'author's own intellectual creation'*) is relatively easy to meet and one can say that, especially in the case of works of language, originality is *de facto* presumed

[2] The exclusive right of reproduction is construed broadly and includes 'direct or indirect, temporary or permanent reproduction by any means and in any form, in whole or in part' (art. 2 of the InfoSoc Directive)

[3] The exclusive right of communication to the public refers to *'any communication to the public (...), by wire or wireless means, including the making available to the public (...) in such a way that members of the public may access [the material] from a place and at a time individually chosen by them'*, (i.e. uploading on the Internet — art. 3 of the InfoSoc Directive)

[4] This authorisation is typically granted in an agreement called 'a licence' (Latin *licentio* — permission).

Pawel Kamocki, Erik Ketzan, Julia Wildgans and Andreas Witt 2019. New exceptions for Text and Data Mining and their possible impact on the CLARIN infrastructure. *Selected papers from the CLARIN Annual Conference 2018*. Linköping Electronic Conference Proceedings 159: 66–71.

related rights in the information society (hereinafter: InfoSoc Directive) contains (in its art. 5) a limitative[5] list of exceptions and limitations that can be adopted in the national laws of the Member States. Apart from one mandatory limitation (that enables the functioning of the Internet)[6], national legislators are free to choose which exception they want to adopt in their legal systems. National implementations of each of these exceptions can be narrower than allowed by the Directive, but they cannot be broader. Art. 5.3 (a) allows Member States to adopt exceptions for *use for the sole purpose of (...) scientific research, as long as the source, including the author's name, is indicated (...) and to the extent justified by the non-commercial purpose to be achieved.*

2. New exceptions for Text and Data Mining in certain EU Member States

Text and Data Mining (or text/data analytics) is the process of deriving new information from unstructured data by means of computational analysis. Since the analysed material is necessarily reproduced in the process (even if these reproductions may be just temporary), mining, in order to be lawful, requires authorisation from rightholders. The necessity to adopt statutory exceptions for Text and Data Mining, especially for research purposes, has been discussed at least since 2011, i.e. the publication of the Hargreaves review[7]. In 2013, a group on Text and Data Mining was created within the Stakeholder's Dialogue *Licenses for Europe*[8]. The academic community, unhappy with the adopted approach (focused on licensing rather than on statutory exceptions), largely withdrew from the process[9]. One of the key arguments in favour of a statutory TDM exception is the fact that TDM for research purposes is allowed under the *'fair use'* doctrine in the US, or covered by statutory exceptions e.g. Japan and other non-European countries. Meanwhile, some EU Member States decided to adopt TDM exceptions within the current legal framework (i.e. art. 5.3(a) of the Infosoc Directive, cf. *supra*).

In 2014, the UK was the first EU country to adopt a statutory TDM exception. Section 29A of the *Copyright, Designs and Patents Act* allows for making copies of works in order to *"carry out a computational analysis of anything recorded in the work for the sole purpose of research for a non-commercial purpose"*. Such copies need to be accompanied by a sufficient acknowledgement (unless this is practically or otherwise impossible) and cannot be transferred or used for any other purpose. The exception is expressly non-overridable by contracts (a contractual clause that purports to restrict the allowed activities is unenforceable)[10], but it only applies to those who have *'lawful access'* to a work. This latter requirement raises questions on whether this access should be expressly authorised (in a license), or simply not resulting from copyright infringement (in which case e.g. everyone with Internet access could mine openly available websites). There seems to be no clear answer to this question, even though, in our opinion, the second interpretation should prevail.

In 2016, France also introduced a TDM exception[11], but its scope remains very unclear. It seems to allow mining of scientific articles for the purposes of non-commercial public research (i.e. research carried out at universities and publicly funded research institutions). Adopted just before presidential and parliamentary elections, the French regulation on TDM is marked by its formal imperfections which an implementing

---

[5] Cf. recital 32 of the InfoSoc Directive: *'This Directive provides for an exhaustive enumeration of exceptions and limitations (...)'*

[6] Art. 5.1 of the InfoSoc Directive (so-called *'temporary acts of reproduction'*)

[7] Hargreaves, I. (2011). "Digital Opportunity. A Review of Intellectual Property and Growth", available at: https://www.gov.uk/government/publications/digital-opportunity-review-of-intellectual-property-and-growth (v. pp. 41-52, esp. p. 48)

[8] European Commission (2013). "Licenses For Europe: Structured stakeholder dialogue 2013", available at: https://ec.europa.eu/licences-for-europe-dialogue/

[9] LIBER (Association of European Research Libraries) (2013). "Stakeholders representing the research sector, SMEs and open access publishers withdraw from Licences for Europe", available at: https://libereurope.eu/blog/2013/05/24/stakeholders-representing-the-research-sector-smes-and-open-access-publishers-withdraw-from-licences-for-europe/

[10] Section 29A, sub-section 5, Copyright, Designs and Patents Act 1988

[11] Art. L. 122-5, 10° of the French Intellectual Property Code

decree was supposed to clarify; unfortunately, a proposal for such a decree was rejected in 2017[12] and, to the best of our knowledge, no progress has been made since. Therefore, it seems that the French TDM law is reduced to dead letter.

A much bolder measure was taken by the German legislator in 2017. New §60d of the German Copyright Act (UrhG) which entered into force on 1 March 2018[13] allows reproductions of copyright-protected content in order to enable automatic analysis of a large number of works for non-commercial scientific research. Furthermore, it also allows *necessary* modifications of mined content[14]. Interestingly, the new law expressly uses the word *corpus* to designate a collection of normalised, structured and categorised data created as part of the TDM process. Such a *corpus* can be shared with a *specifically limited circle of persons* (presumably a research team, also multi-institutional). However, once the research is over, the *corpus* has to be deleted or transferred to a specialised library or an archive for permanent storage[15]. The new German exception is expressly non-overridable by contractual clauses[16], which in practice means that all content openly available on the Internet can be freely mined, even if the terms of service prohibit such uses. On the other hand, the new law requires that flat-rate equitable remuneration be paid to a copyright collecting society for the allowed uses[17]. Moreover, the adopted solution may turn out to be temporary, as it has an 'expiration date': on 1 March 2023, the new rules will cease to apply. However, before that date the German legislator may decide to maintain them in force, or — more likely — adapt them to ensure compatibility with the upcoming EU Directive (cf. *infra*)

It shall also be noted that in some countries, such as Poland, the implementation of the research exception seems broad enough to encompass data mining activities (in Poland: only those carried out in public research institutions[18]). Other Member States, however, seem to lack a research exception exceeding private copying (e.g. Austria). This fragmentation is particularly troublesome from pan-European projects such as CLARIN. A greater degree of harmonisation, achievable only via an intervention at the EU level, seems urgent.

3. New exception for Text and Data Mining in the Digital Single Market Directive

In September 2016, the European Commission proposed a draft for a new Directive on copyright in the Digital Single Market[19]. Art. 3 of the draft proposes a mandatory (i. e. to be implemented in all the Member States) exception for reproductions and extractions *"made by research organisations in order to carry out text and data mining (...) for the purposes of scientific research"*. Only public universities and research institutions can benefit from this exception; however, the exception is no longer limited to non-commercial activities, so public-private partnerships are also within its scope. Like in the UK, the text requires *lawful access* to mined material, which raises the exact same questions as those discussed above.

The proposed exception is, like in the UK and in Germany, non-overridable by contracts. However, it allows rightholders to implement technological protection measures (Digital Rights Management) *"to ensure the security and integrity of the networks and databases"*. Such measures, however, *"shall not go beyond what is necessary to achieve this objective"*.

Many contrasting views on the proposal have been expressed during the discussions in the European Parliament. The Culture and Education Committee (CULT) advocated a solution similar to the one adopted in Germany, requiring payment of equitable remuneration and deletion of the compiled corpus upon the

---

[12] Langlais, P.-C. (2017). "L'exception Text & Data Mining sans décret d'application…", Sciences Communes, 10 May 2017, available at: https://scoms.hypotheses.org/category/data-mining

[13] Introduced by the *Urheberrechts-Wissensgesellschafts-Gesetz* (*UrhWissG*) of 7 September 2017

[14] §23 UrhG (also modified by UrhWissG)

[15] §60d(3) UrhG

[16] §60g UrhG

[17] §60h UrhG; the amount of the remuneration should be specified in an agreement concluded between the German states (Länder) and the relevant collecting society (for text data: VG Wort); to the best of our knowledge, no such agreement has been concluded as of yet, and it is quite impossible to predict its content

[18] Cf. art. 27 of the Polish Copyright Act

[19] European Commission (2016). "Proposal for a Directive of the European Parliament and of the Council on copyright in the Digital Single Market", available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52016PC0593

completion of the project. Its draft also stipulates that *lawful access* to mined works has to be *acquired*, which seems to indicate that a license to use the content (for whatever purpose) is necessary, and that content available on the open Internet is not necessarily concerned by the exception[20]. According to the Committee on the Internal Market and and Consumer Protection (IMCO), the beneficiaries of the exception shall not be limited to research organisations, and mining should be allowed also for other purposes than scientific research[21]. The Industry, Research and Energy Committee (ITRE) took a similar position[22]. Arguably the most important of the Committees, the Committee on Legal Affairs (JURI) expressed a more nuanced opinion. On the one hand, JURI advocates that the exception should concern all users and purposes; on the other hand, it also advocates for a narrow interpretation of *lawful access*. Research organisations, however, shall be allowed to mine databases of scientific publishers even if they do not meet the *lawful access* requirement. Furthermore, corpora mined for research purposes shall be stored securely in designated facilities and re-used only for the purposes of verification of results of the research[23].

On 25 May 2018, the European Council (under the Bulgarian presidency) published its version of the proposal[24]. As far as TDM exceptions are concerned, this version contains three important modifications compared to the Commission's original document. Firstly, the beneficiaries of the mandatory TDM exception include (alongside *research organisations*) also *cultural heritage institutions* (defined as publicly accessible libraries, museums and archives as well as film or audio heritage institutions). Secondly, the Council's version requires that the corpora used for TDM shall be stored *with an appropriate level of security* and not retained *for longer than necessary* (which may imply the necessity to delete them at the end of the research project, cf. *supra* about the solution adopted in Germany). Thirdly, and perhaps most importantly, the Council's proposal adds art. 3a containing an *optional* exception for TDM, allowing Member States to adopt broad TDM exceptions, potentially covering all categories of beneficiaries and purposes; however, these non-mandatory exceptions can only apply if the users have lawful access to the mined works, and if the use for TDM purposes has not been expressly restricted by rightholders (via Digital Rights Management or simply by an appropriate notice). This would change the paradigm from *"TDM only with permission"* to *"open for TDM by default"*, but would not really provide users with means to mine content which its rightholder does not want to be mined.

The final report of the European Parliament's Committee on Legal Affairs, adopted on 29 June 2018[25] was partly inspired by the Council's proposal. JURI advocates that the beneficiaries of the TDM exception shall include research institutions, but also educational establishments and cultural heritage institutions, to the extent that they conduct scientific research the results of which are publicly accessible. Secondly, JURI also added an optional TDM exception, similar to the one proposed by the Council.

---

[20] CULT (2017). Draft Opinion for the Committee on Legal Affairs on the proposal for a directive of the on copyright in the Digital Single Market, available at: http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-%2F%2FEP%2F%2FNONSGML%2BCOMPARL%2BPE-595.591%2B01%2BDOC%2BPDF%2BV0%2F%2FEN

[21] IMCO (2017). Draft Opinion for the Committee on Legal Affairs on the proposal for a directive on copyright in the Digital Single Market: http://www.europarl.europa.eu/sides/getDoc.do?type=COMPARL&reference=PE-599.682&format=PDF&language=EN&secondRef=01

[22] ITRE (2017). Draft Opinion for the Committee on Legal Affairs on the proposal for a directive on copyright in the Digital Single Market: http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-%2F%2FEP%2F%2FNONSGML%2BCOMPARL%2BPE-592.363%2B01%2BDOC%2BPDF%2BV0%2F%2FEN

[23] JURI (2017). I Draft Report on the proposal for a directive on copyright in the Digital Single Market: http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-%2F%2FEP%2F%2FNONSGML%2BCOMPARL%2BPE-601.094%2B01%2BDOC%2BPDF%2BV0%2F%2FEN

[24] European Council (2018). Notice from Presidency to Delegations on the Proposal for a Directive of the European Commission and the Council on copyright in the Digital Single Market, 2016/0280 (COD), available at: http://www.consilium.europa.eu/media/35373/st09134-en18.pdf

[25] JURI (2018). I Report Plenary sitting on the proposal for a directive of the European Parliament and of the Council on copyright in the Digital Single Market (COM(2016)0593 – C8-0383/2016 – 2016/0280(COD)): http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-%2f%2fEP%2f%2fNONSGML%2bREPORT%2bA8-2018-0245%2b0%2bDOC%2bPDF%2bV0%2f%2fEN

JURI's final report was rejected by the European Parliament during a plenary vote on 5 July 2018 (mostly because of other controversial provisions of the Directive), but approved its slightly modified version in the second vote on 12 September 2018. The final stage of the adoption process: three-party negotiations (trilogue) could officially start; however, it did not run very smoothly. The compromise text presented by Romanian presidency as 11 countries (including e.g. Germany, the Netherlands, Italy, Finland, Poland, Portugal and Slovenia) rejected the proposal, and the final vote (initially scheduled for 21 January 2019) was postponed. This was mostly due to the controversies concerning other articles of the proposed Directive (especially 11 and 13), and not the TDM exceptions.

Somewhat unexpectedly, the trilogue reached compromise on 13 February 2019[26]. The text was then debated at JURI and presented for a plenary vote by the European Parliament. On 26 March 2019, the Parliament adopted the Directive (with 348 MEPs votes for, 274 votes against and 36 abstentions)[27]. At the moment (as of 1 April 2019), it still has to be approved by the Council before it can enter into force, but this is usually a formality.

The TDM exceptions in the adopted text are similar to those proposed by the Council and approved by the Parliament in 2018. The mandatory exception (in article 3) benefits only (public) research organisations and cultural heritage institutions, and it is limited to research purposes (including commercial research). What has changed, however, is that the copies (which still have to be stored *with appropriate level of security*) may be retained for research purposes (so, unlike in the previous versions and in the German exception, they do not have to be deleted upon the completion of the project). Like in the original proposal, rightholders may use Digital Rights Management *"to ensure the security and integrity of the networks and databases"*, but without going beyond what is necessary to achieve this objective. The exception is not overridable by contracts.

The newly added (and renumbered) article 4 contains an optional exception with potentially unlimited beneficiaries and scope of purposes, the only limitation being that this exception can only apply to the content for which the rightholders have not expressly reserved the right to mine (so, potentially everything can become '*mineable by default*'). This leaves a lot of leeway to Member States in allowing TDM for other purposes than research, and to other actors than public research organisations and cultural heritage institutions. However, these optional exceptions, unlike the mandatory one, will probably be overridable by contracts.

The Directive will have to be implemented within two years of its entry into force (article 24). However, the transposition process may not run very smooth (because of the aforementioned controversial provisions unrelated to TDM) and may be significantly delayed.

4. The possible impact of the new exceptions on CLARIN infrastructure

Language researchers will receive substantial benefits and some legal certainty from the new TDM exceptions. However, even if certain research activities are exempted from the rules of copyright, proper licensing is still necessary to efficiently and widely share the fruits of researchers' work. In this sense, paradoxically, the new exception can have negative consequences on infrastructures such as CLARIN ERIC. In a world where intellectual property rights are *prima facie* no longer a barrier to access content and conduct (*in-house*) research, researchers have fewer incentives to care about proper licensing and sharing their datasets and results (e.g. within research infrastructures)[28]. This may in turn considerably reduce the *knowledge commons* (i.e. immaterial resources that — due to proper licensing — can be freely accessed and re-used by anyone and for any purpose[29]) and in a long run hamper the development of Open Science. In such circumstances, even if research activities freed from the requirement to obtain permission from rightholders can flourish, knowledge transfer, citizen science and user innovation[30] may paradoxically become more difficult, as they require sharing of data between various groups of stakeholders. In order to

---

[26] http://www.europarl.europa.eu/news/en/press-room/20190212IPR26152/agreement-reached-on-digital-copyright-rules

[27] http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//NONSGML+TA+P8-TA-2019-0232+0+DOC+PDF+V0//EN

[28] On incentives for Open Access in the academic community, see esp. Suber, P. (2012). Open Access, MIT Press

[29] Hess, Ch. and E. Ostrom (2006). Understanding Knowledge as a Commons, MIT Press

[30] Von Hippel, E. (2017). Free Innovation, MIT Press

avoid this, it is important to remember that even if certain research activities are exempted from the rules of copyright, proper licensing is still necessary to efficiently and widely share the fruits of researchers' work.

An alternative incentive (other than removing access barriers to primary material) for contributing to knowledge commons shall perhaps be provided by policymakers and research funding agencies. CLARIN ERIC, who declared its dedication to the principles of Open Science, has an important role to play in guaranteeing that language science remains truly open not only for researchers, but for all citizens.

References

Hargreaves, I. (2011). "Digital Opportunity. A Review of Intellectual Property and Growth", available at: https://www.gov.uk/government/publications/digital-opportunity-review-of-intellectual-property-and-growth

European Commission (2013). "Licenses For Europe: Structured stakeholder dialogue 2013", available at: https://ec.europa.eu/licences-for-europe-dialogue/

LIBER (Association of European Research Libraries) (2013). "Stakeholders representing the research sector, SMEs and open access publishers withdraw from Licences for Europe", available at: https://libereurope.eu/blog/2013/05/24/stakeholders-representing-the-research-sector-smes-and-open-access-publishers-withdraw-from-licences-for-europe/

Langlais, P.-C. (2017). "L'exception Text & Data Mining sans décret d'application…", Sciences Communes, 10 May 2017, available at: https://scoms.hypotheses.org/category/data-mining

European Commission (2016). "Proposal for a Directive of the European Parliament and of the Council on copyright in the Digital Single Market", available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52016PC0593

European Council (2018). Notice from Presidency to Delegations on the Proposal for a Directive of the European Commission and the Council on copyright in the Digital Single Market, 2016/0280 (COD), available at: http://www.consilium.europa.eu/media/35373/st09134-en18.pdf.

CULT (2017). Draft Opinion for the Committee on Legal Affairs on the proposal for a directive of the on copyright in the Digital Single Market: http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-%2F%2FEP%2F%2FNONSGML%2BCOMPARL%2BPE-595.591%2B01%2BDOC%2BPDF%2BV0%2F%2FEN

IMCO (2017). Draft Opinion for the Committee on Legal Affairs on the proposal for a directive on copyright in the Digital Single Market: http://www.europarl.europa.eu/sides/getDoc.do?type=COMPARL&reference=PE-599.682&format=PDF&language=EN&secondRef=01

ITRE (2017). Draft Opinion for the Committee on Legal Affairs on the proposal for a directive on copyright in the Digital Single Market: http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-%2F%2FEP%2F%2FNONSGML%2BCOMPARL%2BPE-592.363%2B01%2BDOC%2BPDF%2BV0%2F%2FEN

JURI (2017). I Draft Report on the proposal for a directive on copyright in the Digital Single Market: http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-%2F%2FEP%2F%2FNONSGML%2BCOMPARL%2BPE-601.094%2B01%2BDOC%2BPDF%2BV0%2F%2FEN

JURI (2018). I Report Plenary sitting on the proposal for a directive of the European Parliament and of the Council on copyright in the Digital Single Market (COM(2016)0593 – C8-0383/2016 – 2016/0280(COD)): http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-%2f%2fEP%2f%2fNONSGML%2bREPORT%2bA8-2018-0245%2b0%2bDOC%2bPDF%2bV0%2f%2fEN

Suber, P. (2012). Open Access, MIT Press.

Hess, Ch. and E. Ostrom (2006). Understanding Knowledge as a Commons, MIT Press.

Von Hippel, E. (2017). Free Innovation, MIT Press.

Processing personal data without the consent of the data subject for the development and use of language resources

Aleksei Kelli
University of Tartu
Estonia
aleksei.kelli@ut.ee

Krister Lindén
University of Helsinki
Finland
krister.linden@helsinki.fi

Kadri Vider
University of Tartu
Estonia
kadri.vider@ut.ee

Pawel Kamocki
ELDA, France /
IDS Mannheim,
Germany
pawel.kamocki@gmail.com

Ramūnas Birštonas
Vilnius University
Lithuania
ramunas.birstonas@tf.vu.lt

Silvia Calamai
University of Siena
Italy
silvia.calamai@unisi.it

Penny Labropoulou
ILSP/ARC, Greece
penny@ilsp.gr

Maria Gavrilidou
ILSP/ARC, Greece
maria@ilsp.gr

Pavel Straňák
Charles University, Czechia
stranak@ufal.mff.cuni.cz

Abstract

The development and use of language resources often involve the processing of personal data. The General Data Protection Regulation (GDPR) establishes an EU-wide framework for the processing of personal data for research purposes while at the same time allowing for some flexibility on the part of the Member States. The paper discusses the legal framework for language research following the entry into force of the GDPR. In the first section, we present some fundamental concepts of data protection relevant to language research. In the second section, the general framework of processing personal data for research purposes is discussed. In the last section, we focus on the models that certain EU Member States use to regulate data processing for research purposes.

1 Introduction[1]

Language resources (LRs) contain material subject to various legal regimes. For instance, they may contain copyright protected works, objects of related rights (performances) and personal data. This affects the way language resources are collected and used. Intellectual property issues relating to language resources have been previously addressed (see Kelli et al. 2015). The focus of this article is on personal data protection. More precisely on the processing of personal data for research purposes without the data subject's consent within the framework of language research. Personal data issues are relevant for language resources, given that they potentially contain oral speech or written text which relates to a natural

Aleksei Kelli, Krister Lindén, Kadri Vider, Pawel Kamocki, Ramūnas Birštonas, Silvia Calamai, Penny Labropoulou, Maria Gavrilidou, and Pavel Straňák 2019. Processing personal data without the consent of the data subject for the development and use of language resources. *Selected papers from the CLARIN Annual Conference 2018*. Linköping Electronic Conference Proceedings 159: 72–82.

person.[2] In the CLARIN Virtual Language Observatory (VLO), approximately 95,502 language resources[3] could contain personal data.[4]

Although the General Data Protection Regulation[5] (GDPR) provides a general framework for personal data protection, it leaves a certain degree of freedom for the EU Member States to regulate the processing of personal data in different contexts (including research, see GDPR Art. 89 (3)). Even the duration of personal data protection is up to the Member States.[6] For instance, according to the Estonian Personal Data Protection Act, the data subject's rights are protected during the lifetime of the data subject and for ten years after the death of the data subject. In the case of minors, the duration is the lifetime and twenty years (§ 9).[7] This means that the Member States can adopt different regulatory models.

This article preliminarily maps the regulatory framework for processing personal data for research purposes. It also provides insights into different national models.[8] The picture is further complicated by the fact that, in addition to the GDPR and national laws directly related to data protection, other national legislation may add regulations to data protection and privacy in particular contexts, e.g. health care. Before concentrating on the data processing for research purposes, key concepts of the data protection framework are addressed.

2 Data subject, personal data and data processing

The data subject is defined through the concept of personal data. Personal data is "any information relating to an identified or identifiable natural person ('data subject')" (GDPR Art. 4). Publicly available personal data is also protectable (C-73/07). According to the Article 29 Working Party[9] (WP29), information contained in free text in an electronic document may qualify as personal data. It does not have to be in a structured database (2007: 8).

The identifiability is a crucial issue since data not relating to a natural person (incl. anonymous data) is not subject to the GDPR requirements (See GDPR Recital 26). One option to avoid problems with personal data protection is the anonymisation of data used for language research. However, it should be kept in mind that the process of rendering personal data anonymous is an instance of further processing which has to follow the data protection requirements (WP29 2014a: 3). It is also slightly complicated as combining already anonymised data sources may again make their data personal, and in some cases, anonymisation may render the data useless for research purposes. For other protective measures, see Section 3.2 below.

A natural person can be identified by reference to an identifier (e.g., name, identification number), location data and physiological, genetic, mental, economic, cultural or social information (GDPR Art. 4). According to WP29 sound and image data qualify as personal data insofar as they may represent information on an individual (WP29 2007: 7). It means that LRs containing oral speech are subject to the GDPR. A question can be raised whether speech and voice as such constitute personal data where there is no additional information leading to a specific individual. It is a question related to identifiability. As suggested in the literature, data that are not identifiable for one person may be identifiable for another. Data can also become identifiable through combination with other data sets. Identifiability is a broad category depending on how much effort must be deemed 'reasonable' (Oostveen 2016: 306).

[2] For instance, according to the Court of Justice of the European Union (CJEU) the concept of personal data covers the name of a person (C-101/01).

[3] Resource type: Audio, Radio, Sound, Speech, Spontaneous, Television or Video.

[4] Language resources with written text may also contain personal data, but this is not as prominent as in the case of audio and/or visual material (e.g. interviews or photos of a certain person).

[5] The GDPR is applicable in all EU Member States from 25 May 2018. It replaces the Data Protection Directive.

[6] The GDPR does not apply to the personal data of deceased persons. Member States may establish the relevant regulation (GDPR Recital 27).

[7] The duration of personal data protection is rather complicated issue since the deceased person's data may still refer to a living person (WP29 2007: 22).

[8] For lack of space not all the EU countries are addressed in the present paper.

[9] According to the Data Protection Directive the Working Party on the Protection of Individuals with regard to the Processing of Personal Data (WP29) is composed of a representative of the supervisory authority or authorities designated by each Member State and of a representative of the authority or authorities established for the Community institutions and bodies, and of a representative of the Commission.

Voice can be considered biometric data (see González-Rodríguez et al. 2008; Jain et al. 2004).[10] Biometric data for uniquely identifying a natural person belongs to a special category of personal data[11] the processing of which is even more restricted than for other personal data. A similar case is that of photos depicting people. Here the GDPR provides a clarification: "The processing of photographs should not systematically be considered to be processing of special categories of personal data as they are covered by the definition of biometric data only when processed through a specific technical means allowing the unique identification or authentication of a natural person" (Recital 51). This should be applicable in case of speech and video as well. Therefore, the requirements concerning the processing of special categories of personal data apply in case oral speech contained in language resources is used for the identification of natural persons.

The GDPR defines processing very broadly. It includes, among other things, collection, structuring, storage, adaptation, use, making available or destruction (GDPR Art. 4). It means that the development and use of LRs containing personal data constitutes processing.

Personal data protection requirements do not have to be followed in case the processing of personal data is done by a natural person in the course of a purely personal or household activity (GDPR Art. 2 (2)). It is debatable if the private use exemption is applicable for research as well.

3 Processing personal data for research purposes

3.1 General framework

The General Data Protection Regulation sets forth the following principles relating to processing of personal data (incl. for research purposes): 1) lawfulness, fairness and transparency; 2) purpose limitation (data is collected for specified, explicit and legitimate purposes); 3) data minimisation (the collection and use of data is as limited as possible); 4) accuracy; 5) storage limitation (kept for no longer than is necessary); 6) integrity and confidentiality; 7) accountability (Art. 5). It is explained that further processing for research is compatible with the initial purposes. Personal data can be stored for more extended periods for research purposes (GDPR Art. 5).

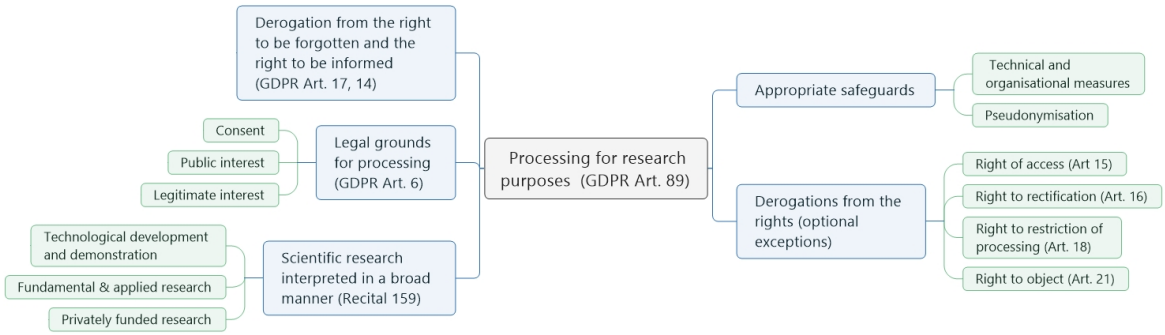


Figure 1: Processing of personal data for research purposes.

The GDPR provides six legal grounds for processing personal data: 1) consent; 2) performance of a contract; 3) compliance with a legal obligation; 4) protection of the vital interests; 5) the public interest or in the exercise of official authority; 6) legitimate interests (Art. 6).

As seen, the processing for research purposes is not an individual legal ground. Therefore, the processing for research purposes has to take place within the existing six grounds. The processing can rely on consent (for further discussion on consent see WP29 2017), the performance of a task carried out in the public interest or the legitimate interests.

[10] The GDPR defines biometric data as "personal data resulting from specific technical processing relating to the physical, physiological or behavioural characteristics of a natural person, which allow or confirm the unique identification of that natural person, such as facial images or dactyloscopic data" (Art. 4).

[11] The GDPR defines special categories of personal data as "data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation".

It is not entirely clear when the processing for research purposes must rely on consent and when the public interest and legitimate interest can be used as grounds. Note, however, that consent is needed at least if the aim is to make personal data publicly available because public or legitimate interest require protective measures limiting access. Consent may also be mandated by national legislation in particular cases, e.g. when collecting health data.

It can be presumed that the processing based on the data subject's consent provides the highest protection of his/her fundamental rights (privacy, integrity, self-realisation). The data subject may even withdraw his/her consent without any legal consequences (GDPR Art. 7 (3)). The controller[12] has to be able to prove the existence of the consent (GDPR Art. 7 (1)). WP29 explains that consent "focuses on the self-determination of the data subject as a ground for legitimacy. All other grounds, in contrast, allow processing – subject to safeguards and measures – in situations where, irrespective of consent, it is appropriate and necessary to process the data within a certain context in pursuit of a specific legitimate interest" (2014: 13).

In case where the acquisition of consent is very complicated or administratively burdensome (e.g., anonymous web posts, legacy resources, public videos and so forth) the question arises which legal ground is relevant. According to WP29, the performance of a task carried out in the public interest is another ground for processing personal data in the research context (2014b: 21-23). The concept of research in the public interest[13] can usually be invoked by research projects affiliated with universities or research institutions having a legal mandate to do research in the public interest[14], i.e. agencies acting on behalf of a Member State.

The GDPR also names the legitimate interests as a legal ground for processing. The concept of legitimate interest is rather complicated and requires weighing different interests.[15] According to WP29, legitimate interest can serve as a legal ground for processing personal data in the research context (2014b: 24-25). The legitimate interest is most likely relevant for commercial research.

Before addressing specific requirements concerning the processing of personal data for research, it is necessary to outline the concept of research in the data protection context. The GDPR defines research broadly so that it covers "technological development and demonstration, fundamental research, applied research and privately funded research" (Recital 159).

The GDPR provides the following requirements for processing data for research purposes (Art. 89):

1. processing for research purposes is subject to appropriate safeguards. The safeguards ensure that technical and organisational measures are in place in particular to ensure respect for the principle of data minimisation. Those measures may include pseudonymisation provided that those purposes can be fulfilled in that manner;
2. the Member States may limit the following data subject's rights for research purposes (optional exceptions):
    a) the right of access by the data subject (Art. 15);
    b) the right to rectification (Art. 16);
    c) the right to the restriction of processing (Art. 18);
    d) the right to object (Art. 21);

[12] The GDPR defines the controller as "the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data" (Art. 4 (7)).

[13] According to the GDPR, processing is lawful if it is "processing is necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller" (Art. 6e).

[14] For instance, according to the Estonian Organisation of Research and Development Act (ORDA) a research and development institution is a legal person or an institution in the case of which the principal activity is carrying out basic research, applied research or development, or several of the aforementioned activities (§ 3 (1) clause 1).

[15] According to the GDPR, processing is lawful if it is "necessary for the purposes of the legitimate interests pursued by the controller or by a third party, except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject which require protection of personal data" (Art. 6f).

There is also a mandatory exception[16] concerning the right to be forgotten[17] and right to be informed about the processing:

1. the right to be forgotten is limited to the extent that processing is necessary for research purposes in so far as the right to be forgotten is likely to render impossible or seriously impair the achievement of the objectives of that processing (GDPR Art. 17 (3)d);
2. the right to be informed about the processing of personal data is limited insofar as the provision of such information proves impossible or would involve a disproportionate effort, in particular for processing for research purposes and it is likely to render impossible or seriously impair the achievement of the objectives of that processing. In such cases, the controller shall take appropriate measures to protect the data subject's rights and freedoms and legitimate interests, including making the information publicly available (GDPR Art. 14 (5) b).

Safeguards are described in the next section. The implementation of optional exceptions is outlined in the section dedicated to national models.

3.2 Appropriate safeguards

Protective measures may be of a technical or organisational nature. The technical measures may concern the data, medium or procedure, and the organisational measures may concern the staff, documentation or procedures. Examples of **technical measures** concerning 1) the *data* are pseudonymization, anonymization or aggregates of personal data; 2) the *medium* are encryption of personal data, internal measures by the data controller and data processor to prevent access to personal data, or measures to verify and prove who has registered, changed or transferred personal data; 3) the *procedure* are measures to continuously safeguard confidentiality, integrity, availability and resilience of processing systems and services in relation to the processing of personal data including the capacity to restore the availability or to safeguard access to personal data in a timely manner in the event of a physical or technical incidents.

Examples of **organisational measures** concerning 1) the *staff*: are appointing a data protection officer, or measures to raise the competence of the staff dealing with personal data, 2) the *documentation* are risk assessments, controller's record of processing activities, data processor agreements, guidelines, or non-disclosure agreements, 3) the *procedures* are a process for regularly testing, assessing and evaluating the effectiveness of technical and organisational measures for ensuring the security of the processing, establishing specific procedures to ensure that Union and Member State law are adhered to in case personal data is transferred or processed for some other purpose, or carrying out a data protection impact assessment.

4 National models

In **Czechia** application of the DGPR is still in progress. At the time of writing of this article, there is a mixed model of the previous Personal Data Protection Law (Czech law) https://zakonyprolidi.cz/cs/2000-101 and the GDPR regulation that overrides some parts of it. Parts that are not overridden by the regulation are still valid – e.g. existence and duties of the Office of Personal Data Protection established by law 101/2000 – until new "adaptation law" that replaces law 101/2000 is passed. The proposal of such a new law adopting the GDPR is now in the legislative process. There was a government proposal in March 2018, and after going through committees and debates in the Chamber of Deputies (lower chamber) of the parliament where it went through 29 amendments, it was passed to the Senate (upper chamber) on 8 January 2019. Currently, it is in the Senate committees, collecting more proposals for amendments. The proposal will be debated on the Senate floor on 30 January.[18] Several of the proposed amendments relate to research exceptions. At the time of passing the proposal to Senate, some deputies added § 16 that was not present in the government proposal. It is titled *"Collecting personal data for scientific or historical research or for statistical purposes":*

- Processing for these purposes is allowed provided that various protecting measures incl. pseudonymisation, maintaining processing logs according to Art. 5 of the GDPR, regular audits,

[16] Mandatory exceptions are directly applicable. They do not need to be incorporated into the national laws.

[17] The data subject shall have the right to obtain from the controller the erasure of personal data concerning him or her (GDPR Art. 17 (1)).

[18] The current status of the proposal and all suggested changes can be followed (in Czech) at the website of Czech Parliament: http://public.psp.cz/en/sqw/historie.sqw?o=8&T=138 (28.1.2019).

etc., are followed. The measures shall be "commensurate with state of the art, the cost of execution, the nature, scope, context and purposes of the processing."[19]

- § 16 ends with this sentence: "Article 15[20] and, to its corresponding extent, Article 5[21] of the GDPR […] shall not apply where processing is necessary for the purposes of scientific research, and the provision of information would require a disproportionate effort."

Thus, the current proposed law would allow scientific processing including large scale data collection for Natural Language Processing provided that best effort is taken to protect personal data. However, the version has to be adopted yet.

The **Estonian** Personal Data Protection Act (PDPA 2018a) sets the following requirements for the processing of personal data for scientific research (§ 6):

1) Personal data may be processed without the consent of the data subject for research purposes mainly if data has undergone pseudonymisation.
2) Processing of data without consent for scientific research in a format which enables identification of the data subject is permitted only if the following conditions are met:
   a) after removal of the data enabling identification, the goals of data processing would not be achievable, or achievement thereof would be unreasonably difficult;
   b) the person carrying out the scientific research finds that there is a predominant public interest for such processing;
   c) obligations of the data subject are not changed by the processed personal data, and the rights of the data subject are not excessively damaged in any other manner.
3) The data controller may limit the data subject's right of access, right to rectification, right to the restriction of processing and right to object in so far as the exercise of these rights are likely to render impossible or seriously impair the achievement of the objectives of the processing for research purposes.
4) In case of processing of special categories of personal data an ethics committee in the corresponding area verifies, before the commencement of the processing, compliance with the requirements set out in this section. In the absence of an ethics committee in a specific area, the Data Protection Authority verifies the fulfilment of requirements.

According to the **Finnish** model, the Data Protection Act (DPA 2018) and the preamble of the Government Proposal for Data Protection Act (Draft PDPA 2018b) outline the following conditions for processing personal data for scientific research:

1) **Data protection in general:** The legal basis for processing personal data by scientific researchers is, according to GDPR §6.1e, i.e. *performance of a task carried out in the public interest* based on the research organisation's legal mandate to do research as long as, according to GDPR §5.1f, the data is processed in a manner that ensures appropriate security of the personal data. Research organisations also have the right to store personal data as long as necessary and reuse them for secondary research purposes based on GDPR §5.1b, i.e. *further processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes shall … not be considered to be incompatible with the initial purposes.*
2) **Data protection in special categories:** According to DPA 2018 §6.7 and §6.8, the above also applies to personal data for special categories mentioned in GDPR §9.1 (with the exception of archiving genetic data) provided that suitable and specific technical and organisational measures are implemented to safeguard the fundamental rights and the interests of the data subject.
3) **Limitations to the rights of data subjects:** According to DPA 2018 §31.1, the data subjects may have limited rights to stop processing of personal data for scientific and historical research if the processing is necessary for carrying out the research, in which case the motivation for why the processing is necessary should be included in a research plan identifying the Principal Investigator.
4) **Limitations to the rights of data subjects in special categories:** According to DPA 2018 §31.3, if limitations to the rights of the research subject are applied to personal data in special categories,

[19] http://www.psp.cz/sqw/text/tiskt.sqw?o=8&ct=138&ct1=0&v=PZ&pn=12&pt=1 (28.1.2019)
[20] Right of access by the data subject.
[21] Principles relating to processing of personal data.

the research plan should assess how the limitations impact the rights and freedoms of the data subject. The written assessment must be delivered to the Data Protection Ombudsman ahead of starting the processing.

In **France**, the national law completing the GDPR initially proposed on 13 December 2017, has finally been adopted on 20 June 2018. It has since been amended twice: by the Decree n° 2018-687 of 3 August 2018, and by the Ordinance n°2018-1125 of 12 December 2018 (which will enter into force on 1 June 2019 at the latest).

Unlike the German legislator, who adopted a whole new statute to comply with the GDPR, the French chose to modify the "*Loi informatique et libertés*" (LIL) which was one of the first comprehensive data protection laws in Europe (dating back to 1978).

The processing of personal data for scientific and archiving purposes is regulated in articles 78 and 79 (according to the new numbering, which will enter into force on 1 June 2019). Article 78 provides that when data are processed for scientific purposes, certain rights of data subjects (access, rectification, restriction and the right to object) can be limited. The exact conditions in which such limitations are possible are to be specified by a Decree (*Décret en Conseil d'Etat*) which to the best of our knowledge has not yet been adopted.

Article 79 concerns purpose extension. It specifies that when data were collected for a different purpose and then re-used for research purposes (according to the purpose extension principle), the obligation to provide information to data subjects (art. 14 GDPR) does not apply.

Germany is probably the first country to have adopted a comprehensive national law to complete the General Data Protection Regulation. The new Bundesdatenschutzgesetz (BDSG) was adopted on June 30, 2017.

It shall be kept in mind that BDSG only applies to the processing of personal data by private entities and by public bodies of the German Federation (Art. 1 of the BDSG). Processing of personal data by public bodies of the Länder (such as universities) is governed by regional norms (Landesdatenschutzgesetze, LDSG). To the best of our knowledge, no LDSGs has yet been updated to conform to the GDPR. Therefore, for now, the situation regarding the processing of personal data for research purposes in German universities is not entirely clear.

As far as public bodies of the Federation (such as certain research institutes) and private entities are concerned, the processing of personal data for research purposes will be governed by Art. 89 of the GDPR, completed by section 27 of the new BDSG. The latter contains four paragraphs.

Firstly, section 27(1) of the new BDSG allows for processing of special categories of personal data for research purposes "if such processing is necessary for these purposes and the interests of the controller in processing substantially outweigh those of the data subject in not processing the data". The provision is based on Art. 9(2)(j) of the GDPR, which seems to leave the Member States the decision on whether to allow processing of special categories of data for research purposes based on the balance of interests. Interestingly, the new German law also contains a list of possible 'appropriate safeguards' for such processing[22]. The list is not meant to be exclusive, and other safeguards are also possible; moreover, it only expressly applies to the cases where special categories of data are processed. Moreover, as the GDPR does not expressly do it, section 27(3) of the new BDSG (still based on Art. 9(2)(j) of the GDPR) states that (according to the general principle of s. 89(1) of the GDPR) special categories of personal data processed for research purposes shall be pseudonymised, and then anonymised as soon as the purposes allow it.

[22] The safeguards "may include in particular the following: 1. technical organizational measures to ensure that processing complies with Regulation (EU) 2016/679; 2. measures to ensure that it is subsequently possible to verify and establish whether and by whom personal data were input, altered or removed; 3. measures to increase awareness of staff involved in processing operations; 4. designation of a data protection officer; 5. restrictions on access to personal data within the controller and by processors; 6. the pseudonymization of personal data; 7. the encryption of personal data; 8. measures to ensure the ability, confidentiality, integrity, availability and resilience of processing systems and services related to the processing of personal data, including the ability to rapidly restore availability and access in the event of a physical or technical incident; 9. a process for regularly testing, assessing and evaluating the effectiveness of technical and organizational measures for ensuring the security of the processing; 10. specific rules of procedure to ensure compliance with this Act and with Regulation (EU) 2016/679 in the event of transfer or processing for other purposes".

Secondly, section 27(2) provides for derogations from certain rights of data subjects, i.e., the right of access, rectification, restriction of processing and right to object. As suggested by Art. 89(2) of the GDPR the derogations apply when these rights are likely to render impossible or seriously impair the achievement of the research purposes and are necessary for their fulfilment. The German federal legislator has therefore taken full advantage of the leeway left by Art. 89(2) of the GDPR and legislated in favour of freedom of research.

Moreover, the legislator even went further than expressly allowed by this article and allowed for a derogation from the right of access when the provision of information listed in Art. 15(1) of the GDPR would involve a disproportionate effort. This derogation seems to be based on recital 62 of the GDPR.

Finally, section 27(4) of the new BDSG states that the controller may publish personal data (processed for research purposes) only if the data subject has provided consent or if doing so is indispensable for the presentation of research findings on contemporary events. This seems to serve as a limit to Art. 89 of the GDPR by stating that, in principle, special rules concerning research stop where publication of personal data starts.

In **Greece**, a Draft Bill for Personal Data (PDPA 2018c) implementing the GDPR after public consultation (which was completed on March 5, 2018), has been adopted and put into force as of 25 May 2018. The Bill contains an article dedicated to the processing of PD for "scientific or historical research or for statistical data". Processing of PD is allowed *if the subjects have given their consent for this or previous studies on the same data, if the data come from publicly accessible sources* or *if the processing can be proven to be required for the research*. For the processing of the special categories, the Bill is more restrictive; especially for research on genetic data prior consultation with the Data Protection Authority is mandatory. Medical data processing is allowed, provided the researchers involved are legally or professionally bound by confidentiality. Pseudonymisation or anonymisation are recommended but only when they do not hinder the purposes of the research. Overall, this draft Bill can be considered favourable towards research purposes.

The **Italian** Republic transposed the GDPR by legislative decree No 101/2018, which entered into force on 19 September 2018 (Italian law). According to that, personal data for scientific research can be processed without the consent of the data subject in the following cases: i) scientific research has been pursued according to the provision of law, provided that the data controller carries out an impact assessment and makes it publicly available, analysing the necessity and proportionality of the processing, the risks with respect to the rights and freedoms of data subjects, and safety measures to deal with these risks; ii) due to particular reasons, informing the data subject about the processing of personal data proves impossible or would involve a disproportionate effort, and it is likely to render impossible or seriously impair the achievement of the research objectives, provided that: a) the data controller shall take appropriate measures to protect the data subject's rights and freedoms and legitimate interests, b) the research project has received favourable and motivated opinion from the Ethics Committee, c) the research project has been submitted to preventive consultation with the Italian Data Protection Authority (It. Garante per la protezione dei dati personali) and to an impact assessment, in accordance with Art. 35 and 36 of GDPR.

In accordance with Art. 110-bis of the Privacy code – as modified by legislative decree No 101/2018 – the reuse of data for research purposes is allowed when: i) it is carried out by third parties that mostly deal with research activities, ii) the information about the processing of personal data proves impossible or would involve a disproportionate effort, and it is likely to render impossible or seriously impair the achievement of the research objectives, iii) it is subject to prior authorization by the Italian Data Protection Authority, made dependent on the adoption of appropriate action in compliance with Art. 89 of GDPR. With specific reference to prior authorisation by the Italian Data Protection Authority, decisions on an application submitted in accordance with Art. 8 of legislative decree No 101/2018 shall be adopted and communicated to the applicant within 45 days after its receipt. The absence of delivery shall take the place of refusal. Also, Art. 8 of legislative decree No 101/2018 provides for the Italian Data Protection Authority to allow the reuse of data for research purposes also by means of general measures.

The Italian Data Protection Authority has been organising several information meetings with Italian universities and public research bodies to raise awareness among the different research communities

and university administrative staff on the changes introduced by the GDPR and their impact on research activities[23].

The next example is **Lithuania**. To duly comply with GDPR the new version of Lithuanian Law on Legal Protection of Personal Data (LLPPD 2018) was enacted and entered into force since 16 July 2018. The previous version of the law included a special exemption for scientific research in Art. 12, which contained quite detailed requirements for the procession of personal data without the data subject's consent. Among other things, the prior checking procedure by the State Data Protection Inspectorate was required. In contrast with the previous regulation and with Estonian and Finnish models as described above, the newly enacted LLPPD 2018 contains no special provisions dealing with the research exemption. The requirement of the prior checking procedure was abandoned as well. It means that Lithuania has not used the opportunities and flexibilities provided in Art. 89 of GDPR. It also means that after the implementation of GDPR, the persons using personal data for scientific research have to rely directly on and comply with the general provisions of GDPR, especially Art. 6, Art. 17.3 and Art. 89. Following the new regulation, Lithuanian universities and other research institutions have enacted their own internal rules, dealing, *inter alia*, with the research exception. For example, Vilnius University, which is the leading research and study institution in Lithuania, enacted the rules on the data protection, which prescribes, that university has a right to process personal data for scientific or historical research purposes. The same rules also state, that, in line with Art. 17.3 of GDPR, the right to be forgotten is not applicable when processing is necessary for, among others, scientific research purposes.

Since the legislative changes were enacted very recently, so far there are no reported cases of application or conflicts concerning the new regulation of research exception. Therefore, the real impact of GDPR on scientific research is yet to be seen.

5 Conclusion

The development and use of language resources often involve the processing of personal data. Several aspects of personal data may be confusing. For instance, it is arguable whether human voice as biometric data should be considered to belong to special categories of personal data (sensitive data). It should also be emphasised that publicly available data are protected by the GDPR as well.

The legal framework setting for requirements for processing personal data for research purposes is based on the GDPR and national laws of the EU Member States. This means that in addition to the GDPR, researchers that wish to develop and use LRs for language research must further follow national requirements.

References

[BDSG] Bundesdatenschutzgesetz. Available at https://www.gesetze-im-internet.de/bdsg_2018/index.html (5.9.2018)

[C-101/01] Case C-101/01. Criminal proceedings against Bodil Lindqvist (6 November 2003). Available at http://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1521039149443&uri=CELEX:62001CJ0101 (3.4.2018)

[C-73/07] Case C-73/07. Tietosuojavaltuutettu vs. Satakunnan Markkinapörssi Oy and Satamedia Oy (16 December 2008). Available at https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:62007CA0073&qid=1536154290371&from=EN (5.9.2018)

[Czech law] Zákon č. 101/2000 Sb. Zákon o ochraně osobních údajů a o změně některých zákonů. Available at https://zakonyprolidi.cz/cs/2000-101 (28.1.2019)

[Data Protection Directive] Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. Official Journal L 281, 23/11/1995 p. 0031 – 0050. Available at http://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:31995L0046&qid=1522340616101&from=EN (29.3.2018)

[DPA 2018]. Data Protection Act (Finland). Entry into force 01.01.2019. Available in Swedish at: http://www.finlex.fi/sv/laki/alkup/2018/20181050 (20.1.2019)

[23] https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/8318508 and https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/7977380 [accessed 22.03.2019]

[PDPA 2018a] Estonian Personal Data Protection Act (Isikuandmete kaitse seadus). Entry into force 15.01.2019. Available in Estonian at https://www.riigiteataja.ee/akt/104012019011 (21.1.2019)

[Draft PDPA 2018b] Finnish Draft Act on Personal Data Protection (Hallituksen esitys eduskunnalle EU:n yleistä tietosuoja-asetusta täydentäväksi lainsäädännöksi) (01.03.2018). Available at https://www.eduskunta.fi/FI/vaski/HallituksenEsitys/Sivut/HE_9+2018.aspx (4.4.2018)

[Draft PDPA 2018c] Greek Draft Bill on Personal Data Protection (Νόμος για την Προστασία Δεδομένων Προσωπικού Χαρακτήρα). Available at http://www.opengov.gr/ministryofjustice/wp-content/uploads/downloads/2018/02/sxedio_nomou_prostasia_pd.pdf (18.4.2018)

[French law] Loi n° 2018-493 du 20 juin 2018 relative à la protection des données personnelles, modifying the French Data Protection Act (loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés)

[GDPR] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). OJ L 119, 4.5.2016, p. 1-88. Available at http://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1515793631105&uri=CELEX:32016R0679 (29.3.2018)

[González-Rodríguez et. al. 2008] Joaquín González-Rodríguez, Doroteo Torre Toledano, Javier Ortega-García (2008). Voice Biometrics. In Handbook of Biometrics edited by Anil K. Jain, Patrick Flynn, Arun A. Ross. Springer

[IPDPC] Italian Personal Data Protection Code. Legislative Decree 30.06.2003 No. 196. English version available at: http://194.242.234.211/documents/10160/2012405/Personal+Data+Protection+Code+-+Legislat.+Decree+no.196+of+30+June+2003.pdf (11.4.2018)

[Italian law] DECRETO LEGISLATIVO 10 agosto 2018, n. 101 Disposizioni per l'adeguamento della normativa nazionale alle disposizioni del regolamento (UE) 2016/679 del Parlamento europeo e del Consiglio, del 27 aprile 2016, relativo alla protezione delle persone fisiche con riguardo al trattamento dei dati personali, nonche' alla libera circolazione di tali dati e che abroga la direttiva 95/46/CE (regolamento generale sulla protezione dei dati).(18G00129) (GU n.205 del 4-9-2018). The Italian version of the law available at http://www.gazzettaufficiale.it/eli/id/2018/09/04/18G00129/sg (27.1.2019).

[Jain et. al. 2004] Anil K. Jain, Arun Ross, Salil Prabhakar (2004). An Introduction to Biometric Recognition. - IEEE Transactions on Circuits and Systems for Video Technology 14(1). Available at https://www.cse.msu.edu/~rossarun/BiometricsTextBook/Papers/Introduction/JainRossPrabhakar_BiometricIntro_CSVT04.pdf (31.3.2018)

[Kelli et al. 2015] Aleksei Kelli, Kadri Vider, Krister Lindén (2015). The Regulatory and Contractual Framework as an Integral Part of the CLARIN Infrastructure. 123: Selected Papers from the CLARIN Annual Conference 2015, October 14–16, 2015, Wroclaw, Poland. Ed. Koenraad De Smedt. Linköping University Electronic Press, Linköpings universitet, 13−24. Available at http://www.ep.liu.se/ecp/article.asp?issue=123&article=002 (28.3.2018)

[LED] Law of European Delegation. Law No. 25.10.2017 No 163. Available at http://www.gazzettaufficiale.it/eli/id/2017/11/6/17G00177/sg (11.4.2018)

[LLPPD 2018] Lithuanian Law Amending the Law on Legal Protection of Personal Data (Lietuvos Respublikos asmens duomenų teisinės apsaugos įstatymo pakeitimo įstatymas). Available at https://www.e-tar.lt/portal/legalAct.html?documentId=43cddd8084cc11e8ae2bfd1913d66d57 (30.8.2018)

[Oostveen 2016] Manon Oostveen (2016). Identifiability and the applicability of data protection to big data. International Data Privacy Law 6 (4), 299-309

[ORDA] Organisation of Research and Development Act. Entry into force 2.05.1997. English translation available at https://www.riigiteataja.ee/en/eli/513042015012/consolide (21.1.2019)

[Privacy Code] Code of conduct and professional practice Regarding the processing of personal data for historical purposes. English version available at http://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/export/1565819 (11.4.2018)

[VLO] CLARIN Virtual Language Observatory. Available at https://vlo.clarin.eu/ (18.4.2018)

[WP29 2017] WP29. Guidelines on Consent under Regulation 2016/679. Adopted on 28 November 2017 [adopted, but still to be finalized]. Available at http://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=615239 (2.4.2018)

[WP29 2014a] WP29. Opinion 05/2014 on Anonymisation Techniques Adopted on 10 April 2014. Available at http://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf (3.4.2018)

[WP29 2014b] WP29. Opinion 06/2014 on the notion of legitimate interests of the data controller under Article 7 of Directive 95/46/EC. Available at http://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp217_en.pdf (3.4.2018)

[WP29 2007] WP29. Opinion 4/2007 on the concept of personal data. Adopted on 20th June. Available at http://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2007/wp136_en.pdf (29.3.2018)

TalkBankDB: A Comprehensive Data Analysis Interface to TalkBank

John Kowalski
Carnegie Mellon University, USA
jkau@andrew.cmu.edu

Brian MacWhinney
Carnegie Mellon University, USA
macw@andrew.cmu.edu

Abstract

TalkBank, a CLARIN B Centre, is the host for a collection of multilingual multimodal corpora designed to foster fundamental research in the study of human communication. It contains tens of thousands of audio and video recordings across many languages linked to richly annotated transcriptions, all in the CHAT transcription format. The purpose of the TalkBankDB project is to provide an intuitive on-line interface for researchers to explore TalkBank's media and transcripts, specify data to be extracted, and pass these data on to statistical programs for further analysis.

1 Introduction

The origins of TalkBank trace back to 1984 with the creation of the CLAN (Child Language Analysis) tools and the associated CHAT transcription format (MacWhinney, 2000). The corpus began with annotated media of child language acquisition (CHILDES database) and has expanded to include fourteen annotated media language databases including SLABank for studying second-language acquisition, CABank for conversational data, ClassBank for study of language in the classroom, SamtaleBank for the study of Danish conversations, and a series of clinical databanks for aphasia, stuttering and other disorders. The size and scope of TalkBank continues to expand. As of this writing, TalkBank includes over 8TB of annotated media.

	CHILDES	**AphasiaBank**	**PhonBank**	**FluencyBank**	**HomeBank**	**TalkBank**
Age (years)	30	10	7	1	2	14
Words (millions)	59	1.8	0.8	0.5	audio	47
Linked Media (TB)	2.8	0.4	0.7	0.3	3.5	1.1
Languages	41	6	18	4	2	22
Publications	7000+	256	480	5	7	320
Users	2950	390	182	50	18	930
Web hits (millions)	5.0	0.5	0.1	0.1	0.4	1.7

Table 1: TalkBank Usage

John Kowalski and Brian MacWhinney 2019. TalkBankDB: A Comprehensive Data Analysis Interface to TalkBank. *Selected papers from the CLARIN Annual Conference 2018*. Linköping Electronic Conference Proceedings 159: 83–91.

Currently, most users interact with TalkBank data by using the CLAN program. CLAN consists of a set of tools for transcribing media in CHAT format, playing back the media with time-stamped annotations, and extracting statistics and metadata from a set of transcripts. CLAN has been refined for decades and is highly capable. However, using it requires a significant effort from the researcher to study the CLAN manual and learn CHAT annotations. Moreover, CLAN is mostly tuned for creating new transcripts and for working with single corpora. It is not designed for systematic queries of the entire TalkBank database to extract general patterns and statistics. Because of these limitations, CLAN is not an ideal tool for researchers who want to conduct wider corpus analyses on the existing database. Here we report on a new system, called TalkBankDB, designed to provide this additional functionality.

2 Increasing the Accessibility of TalkBank Corpora

Previously, browsing TalkBank required knowing the name of a corpus or area of research, finding its location within the talkbank.org domain (ex: fluency.talkbank.org), and then browsing/downloading the media and annotations.

Without prior knowledge of how TalkBank is structured and what corpora exist within each TalkBank collection, users may be unaware that particular resources exist. TalkBankDB provides a single online interface to query across all of TalkBank to find the names and categories of relevant corpora and links to media and transcripts. For example, a query for the Spanish language yields a list of transcripts within TalkBank spanning many separate corpora. Further queries can limit by date of recording, native language of speakers, age of participants, media type (audio/video/none), and others. The user will then have a list of all media and descriptive metadata matching their query, with links to each directly playable from the browser. After a query is submitted, clickable tabs appear to show descriptive lists of participants in matched transcripts, word tokens spoken, tokens grouped by type, and statistics for each speaker (number of words spoken, mean utterance length, and others.) TalkBankDB allows users to construct new combinations of corpora or subparts of corpora based on features they define (Figure 1 and Figure 2).

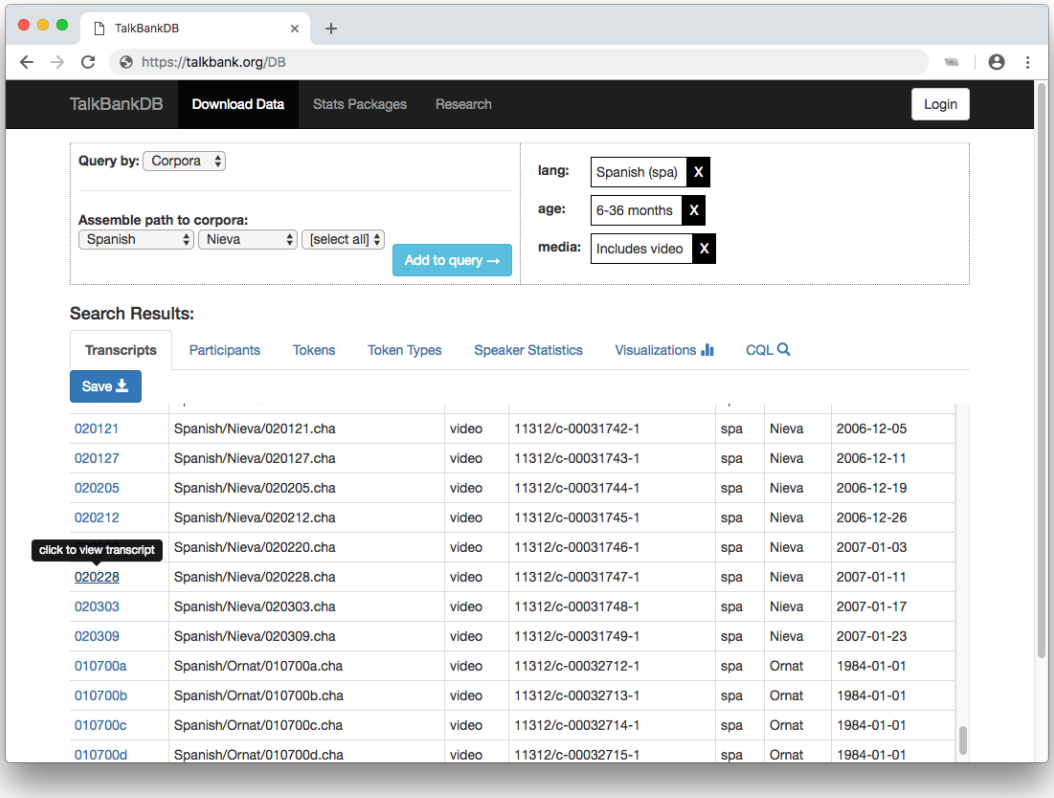


Figure 1. A query yields a table of all matching documents with metadata for each, allowing the user to further refine the query.

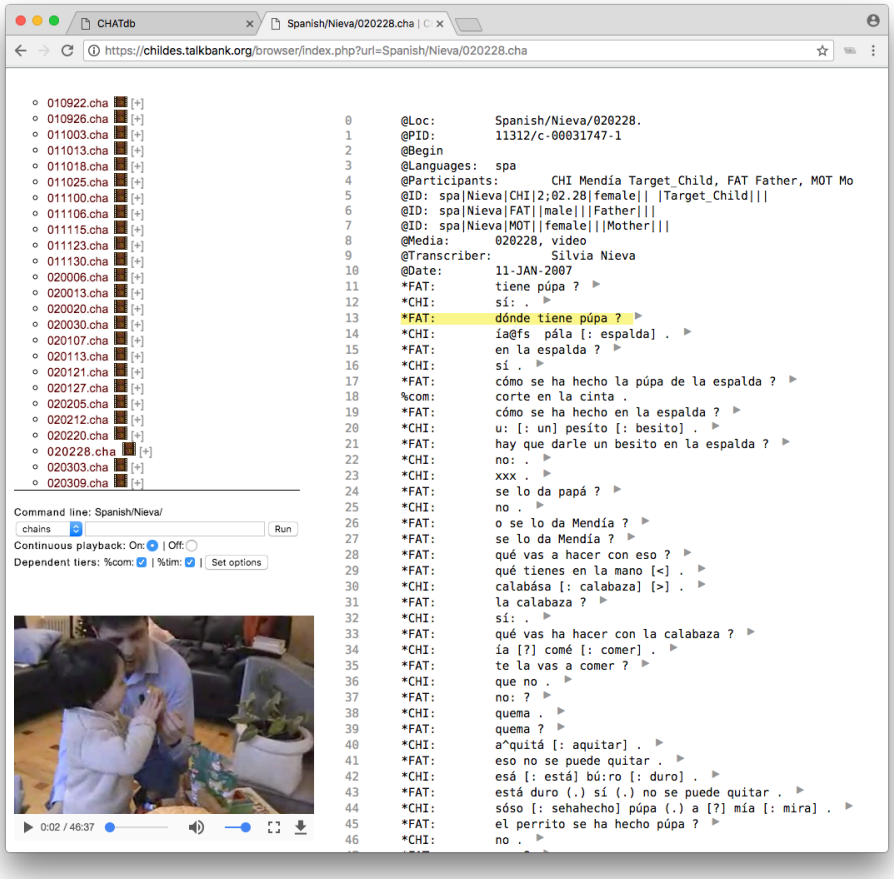


Figure 2. Clicking on the name of the transcript loads the corresponding media and annotations in the browser and allows for direct playback of the media.

In addition to using TalkBankDB to locate transcripts and media with specific features across TalkBank, researchers can derive statistical summaries of the annotations in the transcripts. A pulldown of variables to be extracted includes the age range of participants, the roles of speakers (mother, father, child, teacher, etc), the number of words spoken, mean utterance length, specific words used, and others. For instance, one can make a plot of frequencies of English article usage (a/an, the) by mothers speaking to their children in relation to their child's age. The exploration space enabled by this simple interface is huge.

Child language researchers had already built two systems designed to achieve this type of functionality. These are the childes-db project (Sanchez et al., 2018) and the LuCiD Toolkit (Chang, 2017). Both of these projects were created to analyse only the portion of TalkBank dealing with child language acquisition (CHILDES corpus), whereas TalkBankDB encompasses the whole of TalkBank. The principle goal of these systems is to output spreadsheets which can then be passed on to statistical analysis by systems such as R, NumPy, or Excel. TalkBankDB also provides this functionality.

The childes-db project (childes-db.stanford.edu) offers both a web interface and R package to analyse CHILDES. Downloaded CHILDES data are stored in a MySQL database. There are six main functions in the childes-db R library: get_transcripts(), get_participants(), get_tokens(), get_types(), get_utterances(), and get_speaker_statistics(). For the web interface, childes-db employs R Studio's Shiny Server enabling the plotting of variables also accessible from the aforementioned R library functions. The LuCiD Toolkit offers similar facilities to childes-db for exploring the CHILDES corpus. It also employs a Shiny server (gandalf.talkbank.org:8080) to offer a web interface to extract and analyse variables from the transcripts. However, this facility is based on a 140GB spreadsheet set that contains data and precomputed statistics from CHILDES. TalkBankDB differs from these facilities by creating

an editable document database from which statistics are computed dynamically, creating a more scalable and flexible system.

3 Database Architecture and Implementation Details

Creation of the TalkBankDB database relies on the fact that all TalkBank transcripts are pure UTF-8 text files that explicitly implement the CHAT annotation format. These files are then processed by the Chatter Java program, available from https://talkbank.org/software/chatter.html. Chatter can convert a CHAT file to XML that can be round-tripped back to the file's original CHAT format. The XML format and the associated schema facilitates use of TalkBank corpora by third party programs and systems, eliminating the need to parse complex raw strings.

Since JSON can be used directly by front-end web apps, TalkBankDB eliminates the need for the app to constantly convert XML to JSON and back again by first converting the XML transcripts outputted by Chatter to JSON using xml-js (Nashwaan, 2018). This tool supports bidirectional XML/JSON conversion. So, combined with Chatter, the system can provide a verifiable round-trip from JSON to the original CHAT formatted transcript.

Since much of the data and metadata contained within the TalkBank CHAT transcripts are subject to change and amplification, using a relational database like mySQL with a strict tabular schema is not as practical as a more flexible document database. The effort to pre-set a fixed schema with a normalized relational database can cause problems when the schema needs to be modified and extended with new phonology, sequence numbers for tiers, adding TEI annotations, etc.

To store our collection of JSON documents, we use MongoDB, a widely-used free and open-source document database. An added benefit of this document database is that it makes it easy to scale up to increasing data demands by allowing the database to be encoded across multiple inexpensive machines through "sharding". This can be very difficult to do with relational databases, where often the only option is to "scale up" by purchasing increasingly powerful machines. The strategy of scaling up through use of a single larger machine is not always possible, and it may eventually be unable to meet the growing size and computational demands of the database.

The front end web interface is written in standard HTML, CSS, and JavaScript to ensure cross-browser support. Care is taken so the JavaScript code is clearly commented and maintainable, following the popular "web component" design pattern common in many large-scale web apps.

Initially, TalkBankDB will include only publicly accessible data. Access will be controlled by the CLARIN single sign-on authentication system. Access to private clinical data will require a second-level of authentication.

4 Database JSON Format

The JSON format derived from a CHAT file (via CHATTER XML) is very simple and extensible. An example JSON representation of the utterance "talking to the tape recorder" is below:

```
{
  "who":"FAT",

  "uID":"u8",

  "words":[
    {
       "w":"talking",
       "mor":{
          "stem": "talk",
          "pos": "part"
       }
    },
    {
       "w":"to",
       "mor":{
```

```
            "stem": "to",
            "pos": "prep"
         }
      },
      {
         "w":"the",
         "mor":{
            "stem":"the",
            "pos": "det"
         }
      },
      {
         "w":"tape",
         "mor":{
            "stem":"tape",
            "pos": "n"
         }
      },
      {
         "w":"recorder",
         "mor":{
            "stem":"record",
            "pos": "n"
         }
      }
   ],
   "media":{
      "start": 20.062,
      "end": 20.805"
   }
}
```


Here we see that for each utterance:

- A speaker is defined, here as the father (`who: "FAT"`).
- The utterance's sequence number within the transcript (`uID: "u8"`).
- An array of words (`words: []`).
- A start/end time of the recording (`media: {start: 20.062, end: 20.805}`).

Each word object consists of:

- A word as it appears in the recording (`w: "talking"`).
- A morphology object consisting of an extensible number of properties. (`mor: {}`).
    - Here we define:
        - The stem or lemma of the word (`stem: "talk"`).
        - Part of speech of the word, here participle (`"pos: "part"`).
        - Other key/value pairs.

This simple utterance object is the fundamental building block of the JSON representation of a CHAT transcript and thus of TalkBankDB. All tab-delimited data downloads, statistics, visualizations, and token/grammatical pattern searches mentioned in this manuscript are derived from this repeating structure.

As of this writing, TalkBankDB is still less than a year old and we are in the process of adding additional information regarding specific coding features in CHAT transcripts for the %pho phonology line, the %mor morphology line, and the %gra grammatical relations line . However, the fundamental JSON-based structure will stay the same. Corpora outside TalkBank that have words tagged with stem and part of speech could also be converted to this simple format. They would then immediately get the benefits of the analysis, visualization, and search tools of TalkBankDB.

5 Searching for Token and Grammatical Patterns

TalkBankDB provides a toolkit to search for token and grammatical structures across TalkBank corpora. The toolkit interface is a language composed of a combination of regular expressions and a syntax to specify patterns of grammatical and other semantic tags. This query language is based on the popular Corpus Query Language (CQL) used by SketchEngine software (www.sketchengine.eu) (Kilgarriff, 2004), which in turn is based on the query language of the Corpus Work Bench (CWB) project developed at the University of Stuttgart (Christ, 1994). Since CQL syntax is familiar to many researchers, TalkBankDB implements a large subset of it.

In CQL, searching for a token takes the form of:
`[attribute="value"]`

Where `attribute` is often one of:

- "`word`" to match a particular word or set of words defined with a regular expression.
- "`lemma`" to match all forms of a lemma. Ex: jump → jump, jumps, jumping, jumped.
- "`pos`" to match a word based on its part of speech tag.

Any attribute/value defined on a word in the database can be searched this way. The full search language grammar along with all attribute/value pairs are defined on the TalkBankDB website. Listed below are examples of some common CQL search patterns.

To find all instances of...

- The word "bring":
  `[word="bring"]`

- Verb forms of "bring" (bring, brings, bringing, brought):
  `[lemma="bring"]`

- The sequence "bring your hat":
  `[word="bring"] [word="your"] [word="hat"]`

- The sequence "bring your" followed by a noun:
  `[word="bring"] [word="your"] [pos="N"]`

- The sequence "bring your" followed by an adjective and noun:
  `[word="bring"] [word="your"] [pos="ADJ"] [pos="N"]`

- The sequence "bring your" followed by one or more adjectives and a noun:
  `[word="bring"] [word="your"] [pos="ADJ"]+ [pos="N"]`

- The sequence "bring a/an/the" followed by one or more adjectives and a noun:
  `[word="bring"] [word="an?|the"] [pos="ADJ"]+ [pos="N"]`

To search for keyword/grammatical patterns in TalkBankDB, first define the corpus to search (by name, language, media present, etc.), then click on the "CQL" tab, enter a CQL query in the text field, and click "Submit". This generates a list of "concordances", matched keywords along with the words surrounding each. The concordances listed are in the format of a "Key Word in Context" (KWIC), where the matched keywords are highlighted in blue surrounded by the text that contains it in the transcript. In addition, the transcript path, speaker, and utterance ID within the transcript are also listed to provide additional context (Figure 3).

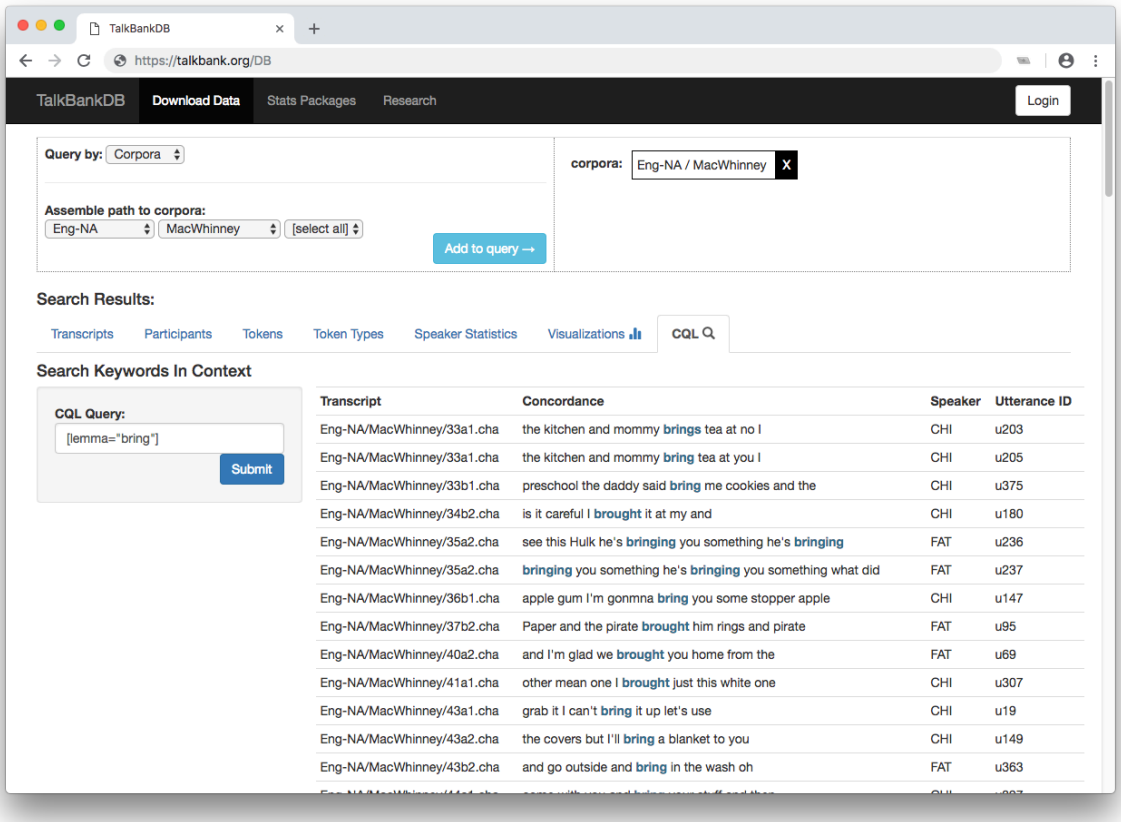


Figure 3. Selecting the MacWhinney corpus, then choosing "CQL" tab to generate KWIC concordances for the lemma of verb "bring".

6 Visualizations

TalkBankDB also provides methods for client-side data visualization. After selecting a corpus, clicking on the "Visualizations" tab opens a collection of options for generating plots and for downloading the data used to generate these plots. The layout for each visualization UI is similar. On the left is a UI for specifying the options and information needed to generate a plot; on the right is the interactive plot generated from this information. At the top is the default "Plot" tab just described and a "Table" tab that displays the downloadable data for the current plot. Figure 4 shows the "Word Frequency by Age" visualization for the MacWhinney corpus, choosing the English articles (a, an, the).

In contrast to servers like the RStudio Shiny Server (shiny.rstudio.com) that process data for creation of a static image to be displayed in the browser, TalkBankDB's server simply sends back data, thereby giving the client freedom to plot with any visualization library. Many visualization libraries are freely available today, such as Chartist, Highcharts, Google Charts, Chart.js, and others. New libraries can be swapped in and out of TalkBankDB as required. We chose the open-source C3.js library, because it is easy to use and covers all the plots we currently need. Moreover, the plots it generates can be explored by zooming in/out and dragging along values of the x-axis. Another benefit of C3 is that it is based on D3, a very powerful graphical toolkit for browsers. If a visualization is desired but not covered by C3, one can extend the code using the powerful toolkit that D3 provides.

TalkBankDB currently provides visualizations for exploring word frequency by age, number of utterances/words by age, mean word length of utterances (MLUw), and type-token ratio (TTR). Adding more visualizations to TalkBankDB is straightforward. Each client-side visualization is processed by its own JS module that follows a 3-step pattern of:

1. Make API request to server to retrieve data on selected corpus.
2. Process or compute stats on data.
3. Call a visualization function to plot processed data.

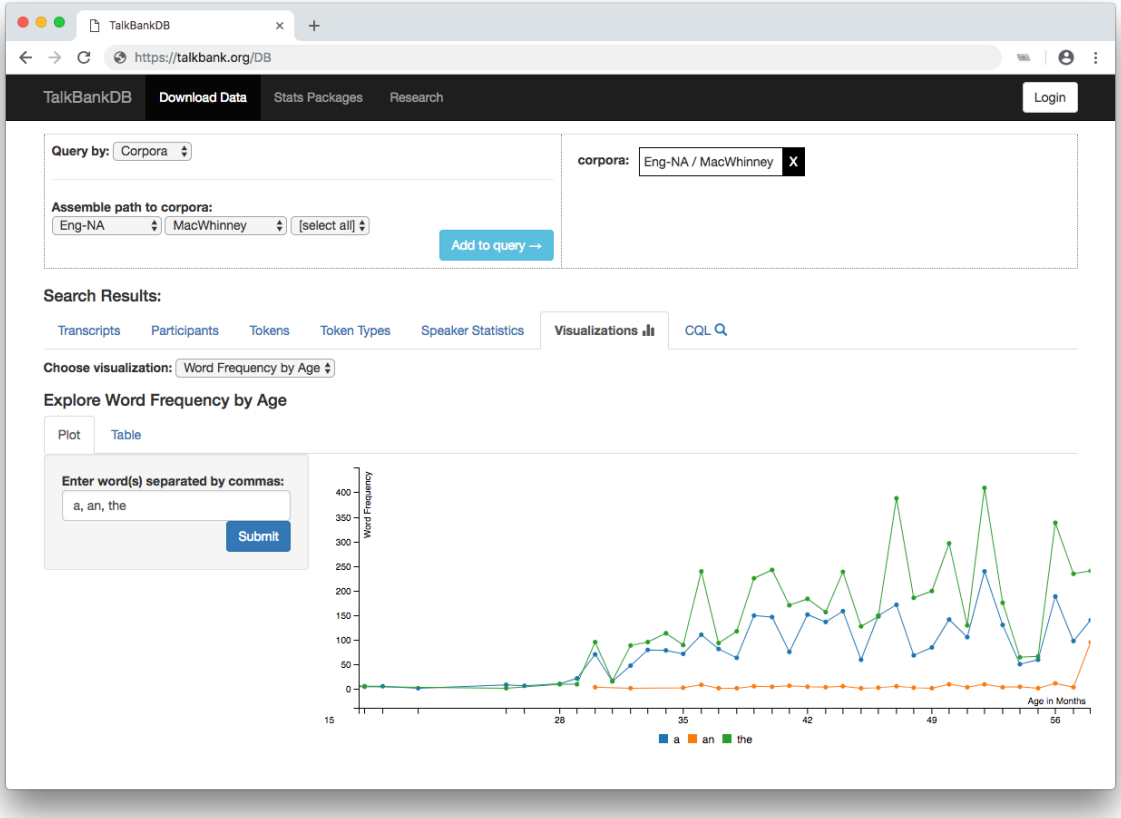


Figure 4. Plotting word frequency by age of English articles (a, an, the) by age within the MacWhinney corpus.

7 Other Features

A beta version of TalkBankDB is currently at https://talkbank.org/DB. The features offered will be refined and expanded on the basis of input from the extended CLARIN community. Below we list some features in the current beta specification:

- Button to download local copies of tab-delimited tables generated by TalkBankDB queries for use in further statistical analysis.
- Include links in tables returned by queries to open and play audio/video transcripts in browser.
- Option to upload new files, define new TalkBank corpora branches.
- Option to view/edit transcripts.
- Maintain state in state of user's queries and analyses in URL so that analyses can be shared with others by sending a unique URL.

8 Related Work

The design and scope of TalkBankDB has been influenced by our work with several related projects, including SketchEngine (sketchengine.eu), Corpus Workbench (cwb.sourceforge.net), EXMARaLDA (exmaralda.org), MTAS (meertensinstituut.github.io/mtas), ANNIS (corpus-tools.org/annis), and Alpheios (alpheios.net), as well as the childes-db and LuCiD Toolkit projects mentioned earlier. These systems have features that continue to influence the development of TalkBankDB, especially for facilitating the creation and interpretation multi-tier annotations of multimedia corpora.

9 Additional Applications and Expansions

Although TalkBankDB is designed around the CHAT format, it can be applied to other formats and projects in the CLARIN ecosystem. Since the format stored in TalkBankDB is not CHAT, but a simplified JSON representation, including documents in TalkBankDB only requires a script to convert from another (non-CHAT) CLARIN format to this JSON format. The JSON schema currently includes entries for metadata such as document name, version number, corpus name, and media type. In addition, it has a list of participants, and an "utterances" array with an entry for each word, with each word supplemented with metadata including speaker ID, token morphology, and utterance number. Any format encoding transcripts of spoken text with morphological tagging can easily be adapted for inclusion in TalkBankDB.

10 Conclusion

A main goal of TalkBankDB is to provide the CLARIN/TalkBank community with easier access to TalkBank data and analysis. Features such as word usage, utterance length, measures of language acquisition speed and ability by demographics can easily be selected, output, plotted, and analyzed through the web interface. The TalkBankDB interface can also be used in classroom demonstrations and project assignments for humanities or data analysis students, increasing awareness of the CLARIN community and inspiring future members.

References

[Chang 2017] Chang, F. (2017) The LuCiD language researcher's toolkit [Computer software]. Retrieved from http://www.lucid.ac.uk/resources/for-researchers/toolkit/

[Christ 1994] Christ, O. (1994). A modular and flexible architecture for an integrated corpus query system. *arXiv preprint cmp-lg/9408005*.

[Kilgarriff 2004] Adam Kilgarriff, Pavel Rychlý, Pavel Smrž, David Tugwell. Itri-04-08 the sketch engine. Information Technology, 2004.

[MacWhinney 2000] MacWhinney, B. (2000). The CHILDES Project: Tools for Analyzing Talk. 3rd Edition. Mahwah, NJ: Lawrence Erlbaum Associates

MongoDB [Computer software]. (2018). Retrieved from https://www.mongodb.com.

Node.js [Computer software]. (2018). Retrieved from https://nodejs.org/en.

[Nashwaan 2018] Nashwaan, Yousuf, xml-js, (2018) GitHub repository, https://github.com/nashwaan/xml-js

[Sanchez] Sanchez, A., Meylan, S., Braginsky, M., MacDonald, K., Yurovsky, D., & Frank, M. C. (2019). childes-db: a flexible and reproducible interface to the Child Language Data Exchange System (CHILDES). Behavior Research Methods. 1-14.

DI-ÖSS - Building a digital infrastructure in South Tyrol

Verena Lyding and **Alexander König** and **Elisa Gorgaini**
and **Lionel Nicolas** and **Monica Pretti**
Institute for Applied Linguistics
Eurac Research, Bolzano, Italy
{firstname.lastname}@eurac.edu

Abstract

This paper presents the DI-ÖSS[1] project, a local digital infrastructure initiative for South Tyrol, which aims at connecting institutions and organizations working with language data. It aims to facilitate and increase data exchange, joint efforts in processing and exploiting data and synergies, and thus linking to big European infrastructure initiatives. However, while sharing the overall objectives to foster standardization, increase efficiency and sustainability, a local initiative faces a different set of challenges on the implementation level. It aims to involve institutions which are less familiar with the logic of infrastructure and have less experience and fewer resources to deal with technical matters in a systematic way. The paper will describe how DI-ÖSS addresses the needs for a digital language infrastructure on a local level, lay out the course of action, and depict the targeted short-, mid- and long-term outputs of the project.

1 Introduction

In recent years, the field of Digital Humanities has seen the development of multiple infrastructure projects at European level. Among the most well-known initiatives CLARIN (Krauwer and Hinrichs, 2014) and DARIAH (Edmond et al., 2017) target the needs of researchers, with CLARIN being mostly centered around the discipline of linguistics, and, to a lesser degree, history and literary studies, while DARIAH focuses on the broader field of all the arts and humanities. In some countries, like the Netherlands, CLARIN and DARIAH have even started to merge into a joint CLARIAH[2] project (Odijk, 2016).

Europeana, on the other hand, focuses on the cultural heritage sector (Europeana Foundation, 2015). Its main aim is to strengthen the networks between institutions like galleries, libraries, archives and museums (GLAM), especially by aggregating their metadata as much as possible to make them searchable in an easier and more convenient way. In doing this, Europeana also creates attractive portals to these data.[3]

The large field of smaller institutions, both in the public and the private sector, is not targeted by any of these big infrastructures, even though it could benefit from a close collaboration with Digital Humanities. It contains smaller libraries,[4] archives, cultural associations, and publishing houses; actors that deal with language and contribute to the field of research and heritage, but who are themselves too small to easily participate in one of the big infrastructures. These minor but central players are the target of DI-ÖSS: *Digitale Infrastruktur für das Ökosystem Südtiroler Sprachdaten und –dienste* (Digital infrastructure for the ecosystem of South Tyrolean language data and services).

2 Motivation

The increasing availability and the wide-spreading use of digital data in various academic disciplines, such as the humanities and the social sciences, have progressively led to heighten awareness to issues

[1] *Digitale Infrastruktur für das Ökosystem Südtiroler Sprachdaten und –dienste* - Digital infrastructure for the ecosystem of South Tyrolean language data and services

[2] https://www.clariah.nl

[3] For example, a special portal was launched to commemorate the start of the First World War. https://www.europeana.eu/portal/en/collections/world-war-I

[4] In contrast to the big national libraries targeted by Europeana.

Verena Lyding, Alexander König, Elisa Gorgaini, Lionel Nicolas and Monica Pretti 2019. DI-ÖSS - Building a digital infrastructure in South Tyrol. *Selected papers from the CLARIN Annual Conference 2018*. Linköping Electronic Conference Proceedings 159: 92–102.

about data standardization, preservation, exchange and reuse, thus calling for a shared agenda on workflows for data collection, processing and analysis. Consequently, a number of large-scale research infrastructure initiatives have been launched over the last decade (see Section 1). These have all been conceptually conceived as network communities and have primarily directed their efforts toward setting common standards, identifying best practices and employing technical solutions so as to foster accessibility, interoperability and sustainability whilst sharing and reusing data in national and/or international research contexts.

The project DI-ÖSS borrows potential from the aforementioned humanities infrastructures and functionally attempts to replicate their prospect of establishing connections and deploying synergies. Nonetheless, it theoretically adjusts such potentiality to a local level, namely to the Autonomous Province Bozen/Bolzano-South Tyrol situated in northern Italy. This shifts the operational focus with regard to both the types of participants/contributors and the scope/scale. Hence, small and very small institutions or companies, which are not necessarily connected to the research domain, turn into main actors, whereas the geographical, political and cultural area of South Tyrol becomes the core stage of the project.

The rationale behind this deliberate focusing and the subsequent course of action is twofold: in the first place, the pivotal role played by small organizations in performing fine-grained work on local cultural assets; in the second, the resulting need for notional models, actual practices and, as a linking element, tangible means of optimization in local contexts.

The first reason for upholding a locally-centered infrastructure is the contribution regional organizations make toward strengthening a sense of cultural identity and belonging. In fact, by systematically undertaking a series of data-driven tasks, i.e. collecting, recording, cataloging, processing, analyzing, evaluating, archiving and disseminating existing resources (cf. Figure 1) according to local needs, demands or requests, they buttress the conservation of today's ethnolinguistic legacy. However, given the broad spectrum the aforesaid duties span – from information retrieval to knowledge management – and given the independent recourse to consistent yet individually streamlined workflows, local actors' efforts often translate into heritage preservation and territory enhancement to a degree which may not be proportionate to their investments, and could be greatly facilitated by using synergies with other actors.

Moreover, regional data and services may exhibit particularities which are location-dependent and can, therefore, be effectively accommodated only at a local level.[5]

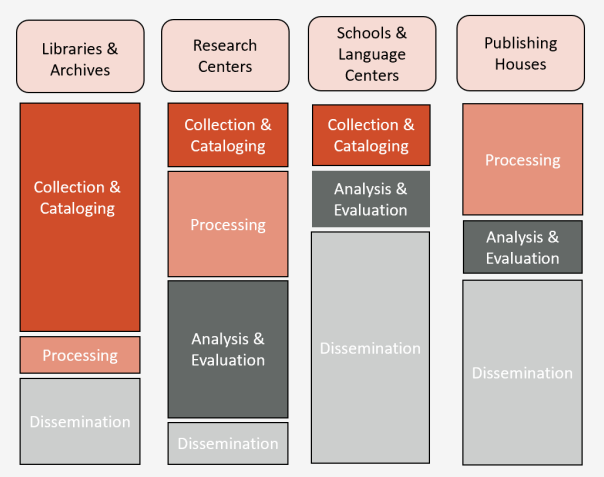


Figure 1: Exemplary overview of typical task allocation

The second motive for "going local" lies with the realization that regional institutions or companies are – as a tendency – prone to be less informed about existing approaches and/or ongoing efforts in the realm of infrastructure creation, development and implementation – which is currently pursued on higher structural levels (research, public sectors on cultural heritage preservation) and by larger frameworks, as

[5]E.g. South Tyrolean German is a local linguistic variety which is mostly documented and studied *in situ*.

explained above. They may, therefore, fail to benefit from the opportunities which arise in the process. Furthermore, should they succeed in keeping up to date, they might still lack some of the skills or resources essential to apply the knowledge acquired. By supporting small public or private organizations in the language sector in making the first steps to conform to bigger initiatives' standards, conventions and technologies, DI-ÖSS intends to sensitize them to the advantages of a digital language infrastructure. These include developing a framework for exchanging theoretical approaches and best practices, implementing specific interfaces for sharing data and/or tools, and coordinating and executing complex multi-step workflows. The primary objective is the promotion of cross-institutional efficient work.

In this regard, DI-ÖSS aims at actualizing the conditions for joining efforts, catalyzing processes and sharing outputs in order to support an interconnected ecosystem of South Tyrolean language data and services. Its synergetic potential is indicated by the existence of overlapping tasks and objectives amongst diversified organizations and by the use of comparable data sets. Having said that, a particularly apt way of exploiting such scope is a task-oriented allocation of work on the basis of each institution's dedicated spheres of competence, which allows for both medium-term quality improvement and long-term cost savings.

A concrete example may clarify this assertion. Libraries and archives specialize in collecting and cataloging text documents, whereas linguistic research institutes focus on analyzing data and developing sophisticated tools to automatically process and evaluate them. By combining and exchanging skills, the former can profit from rigorous high-quality linguistic research, while the latter can evade the elaborate task of data collection in return. Finally, data themselves prove valuable: by sharing them under established copyright rules, each institution can take advantage of a larger database without having to do any additional work to build it up.

3 Project Plan

As DI-ÖSS is a pilot project and deals with an abstract idea of a digital infrastructure, a careful project planning has been put into place to drive its step-by-step actualization and adaptation - if needed. Below this plan is laid out by first stating the aims of the project, then examining the approach to reach these goals and finally listing and illustrating the targeted outcomes.

3.1 Aim

The project aims to design and, in a second step, implement prototypical infrastructure components of a cross-institutional operational infrastructure, which is bound to digitally network South Tyrolean language data and services as integral elements of the local language ecosystem. This comprises different types of institutions and companies as well as four main project partners (see Section 4.1), all of which deal with language and cultural resources either in a commercial or in a non-commercial way.

It is within the consortium, selected to represent the wide-ranging variety of language institutions which the project sets out to cover, that DI-ÖSS plans to pilot the aforesaid prototype by means of concrete institution-specific use cases. They have been designed to meet the needs and reproduce the archetypal application scenarios of each partner, thus underpinning mutual bidirectional exchanges between organizations, permitting reusability in analogous circumstances or contexts and backing structural supportability in the project. Therefore, they ought to showcase how the target infrastructure can enrich and optimize each partner's contribution.

Additionally, the implementation of the infrastructure itself and of the correlated use cases is geared toward time-staggered objectives. In particular, the short-term facilitation of outputs at a local level should augment the quality of the project consortium's work, generate room for further development and lead to the medium-term extension of the infrastructure to the entire South Tyrolean language ecosystem. Then, the long-term evaluation of the infrastructure feasibility will be accompanied by the overarching aim of connecting it to national and pan-European initiatives.

Lastly, the higher-level goal of DI-ÖSS is, among other things, laying the foundations of the South Tyrolean digital cultural heritage, as the project motto epitomizes: "Start local, think big."

3.2 Approach

The aforesaid large-scale projects – CLARIN[6] at the forefront – are transnational initiatives of crucial reference for DI-ÖSS in that they offer efficacious examples of location-independent aggregation, convention-driven harmonization and user-friendly utilization of valuable resources. They, therefore, provide a yardstick by which to draft preliminary aims and orientate expected results. This liaison notwithstanding, DI-ÖSS differs from them in a series of aspects,[7] the foremost being the theoretical approach (for an illustrative exemplification of how CLARIN and DI-ÖSS differ in their specific focus cf. Figure 2).

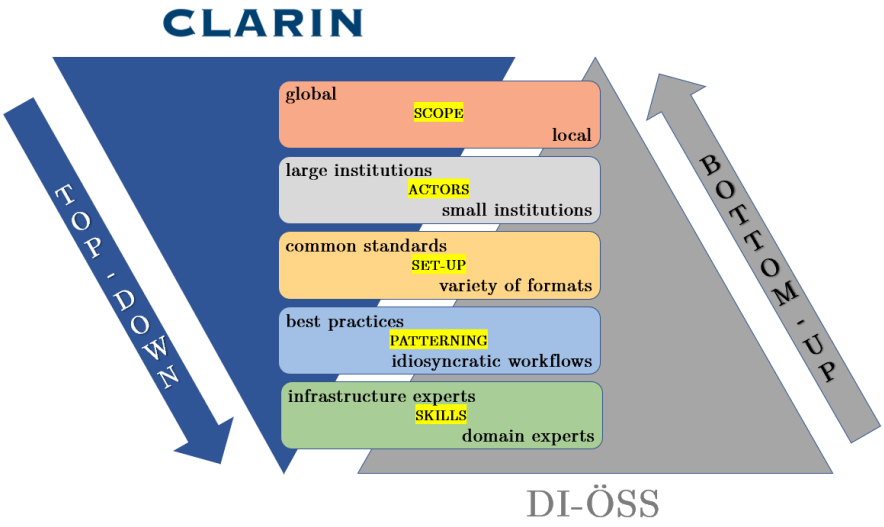


Figure 2: CLARIN vs. DI-ÖSS

Bigger projects employ a top-down method in accordance with each initiative's agenda, developing intelligible research applications. Conversely, DI-ÖSS makes use of a bottom-up strategy, intentionally launched by addressing a more-applied type of participants and a local scope. This anticipates two further points: on the one side, the necessity of *following the actors* (Latour, 2005), i.e. *learn from them* so as to identify data and service overlaps and share solutions by means of functional interplay and technical interoperability. On the other, the challenge of incorporating a heterogeneous ensemble of contributors – consisting of diverse data formats, idiosyncratic workflows and domain experts – into a coherent *assemblage* (Deleuze et al., 1987), i.e. a multiplicity of configurations connected in fluid exchanges. DI-ÖSS intends to pursue the surfacing of plastic articulations amongst the participating institutions and companies so that relevant theoretical and practical resources can flow within the infrastructure and give rise to a dynamic *constellation* of placeable, displaceable and replaceable relationships of *exteriority*.

3.3 Expected Results

As far as the project is concerned, we are targeting the outputs listed below with the intention of finding a balance between workable expectations and practical opportunities:

- Establishing a competence-based purpose-built cooperation which assembles and maintains language data and services;
- Adopting shared procedures and standardized data formats;
- Connecting institutions, companies, their data, services and search portals in integrated fashion and in view of relevant copyright and economic requirements;

[6] As mentioned in Section 1, CLARIN is but one of such projects. Nonetheless, given its exemplarity, it is here regarded as a touchstone to compare and contrast DI-ÖSS with well-established European infrastructures.

[7] The differences in terms of project scope and actors' participation have already been highlighted in Section 2.

- Providing a stable, centralized, flexibly expandable work environment;
- Restructuring redundant work steps and consequently reorganizing personnel responsibilities so as to recuperate misplaced resources and reinvest them.

The coalescence of these potentialities into an interoperable network is anticipated to lead to a collective advancement in terms of innovation. In fact, from a general perspective, DI-ÖSS ought to generate added value in form of increased and improved quality, efficiency, visibility and sustainability of the data and services offered. The overall character of these advantages outlines their versatility and malleability, i.e. leeway for the stakeholders to resourcefully mold or shape them into personalized benefits according to their institutional interests, needs and aims (for an exemplary overview of added value adaptation cf. Table 1).

Type of Stakeholder	**Added Value**
Libraries and archives	Use of computer linguistic tools and data processing stages
Research centers	Access to carefully documented collections of texts
Schools	Access to search portals with enhanced functionality
Language centres and cultural institutes	Access to carefully documented collections of texts and use of data analysis tools
Institutions and organizations responsible for education policy	Access to extensive studies and observations on the language situation in South Tyrol
Publishing houses	Use of automatic data analysis tools

Table 1: Overview of added value adaptation

Finally, the results of every project phase will be compiled into an evaluation report and a plan for a follow-up project aimed at building a comprehensive and sustainable digital language infrastructure.

4 Implementation of the project plan

The project, which was set in motion on January 1st 2017 and has been running since, is organized along a number of phases which build one upon the other. In the beginning, a small project consortium is built, which includes institutions from each of the most relevant target groups within the local language ecosystem (see Section 4.1). At the same time, detailed information regarding language data and services is collected from a wider set of institutions and organizations in South Tyrol (see Section 4.2). Afterwards, building on the previous phases' insights, specific use cases for each partner institution are determined and defined in greater depth. Along these use cases, a set of prototypical infrastructure components is built by implementing the technical setup required to connect partners and their data as needed by each use case. Finally, the infrastructure is piloted by employing it for each of the use cases and improving technical and conceptual aspects within the process (see Section 4.3).

4.1 Consortium

The project consortium was built during the first months of the project. In that initial phase the project partners were chosen carefully so that a range of different types of institutions as wide as possible could be represented in the consortium. As the DI-ÖSS project is aware of and embraces the diversity of relevant language institutions in South Tyrol the various relevant institution types had to be taken into account when considering potential partners for this pilot project and its envisioned infrastructure. The selection process has therefore been looking at a number of different institutions and companies, especially considering their approach to the development, distribution and preservation of language data and services. As explained in more detail later (see Section 5.2), due to its status as a pilot, the project can be perceived as abstract, thus making it challenging to communicate its objective and potential to the targeted partner institutions.

The final project consortium has been established between the following four institutions:

1. the *Institute for Applied Linguistics at Eurac Research* (project lead) as a research institution working with empirical language data and related language technologies,

2. the *Landesbibliothek Dr. Friedrich Teßmann* as a general-purpose library with a large digital collection of texts,

3. the *Sprachstelle* ("language unit") of the South Tyrolean Institute of Culture as a central institution for promoting the German variety of South Tyrol and informing the public about related matters,

4. the news and community portal *salto.bz* as a South Tyrolean publisher of daily news, local content and discussions around it.

4.2 Stocktaking - *Bestandsaufnahme*

The DI-ÖSS project, started of with the "Bestandsaufnahme" (literally *stocktaking*), which designates an initial work phase with a distinctively informational character. It involves collecting and thoroughly categorizing facts and details on the project partners and participating organizations with the intention of mapping the current state and gaining some insight into the nature of their data and *modus operandi*.

With this end in view, a questionnaire concerning five key aspects, i.e. general information on the institution or company in hand, its data collections, services, workflows and target groups, has been developed so as to cover the major fields of interest for the end users.

1. General information: on the one hand, it provides a global overview of the type of organization selected; on the other, it describes its specific needs and wants/intents and purposes while allowing each institution to become part of the infrastructure itself.

2. Collections: it presents the analogue (print) and/or digital data sets typical of each organization and expands upon them by differentiating between content and technical data. The former include criteria, such as a genre-based sorting of the material (principally fiction vs non-fiction), its amount and language of composition; the latter comprise parameters, like a medium-based sorting of the material (e.g. books, newspapers, journals, etc.), its format(s) and the software(s) used internally for working with it in the broadest sense. Moreover, this section contains indications as to which copyright terms and conditions apply for each collection.

3. Services: it labels the main, user-tuned, institution-specific services offered and briefly describes them, i.e. interlibrary loan, archive research inquiries, etc.

4. Workflows: where applicable, it sketches internal procedures concerned with data management, i.e. acquiring, processing and disseminating them, and service provision.

5. Target groups: it categorizes principal and secondary user groups, their approximate size and, if possible, how these typically make use of an institution's data and/or services.

The information collection process has so far taken place in the form of a recorded interview whose content is minuted at a later stage. In a second step, the key information is copied into a CMDI XML[8] document that is based on a profile adapted for the project's specific needs and is then fed into a modified VLO[9] where the data can be browsed via facets. These procedural steps have been carried out factoring in both the plausible future integration of the specifics into a larger CLARIN-like language resource infrastructure and especially the possibility to ingest parts of the data into the actual CLARIN VLO so as to make it more visible to the larger research community.

[8] https://www.clarin.eu/cmdi
[9] http://www.clarin.eu/vlo/

4.3 Use Cases

Specific use cases are identified for each project partner. The use cases are selected and defined in order to best comply with the following four aims:

1. Enhancing a task in the partner's daily workflow;
2. Exploiting a synergy (shared or complementary expertise) with at least one other partner;
3. Being applicable or easily adaptable for future or similar tasks;
4. Allowing to build a generic infrastructure interface for handling them

In the following subsections we will briefly depict the four use cases.

4.3.1 Use Case 1 - Teßmann library: browsing cultural magazines

Use Case 1 addresses the task of serving *enhanced search facilities for digitized content* to library users. It is built on a collection of cultural magazines from the 70s and 80s which contain written content and pictures, and follow an irregular layout structure (e.g. paragraphs of texts are blended with images, articles run over several – not always consecutive – pages, etc.). Content-wise the magazines combine cultural reviews, lyrical and poetic contributions, portraits of artists and artworks as well as announcements, manifests and reports. Accordingly, readers would prefer to browse them based on recurrent themes, figures or also concepts. For example, a library user might want to find all articles related to one artist or a prevalent topic of discussion. In addition, locations, time periods or arts genres might be themes of particular interest.

Approach

The delivery of enhanced search facilities is approached in three steps. First, the digitized texts are processed and annotated for information of interest, such as thematic keywords, persons, locations, arts genres, etc. Second, the individual articles of the cultural magazines as well as smaller text snippets are interlinked, using the annotated information (step 1). Third, a search interface that allows browsing of the cultural magazines based on the presented concepts and their interlinking is created. For example, the interface offers the user access to related articles grouped together and navigation along related concepts or interconnected persons and themes.

Synergies

In order to implement the use case, computational linguists at Eurac Research are closely collaborating with experts in literature and cultural studies at the Teßmann library. The literary and culture study experts work on identifying themes and aspects of relevance and clearly describe their informational needs directed toward the cultural magazines. Based on these pointers, the computational linguists select and apply NLP tools to automatically detect and annotate these types of information in the texts. Finally, in close collaboration they design and implement an interactive interface for searching the texts based on the advanced textual cues.

The use case is non-specific and transferable to the extent that a generic toolchain for annotating digital texts is put into place, and a search interface that builds on the annotated text formats is created so as to access and display segments of text (magazine articles or smaller paragraphs).

4.3.2 Use Case 2 - salto.bz: enhanced tagging of articles

Use Case 2 is concerned with the task of *improving search and discoverability* for the readers of the online news portal *salto.bz*.

Currently, the portal offers readers the built-in search of their CMS, which only performs simple string matching and no ranking of the search results. This makes finding interesting content much more difficult for the readership, both targeted search and browsing by being offered similar articles are not very efficient at the moment. Most readers will likely browse through the various sections without looking for something specific, they could therefore benefit from articles being more closely interlinked so that

related articles could be offered automatically, giving more background or a different view on a subject. But also the targeted search is a valid use case for readers of a news site. Ideally, it would be possible to limit search results to a specific news section (politics, sports, local news, etc.) and also to a specific time-frame.

Approach

The problem is approached in a number of separate, but closely interrelated steps. All the existing articles will undergo a semantic analysis to extract the most relevant keywords and the same will be integrated into the editorial user interface so that the mechanism can also be triggered for newly written articles. At the same time, the semantic analysis will automatically identify related articles and link them directly to each other. This newly generated deeper information about the content of the articles will then enable a much more user-friendly search interface to be implemented. One especially challenging part of this endeavor is the fact that the news portal in question is bilingual in scope. Both articles in German and Italian are being published and a well-designed search has to take into account that the readers would like to find results in both of the languages no matter which language they are using to search. This means whether someone is searching for *elezioni* or *Wahlen* they should obtain the same set of search results.

Synergies

The implementation is taking place at two different ends of the use case. While computational linguists are working on the backend and creating web services that are able to automatically extract keywords and compare articles for relatedness, the editors and technicians of salto.bz will deal with the frontend. The editors are checking and approving the automatically generated keywords and relations between the articles. At the same time, the salto.bz developers will adapt the user interface to make the newly generated keywords accessible to the readers and create an improved search interface that makes use of the additional information on the articles that is now available. As the web services that offer the linguistic services will be implemented using a generic API, they could in principle be used by other interested parties at a later stage.

4.3.3 Use Case 3 - Eurac Research: crowdsourcing of historical letters

Use Case 3 is a cooperation between two institutes of Eurac Research, the Institute for Applied Linguistics (IAL) and the Institute for Minority Rights (IMR). The IMR has been collecting missives (mostly letters and postcards) from the inhabitants of South Tyrol to create a representative corpus of historical letters spanning the 20th century. Within this use case, the still-growing collection of mostly handwritten letters and postcards will be enriched with structured metadata and transcribed by the local population using online crowdsourcing tools. During the crowdsourcing phase, the public will also be involved through public events and the envisioned end result is both a well-curated digital collection and a highly engaged and passionate (part of the) public.

Approach

There is already a huge collection of some twelve thousand missives that have been digitized as pictures, while more material is still being collected. In a first step, the data will be uploaded into an instance of the crowdsourcing software Pybossa where volunteers can extract the metadata (sender, addressee, date, etc.) from each item which is then stored in a structured machine-readable format. After the metadata have been extracted, the missives can be grouped into related collections (e.g. based on location or time) and be ingested into a web-based annotation software. The project is using the web version of Transkribus[10] to crowdsource the transcription.

Synergies

The Institute for Minority Rights is collecting the missives from citizens all across South Tyrol and takes a first step of digitizing them. The whole technical setup of the various crowdsourcing tools is handled by experts at the Institute for Applied Linguistics while the design of the user interface and

[10] `https://transkribus.eu/r/read/projects/`

the accompanying texts that guide the users are jointly developed by both institutes. This collection of historical letters is a great opportunity for such a shared project because the content, while being very interesting from a historical perspective on the eventful 20th century in South Tyrol, also offers an insight into a unique linguistic situation where writers often switch between standard German and their local dialects, sometimes in the same letter. And especially after the annexion of the territory by Italy the language mix that can be found in the texts also includes Italian. Additionally, during the war periods we can find a lot of letters by authors that are not very used to writing longer texts, which also promises to yield some interesting analyses.

With the IMR's focus on making this part of their cultural heritage available to the population of South Tyrol, all data (as far as the obvious privacy concerns allow) will be made freely available on the internet so that also hobby scholars and citizen scientists can use it for their own studies. It is also expected that the experience both institutes will gain in the area of crowdsourcing will be beneficial to future projects.

4.3.4 Use Case 4 - *Sprachstelle*: identifying regional neologisms

The final Use Case is a project aiming at installing an infrastructure for finding and identifying neologisms that are specific to the region of South Tyrol. For this use case, computational linguists automatically harvest South Tyrolean sources on the internet to propose candidates for such neologisms and experts on the local variety of German at the language unit (*Sprachstelle*) of the South Tyrolean Institute of Culture will verify which of those candidates can actually be seen as potential regional neologisms. This feedback will then be used to fine-tune the automatic detection to minimize the amount of manual work that has to be done by language experts further on.

Approach

The starting point for the use case is a carefully curated list of South Tyrolean media that publish original texts online. Among those are web sites of newspapers and local TV and radio stations, but also personal or semi-professional websites or sites that provide information from the local government. This list of websites is then regularly crawled and the resulting word list is checked against a list of standard German words to eliminate known forms. After trying to automatically eliminate as much as possible also errors resulting from the known error-prone process of crawling HTML pages, the list of candidates for possible neologisms is then checked by experts on the local variety of German to determine if a candidate is a neologism and if so, if it is specific to the linguistic variety spoken in South Tyrol. The edited list is then fed back into the algorithm that selects the candidates from the web crawl resulting in a continually improving selection process.

Synergies

The Institute for Applied Linguistics at Eurac Research has implemented a first version of the software that crawls the web and selects possible neologism candidates, called *Styrlogism*. First editing rounds have already carried out internally with experts on South Tyrolean German from the institute itself. Within this use case, this will now be complemented by the expertise coming from the language unit at the South Tyrolean Institute of Culture. As the main task of the language unit is to inform and educate the population about the local variety of German, they can use the results from this process to showcase the newly detected South Tyrolean neologisms in their public relations work. In this presentation, it is often possible to show the original context of this discovery because the Styrlogism tool keeps the whole environment of the detected words and also always stores the originating website. If this has not been taken down by their authors in the meantime, interested users can then even go back to the original source to see the new word in the complete context.

5 Discussion of encountered challenges

Even though conceptually aligned with established large-scale infrastructure initiatives like CLARIN, the actualization of the locally-oriented DI-ÖSS language infrastructure is a step into uncharted territory. Especially the fact that the language partners involved and approached within the course of the project

are relatively small, have few resources at their disposal and possess limited experience with large-scale projects has proven to pose particular challenges, which could not be anticipated to the extent encountered. The DI-ÖSS project is devised as a pilot project that is specifically designed to find the unique challenges inherent in such a local infrastructure. The goal is to learn from the prototypical phase and use it as a facilitator to establish a comprehensive and powerful digital language infrastructure in South Tyrol in the mid to long term and take steps to integrate it as much as possible with larger infrastructures like CLARIN and DARIAH.

Having said that, we will close this article with a discussion of some of the most prominent challenges that we have faced over the course of the project up to now. They concern conceptual, communicative and technical aspects, as laid out in the following three subsections.

5.1 Conceptual challenges

Already when creating the consortium, but especially later when interviewing potentially interesting institutions for the *Bestandsaufnahme*, it became apparent that while everything can be considered potentially interesting linguistic data – from the protocols of the province offices to advertisements of a local company – DI-ÖSS has to tighten its scope to more obvious "language institutions" like libraries, publishing houses and linguistic research centers (see Section 4.1). Generally, the focus was reduced to institutions that 1) deal with language data produced in South Tyrol, 2) consider working with language data their main activity, and 3) work with data that are available digitally, either digitized or born digital. It was also decided to explicitly involve smaller actors that do not already have visibility and power in the South Tyrolean ecosystem in order to make the resulting infrastructure more of a democratic place. Additionally, there was a problem with clearly defining some of the use cases. On the one hand, a very deep understanding of the workings of an institution is fundamental to see whether there are specific needs; on the other, a wide knowledge of the other partners is essential to determine which possible solutions there could be to those problems. This was only solvable by taking the time to have long discussions with each project partner to fully understand their needs and capabilities.

5.2 Communicative challenges

Communicative challenges arose in the process of getting institutions interested in joining the project as it has proven difficult to properly communicate the scope and purpose of it. As described in Section 3.2, the infrastructure can be theoretically seen as a fluid assemblage of actors, which influences the South Tyrolean culture and identity, and sets the basis of a South Tyrolean digital cultural heritage. Translating these theories into graspable concepts for the possible partners is certainly a challenge. It helps to use metaphors of physical infrastructures, like the railway system, and also to focus on concrete use cases early on, so that it becomes easier for the potential partners to see their specific role within the project.

It is necessary to address every possible partner institution with a different approach, trying to anticipate their needs and reservations. While libraries and other public institutions are more readily willing to share their data freely, commercial actors, e.g. publishing houses, are often very protective of their data as this is central to their business model. But even once a potential partner sees the benefits of the project, there are still further issues. Because of the small size of many of the potential partners, there might not be enough resources available that they could bring into the project. Especially, if there is no obvious short-term benefit for the partners, it becomes difficult to justify spending some amount of their often quite limited human resources on this project.

5.3 Technical challenges

This is another point where this small-scale infrastructure differs considerably from its larger counterparts. Many language partners in DI-ÖSS have very limited resources, both on the personnel and on the IT side, so it usually is difficult for them to implement large changes in their data management infrastructure or their typical workflows, while this is more feasible for the bigger institutions involved in infrastructure projects like CLARIN. This means the DI-ÖSS infrastructure has to be constructed in such a way that it integrates the needs of an infrastructure (standardized data formats and APIs) with the

existing working realities, which often involve suboptimal or home-grown solutions that cannot be easily changed or adapted.

References

Gilles Deleuze, PF Guattari, and Felix Guattari. 1987. *A thousand plateaus: Capitalism and schizophrenia*, volume 19. University of Minnesota Press.

Jennifer Edmond, Frank Fischer, Michael Mertens, and Laurent Romary. 2017. The dariah eric: Redefining research infrastructure for the arts and humanities in the digital age. *ERCIM News*, (111).

Europeana Foundation. 2015. Transforming the world with culture: Next steps on increasing the use of digital cultural heritage in research, education, tourism and the creative industries. Technical report, Europeana Foundation, September.

Steven Krauwer and Erhard Hinrichs. 2014. The clarin research infrastructure: resources and tools for e-humanities scholars. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pages 1525–1531. European Language Resources Association (ELRA).

Bruno Latour. 2005. *Reassembling the social: An introduction to actor-network-theory*. Oxford university press.

Jan Odijk. 2016. Clariah in the netherlands. In *LREC*.

A PID is a promise
Versioning with persistent identifiers

Martin Matthiesen
CSC – IT Center for Science
Espoo, Finland
martin.matthiesen@csc.fi

Ute Dieckmann
University of Helsinki
Helsinki, Finland
ute.dieckmann@helsinki.fi

Abstract

We present the update process of a dataset using persistent identifiers (PIDs). The dataset is available in two different variants: for download and via an online web interface. During the update process, we had to fundamentally rethink as to how we wanted to use PIDs and version numbering. We will also reflect on how to effectively use PID assignment in case of minor changes in the large dataset. We discuss the roles of different types of PIDs, the role of metadata, and access locations.

1 Introduction

While other disciplines have been affected by reproducibility concerns as described in Baker (2016), this has so far not been the case in the Humanities. With the increasing use of statistical methods and automated data processing in the Digital Humanities and Computational Linguistics, this is likely to change and manifestos such as Munafò et al. (2017) will become more relevant.

Making data available in a persistent manner is one important aspect of making a dataset reusable for further research, but is also important for reproducibility of existing research. Publication principles such as FAIR (Wilkinson et al., 2016) emphasise the importance of persistent identifiers (PIDs) and descriptive metadata.

In an abstract sense, the role of PIDs is very clear: "Persistent identifiers allow different platforms to exchange information consistently and unambiguously and provide a reliable way to track citations and reuse." (Rueda et al., 2016, 40). In the same article, the authors warn: "Low-quality metadata, uncurated content, and a lack of internal and/or external organisation create repositories that are impossible to navigate or to obtain information from." (Ibid., 41).

Using PIDs consistently to avoid the aforementioned pitfalls turned out to be complex. In this paper, we explore in detail what using PIDs and descriptive metadata records means in practice when updating a large dataset.

The paper addresses the following areas in the design and construction of a CLARIN infrastructure:

- Recent tools and resources added to the CLARIN infrastructure
- Metadata and concept registries, cataloguing and browsing
- Persistent identifiers and citation mechanisms
- Web applications, web services, workflows
- Models for the sustainability of the infrastructure, including issues in curation, migration, financing and cooperation

Martin Matthiesen and Ute Dieckmann 2019. A PID is a Promise - Versioning with Persistent Identifiers. *Selected papers from the CLARIN Annual Conference 2018*. Linköping Electronic Conference Proceedings 159: 103–112.

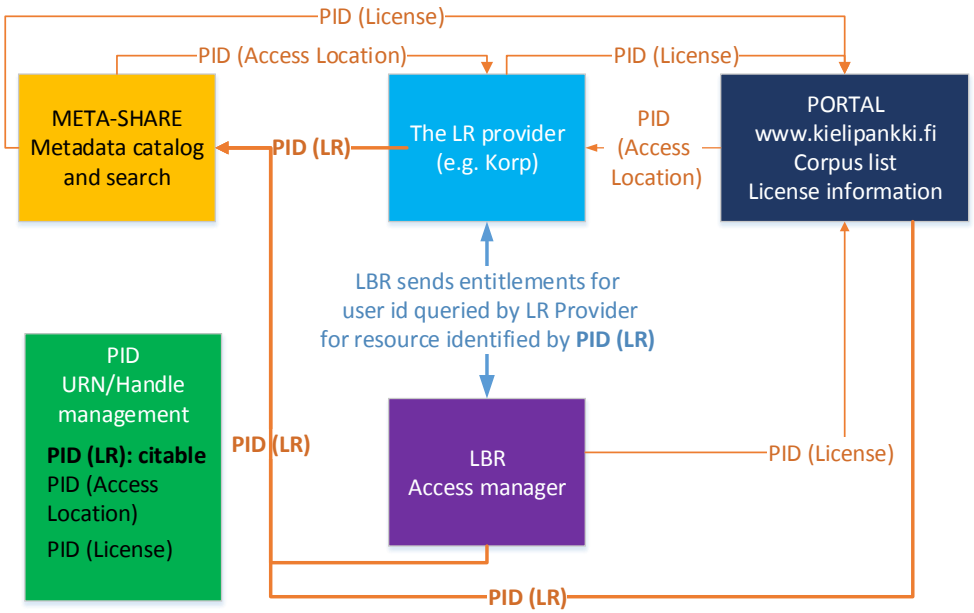


Figure 1: Resource and access management using PIDs

2 The repository

The Language Bank of Finland is a CLARIN B Centre and has therefore all the tools in place to provide data in a FAIR manner.

- A catalog for descriptive metadata using META-SHARE (http://metashare.csc.fi)
- A PID registry providing Handles and URNs.
- A download service (https://korp.csc.fi/download/)
- Corpus analysis tools such as Korp[1] (https://korp.csc.fi/)
- Access management via REMS[2] (Language Bank Rights, https://lbr.csc.fi/)
- Generated citation instructions of resources[3]

PIDs are used to reference language resources and implement access management. Figure 1 gives a general overview over the interaction of the components mentioned above.

In this paper, a resource consists of two main parts: the descriptive metadata in META-SHARE and the data itself. The role of the manually curated descriptive metadata (henceforth: metadata[4]) is to give the resarcher the "context information"[5] of the data itself.

We use the PID pointing to the META-SHARE metadata of a resource as the ID of the given resource. This PID (referenced as *PID(LR)* in figure 1) is citable and used in all services, such as Korp, Download, Language Bank Rights, and our corpus list to identify the resource. This citable PID is in fact the only essential PID needed to publish the dataset. The distinction between citable and non-citable PIDs is discussed further in section 7.

[1] Borin et al. (2012)
[2] See Linden et al. (2013) in Foster (2013)
[3] See "cite" column in our corpus list: https://www.kielipankki.fi/corpora/
[4] What we call a descriptive metadata page is sometimes referred to as "landing page" (See `https://documentation.library.ethz.ch/display/DOID/Landing+pages`.) We avoid the term in this paper.
[5] See Weigel et al. (2013).

At the Language Bank, PIDs are minted manually using a simple csv file as source in Github. Uniqueness is ensured by using the date of minting in reverse and a running number:

```
# Example
201801011 http://example-url.com
```

A script then registers URNs as well as Handles[6]. Handle attributes are not used for two reasons: compatibility with URNs, which do not support attributes, and the increased complexity of keeping the metadata in the attributes up-to-date and in sync.

3 The dataset

The dataset named "Corpus of Finnish Magazines and Newspapers from the 1990s and 2000s" was originally intended as an accruing dataset and therefore not versioned. The goal was to create a contemporary dataset of magazines and newspapers of various origins, such as scientific journals, regional newspapers, company internal circulations, and trade union member journals.

It was available for download licensed as CLARIN ACA +NC[7] and in Korp, licensed as CC BY[8].

We had preliminary policies in place for versioning accruing datasets[9]. In this case, we assumed that we would not need strict versioning for this dataset because changes seemed transparent enough for us as well as the research community. Over time it became apparent that using versioning is nevertheless more transparent than trying to avoid it.

3.1 Creation of the dataset

For this corpus, the original data was mostly harvested (partly automatically with the help of a python script) from the internet in PDF format. The PDF was converted to plain text with OCR software. These PDF and text files are available in our download service. For legal reasons, in a few cases we cannot provide the original PDFs.

The text files were then converted to the Corpus Workbench VRT format[10] using Python scripts. Structural attributes carrying metadata information, such as the name of the magazine, issue and date, were added. Finally the VRT data was enriched with dependency information, part-of-speech and named entity tags using the Turku Dependency Parser[11], and an earlier version of Finnish Tagtools[12]. This enriched VRT data was imported into Korp.

4 The initial update process

Even though the dataset consists of various individually identifiable newspapers and magazines, we had assigned only four PIDs to refer to the variants of the entire dataset: one PID to refer to the metadata of the Korp variant, one PID to refer to the metadata of the downloadable variant, and another two PIDs to point to the access location of the data itself, in Korp and our download service ("Download"), respectively.

Initially the dataset was updated as stated in the metadata and outlined in figure 2: frequently and without changing the PID. Information on the updates of each variant was maintained on a separate wiki page, referenced from the metadata. The metadata did not specify the update process of the variants. Korp and Download were not updated synchronously. Sometimes Korp would get updates before Download, more often it was the other way around.

[6] URNs and Handles can be derived from one another, this method developed at the Language Bank is now part of official GEDE/RDA recommendations, see assertion *PID-45* in Wittenburg et al. (2017, Section 3.3).

[7] `http://urn.fi/urn:nbn:fi:lb-2016050602`

[8] `https://creativecommons.org/licenses/by/4.0/`

[9] See The Language Bank's *Life cycle and metadata model of language resources*: `http://urn.fi/urn:nbn:fi:lb-201710212`

[10] See `https://www.kielipankki.fi/development/korp/corpus-input-format/` for a more detailed description.

[11] `http://turkunlp.github.io/Finnish-dep-parser/`

[12] See University of Helsinki (2018).

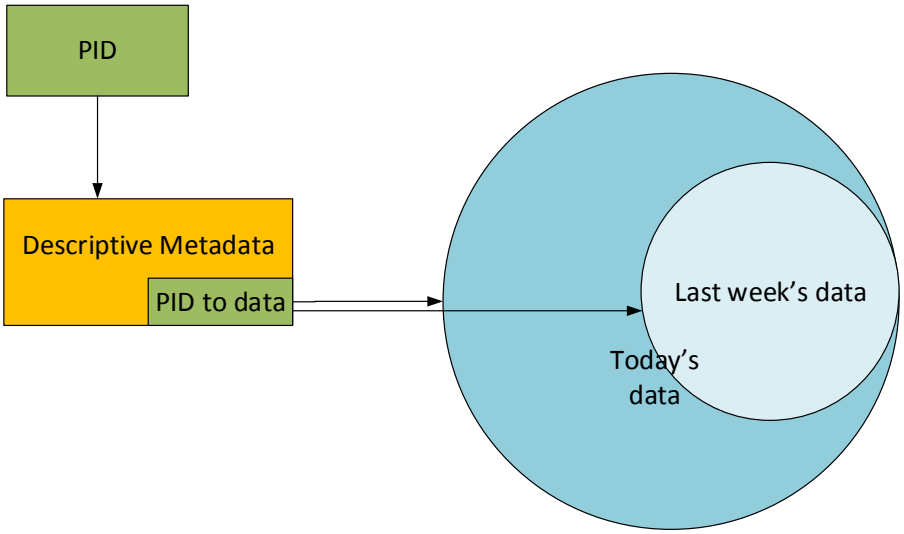
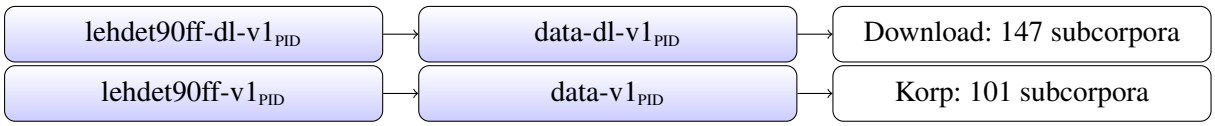


Figure 2: Corpus variant with an unversioned PID

The PID pointing to the top level directory of the Download variant of the dataset would automatically include any added content. This was not the case with Korp. The PID pointing to the Korp collection was not consistently updated, explicitly selecting only parts of the dataset for use in Korp.

However, even the extended versions of the dataset were implicitly addressed, since new subcorpora showed up as additional selectable items under the same collection in Korp. At the time of the update, we had a corpus of 147 subcorpora in Download and 101 subcorpora in Korp. Figure 3 shows how the citable PID points to the metadata and the metadata in turn points to the access location of the actual data.



lehdet...$_{PID}$: A simplified persistent identifier pointing to the descriptive metadata
data...$_{PID}$: A simplified persistent identifier pointing to the data itself
Korp,Download: The access location of the data itself

Figure 3: PIDs and access locations for version 1 before the update

During the update, we discovered issues in both dataset variants:

- Some subcorpora in Korp and Download were missing data, due to previously unnoticed problems with the conversion.
- Some Korp subcorpora were not properly annotated.
- Some Download zip files did not have license and README information.
- Existing README/license.txt files were located in the root path of zip files, and they were overwritten if more than one zip file was unzipped in the same directory.
- The directory structure of the zip files was generally not consistent.
- Files zipped on a Mac had filename encoding problems in Linux.

- Some zip files contained thumbnails and other irrelevant temporary files/directories.

In other words, an update planned as a simple addition of data turned into the curation of an already published dataset.

5 A more consistent approach

At the time of the update, the dataset was by design unversioned. The variants in Korp and Download were not synchronized, and existing data in both variants needed to be curated. We essentially faced a versioning task, as described in appendix A3 in Weigel et al. (2015, 21). The decisions we made are explained below.

5.1 Versioning

First, we abandoned the idea of an accruing dataset behind a single PID. It is clear what the PID denotes at any given point in time, and its general intension stays the same. Determining the concrete extension of such a PID at different times is possible, but impractical and error prone. We therefore now take the temporal component into account and introduce versioning as shown in figure 4.

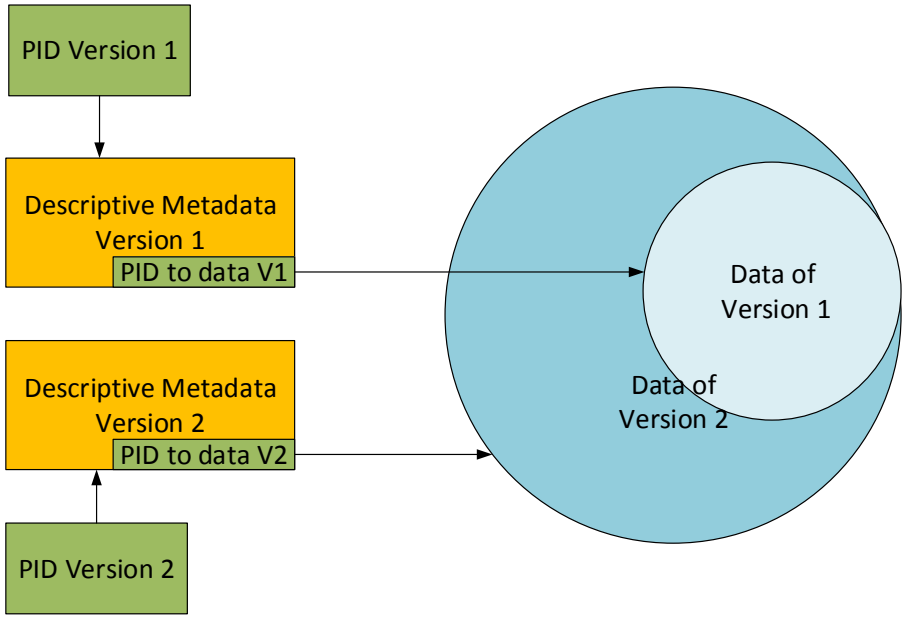


Figure 4: Corpus variant with a versioned PID

Versioning for small datasets in textual format is largely solved if they do not significantly exceed the size of a computer program[13]. Version control for computer software like Github[14] has been around for a long time. In our case, we could not use it for many reasons, the most important one being the size of the dataset. We had to invent our own approach.

As the two variants (Korp and Download) of the dataset were not in sync, we briefly considered synchronizing them to have a well-defined starting point for the update. We abandoned this idea, since it would require even more PIDs and version numbers. So we accepted that we did not have a well defined starting point and aimed for a well-defined end state.

[13] To our knowledge only the *Hamburger Zentrum für Sprachkorpora* uses *git* for versioning of text corpora, see `https://inel.corpora.uni-hamburg.de/wp-content/uploads/jettka_hedeland-2018-HZSK_INEL_Workflows.pdf`

[14] `https://github.com`

We therefore introduced version 1 of each variant and made it explicit that they are overlapping but not absolutely in sync. The updated and synchronized dataset, version 2, now contains 369 subcorpora in either of its two variants.

5.2 Stop-over pages

We use PIDs for metadata resolution and resource resolution, as defined in Weigel et al. (2013). In our case, the Korp variant of version 1 was not worth keeping online unchanged. For example, some subcorpora lacked part-of-speech information in version 1 that was added in version 2, but the content was otherwise unchanged. A search performed on version 1 can thus be repeated with version 2 by simply ignoring the part-of-speech information. In other cases, attributes were renamed. Again, the old search could be repeated by slightly modifying it to work with version 2. Since we had quite a few such changes, we decided not to keep version 1 online, and instead point the resource PID of version 1 to the relevant subset of version 2. To make the changes transparent, we did not point the PID directly to the resource, but to a "stop-over page".

A "stop over page" is a manually curated web page accessed by a resource PID that has pointed to data which is not available in its original form any longer. The changes are explained and the user is directed further to the location of the corrected data, as outlined in figure 5. The stop-over page either gives access to the previously available data or it provides information on how to use the updated data to get comparable results. A stop-over page shares properties with a tombstone page. Both refer to data not directly available anymore. A tombstone page is used when it is hard or impossible to recreate the old data.

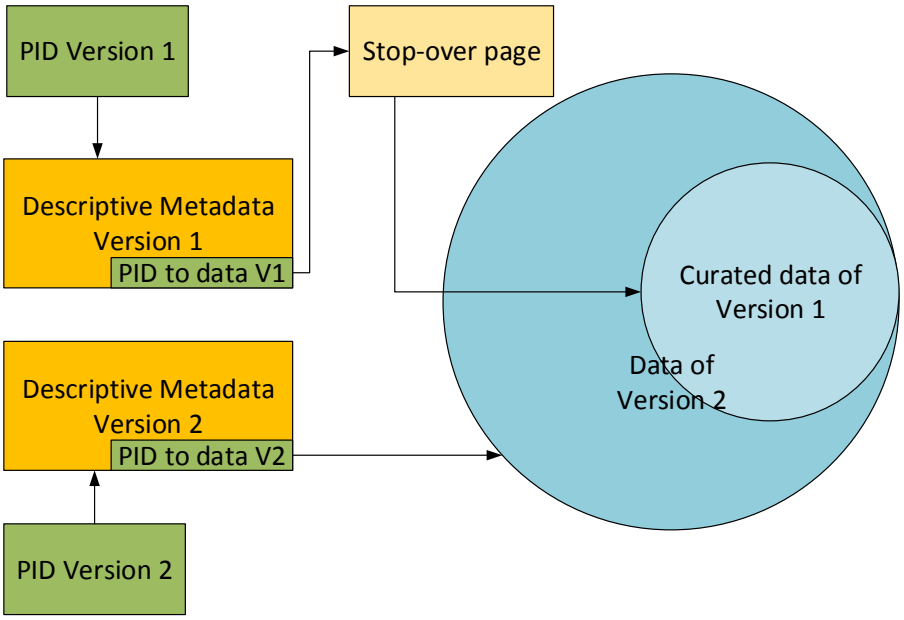


Figure 5: Corpus variant with a versioned and curated PID

The stop-over page makes it possible to take old data offline without excessively compromising reproducibility. This not only saves storage and backup space, but also improves accessibility and structure of the repository as a whole.

5.3 Change Log

The Change Log is a well known concept in software engineering[15]. We use it to describe "minor changes" in the dataset that we do not consider worth a version change. For example[16]:

```
Change Log:
20.2.2018 Name and file type of "tiedelehdet_terminfo.vrt.gz" changed to
          "Terminfo 2010-2015.zip" (subdirectory added in zip file)
```

The Change Log is kept in the metadata record of the dataset and only relates to the version at hand, unlike software changelogs that often contain the whole version history. In our case, that history is kept using relations between metadata records. For those relations (e.g. *IsPreviousVersionOf*), we use a subset of the controlled vocabulary described in DataCite Metadata Working Group (2016).[17]

5.4 PID granularity

During the update we also considered changing the granularity of the PIDs. In the following sections, we explain why we did not opt for increasing the amount of PIDs.

5.4.1 Rejecting data object PIDs

We evaluated the introduction of data object PIDs. The CLARIN B Centre Requirements state that data objects can be assigned a PID if they "are considered to be worth to be accessed directly (not via metadata records) by the data provider" (Wittenburg et al., 2018, Section 7).

Such data object PIDs are obviously useful for machine to machine communication, they can be accessed by scripts and automatic processing pipelines. However, this is of practical use only for small datasets. It is fair to assume that large datasets will hardly ever be processed online, but rather downloaded, decompressed and processed locally. Data object PIDs are of little use in such a scenario.

In our data curation task we had to make at least minor changes to all subcorpora in Korp and Download. Version 1 of the downloadable corpus (University of Helsinki, 2017) alone is a collection of 147 subcorpora consisting of 413 zip files and tens of thousands of individual files.

Had we assigned 413 PIDs to the zip files, most of them would have needed stop-over pages, because we changed the content of the zip files by adding READMEs and subdirectories, correcting typos in filenames, and so on. It would not have been feasible to keep the old zip files online. Any script relying on the PIDs would have stopped working at this point. Even if the stop-over page had been machine readable, the end result would have been that the old zip file would not have been provided automatically.

PIDs to individual files would have required us to provide the content either uncompressed or compress the files individually and would have created a need for even more stop-over pages. Storage and bandwidth considerations also had to be taken into account. Apart from the higher maintenance need for hundreds of PIDs, we did not see an added value for a user using only a subset of the corpus. The subset can still be defined relative to the dataset variant referenced by the PID.

Instead, we use one PID for the Download variant and explain the changes in the metadata in a Change Log. We also maintain a Change Log in the metadata and a stop-over page to explain the changes we made to the already published subcorpora in Korp. We used the stop-over page only for the Korp variant, since we considered the changes in Download minor enough to be described in the Change Log.

5.4.2 Adequate PID granularity

In the previous chapter, we argued that as few PIDs as possible should be used. As shown in figure 6, we are down to four, two per dataset variant, two for each version. Why not even less, why not use one PID for both variants, reducing the number of PIDs further to two, one for version 1 and one for version 2?

The Prague Dependency Treebank is published in this way: The PID points to the metadata from where the data can be dowloaded or accessed with two distinct web based tools[18]. We considered two use cases

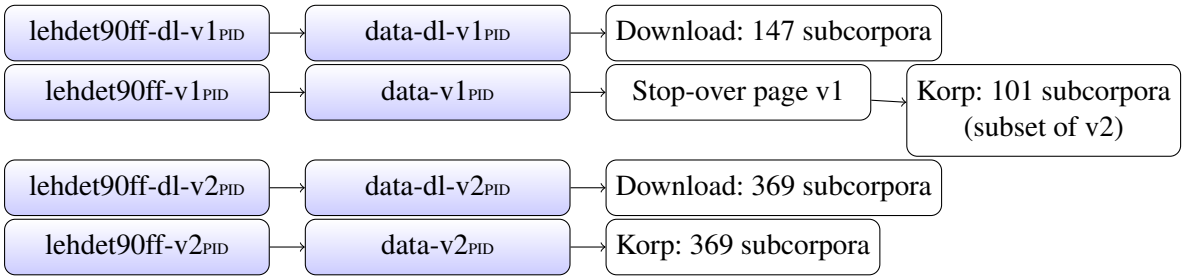
[15] https://www.gnu.org/prep/standards/standards.html#Change-Logs

[16] See `http://urn.fi/urn:nbn:fi:lb-2017091902`

[17] See The Language Bank's *Life cycle and metadata model of language resources*: `http://urn.fi/urn:nbn:fi:lb-201710212`

[18] Cf. (Hajič et al., 2018).

and decided that both warrant separate PIDs: If a user uses only one instance of a corpus, for example only the downloadable version or only the Korp web interface, then there will be only one reference in any case. Should a user use both variants in one paper, different PIDs make it easier to keep track of what has been observed and where. After all, it is always possible that the variants differ in unintended ways.



lehdet...PID: A simplified persistent identifier pointing to the descriptive metadata
data...PID: A simplified persistent identifier pointing to the data itself
stop-over page: As described in section 5.2
Korp,Download: The access location of the data itself

Figure 6: PIDs and access locations after the update

6 Generated PIDs vs. manual PID curation

Automatic management of PIDs has a few advantages: Links and relations are created in a consistent manner. Data changes can be detected automatically and new PIDs can be instantly created, if needed. This is especially important when dealing with a large number of PIDs.

However, the automatic approach cannot easily distinguish between significant and non-significant changes, as for example adding a missing comma in the README.txt file of a downloadable dataset. Keeping the old dataset in this case and minting a new PID makes no sense, at least not in terms of reproducibility or responsible usage of storage space. Changing a character in the tagset of a corpus can be a minor or major change. The computer cannot yet categorize changes and more importantly cannot substantiate and justify such categorizations; this still needs to be done by humans.

The tools we use, such as META-SHARE do not support automatic PID handling. Minting them manually as described in section 2 gives us more flexibility in using them. It does, however, also leave room for inconsistencies, as discussed with inconsistent updates of data PIDs in section 4.

While fully automatic PID handling is not desirable, automatic checking of existing PIDs and their relations would help to ensure more consistency. The usability of META-SHARE would also benefit from better support for PIDs and expressing their relations using controlled vocabularies, as suggested by DataCite Metadata Working Group (2016).

7 Citable vs. non-citable PIDs

We divide PIDs into two major categories, regardless of the underlying PID resolver technology (eg. URN, Handle, DOI): Citable and non-citable. Citable PIDs point to the authoritative metadata of the resource, and therefore are absolutely essential properties of a dataset. It cannot be published without them being assigned. Non-citable PIDs can be used to refer to the access location of the data itself. Non-citable PIDs are not absolutely necessary since the dataset can always be referenced using its citable PID. They can be further subdivided into access location and data object PIDs. While not essential, they can be useful to manage changes in access locations, such as server name changes.

- Citable PID: PID to authoritative metadata of a resource
- Non-citable PID
  - Access location PID (points to the actual data location in services such as Korp or Download)

- Data object PID (directly points to data object, like zip, pdf, wav, mp4)
- PIDs to license pages

Note that stop-over pages are also useful in scenarios where only citable PIDs are used. In that case, the direct access location link to a dataset is replaced by a link to a stop-over page leading further to the updated dataset access location.

8 Discussion and Conclusions

Our aim was to update two variants of a previously unversioned dataset in a way that enables researchers to replicate earlier studies. Transparent information should be provided on any deviations within each version.

We created our own approach to versioning. In section 5.4.1, we showed that it is often not practical to keep earlier versions of large datasets available online. Taking a dataset offline immediately breaks automatic workflows. It would also break data object PIDs, which is one reason why we consider them impractical.

We showed that the inflationary automatic creation of PIDs (usually to data objects) considerably increases curation needs. The consequences in terms of human and technical resources can be significant. By making a clear distinction between mandatory citable and optional non-citable PIDs, we offer a way to keep the focus in PID handling.

A PID is, not unlike a bank note, a promise. Once you create it, you have to make sure it keeps its value.

Also not unlike currency, different people see different values in PIDs. In our opinion, the core value of a PID is the ability to make datasets traceable, even if they change over time and older versions are not available online anymore.

Our aim is not to ensure 100% repeatable runs of scientific software over the span of many years. Our aim is to enable plausible repeatability of research. Such repeats might require changes in the original scientific code or web request to produce similar outputs. To make possible changes transparent we introduced a Change Log and the new concept of stop-over pages.

To sum up, when introducing versioning using PIDs, we tried to find a balance between maintainability, usability, and transparency at every stage of the update process of a large dataset.

9 Outlook

While we do not want to maintain data object PIDs, a direct path to a data object via a general PID is useful. The transparent implementation of part identifiers for URNs and Handles[19] would be a solution to this problem. Automatic validation of manually created PIDs and their relations is another area of improvement. META-SHARE could work more with controlled vocabularies and warn of missing back-references in case of reciprocal relations. The efficient storage, versioning, and dissemination of large binary datasets continues to be a challenge. We intend to evaluate efforts like the German KA3 Project[20] for its applicability to our data management needs.

References

Monya Baker. 2016. 1,500 scientists lift the lid on reproducibility. *Nature*. https://doi.org/10.1038/533452a.

Lars Borin, Markus Forsberg, and Johan Roxendal. 2012. Korp the corpus infrastructure of Språkbanken. In *Proceedings of LREC 2012. Istanbul: ELRA*, page 474478.

DataCite Metadata Working Group. 2016. Datacite metadata schema for the publication and citation of research data v4.0. page 37ff. https://doi.org/10.5438/0012.

[19] See assertion *PID-45* in Wittenburg et al. (2017, section 3.3)

[20] http://dch.phil-fak.uni-koeln.de/ka3.html, in German

David Foster, editor. 2013. *Innovating Together, The 29th Trans European Research and Education Networking Conference, 3 - 6 June, 2013, Maastricht, Netherlands, Selected Papers*. TERENA, August. http://www.terena.org/publications/tnc2013-proceedings/.

Jan Hajič, Eduard Bejček, Alevtina Bémová, Eva Buráňová, Eva Hajičová, Jiří Havelka, Petr Homola, Jiří Kárník, Václava Kettnerová, Natalia Klyueva, Veronika Kolářová, Lucie Kučová, Markéta Lopatková, Marie Mikulová, Jiří Mírovský, Anna Nedoluzhko, Petr Pajas, Jarmila Panevová, Lucie Poláková, Magdaléna Rysová, Petr Sgall, Johanka Spoustová, Pavel Straňák, Pavlína Synková, Magda Ševčíková, Jan Štěpánek, Zdeňka Urešová, Barbora Vidová Hladká, Daniel Zeman, Šárka Zikánová, and Zdeněk Žabokrtský. 2018. Prague dependency treebank 3.5. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. http://hdl.handle.net/11234/1-2621.

Mikael Linden, Tommi Nyrönen, and Ilkka Lappalainen. 2013. Resource Entitlement Management System. In Foster (Foster, 2013). http://www.terena.org/publications/tnc2013-proceedings/.

Marcus R. Munafò, Brian A. Nosek, Dorothy V. M. Bishop, Katherine S. Button, Christopher D. Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J. Ware, and John P. A. Ioannidis. 2017. A manifesto for reproducible science. *Nature Human Behaviour*, 1:21, Jan. http://dx.doi.org/10.1038/s41562-016-0021.

Laura Rueda, Martin Fenner, and Patricia Cruse. 2016. Datacite: Lessons learned on persistent identifiers for research data. *International Journal of Digital Curation*, 11(2). https://doi.org/10.2218/ijdc.v11i2.421.

University of Helsinki. 2017. Corpus of Finnish Magazines and Newspapers from the 1990s and 2000s, Downloadable Version 1. The Language Bank of Finland. Retrieved from http://urn.fi/urn:nbn:fi:lb-2016050401.

University of Helsinki. 2018. Finnish Tagtools. The Language Bank of Finland. Retrieved from http://urn.fi/urn:nbn:fi:lb-2018062101.

Tobias Weigel, Michael Lautenschlager, Frank Toussaint, and Stephan Kindermann. 2013. A framework for extended persistent identification of scientific assets. *Data Science Journal*, 12:10 – 22. https://doi.org/10.2481/dsj.12-036.

Tobias Weigel, Timothy DiLauro, and Thomas Zastrow. 2015. PID Information Types WG final deliverable. Technical report, Research Data Initiative. https://doi.org/10.15497/FDAA09D5-5ED0-403D-B97A-2675E1EBE786.

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J. G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A. C. 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. 2016. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3:160018, Mar. http://dx.doi.org/10.1038/sdata.2016.18.

Peter Wittenburg, Margareta Hellström, Carlo-Maria Zwölf, Hossein Abroshan, Ari Asmi, Giuseppe Di Bernardo, Danielle Couvreur, Tamas Gaizer, Petr Holub, Rob Hooft, Ingemar Häggström, Manfred Kohler, Dimitris Koureas, Wolfgang Kuchinke, Luciano Milanesi, Joseph Padfield, Antonio Rosato, Christine Staiger, Dieter van Uytvanck, and Tobias Weigel. 2017. Persistent identifiers: Consolidated assertions. Status of November, 2017., December. https://doi.org/10.5281/zenodo.1116189.

Peter Wittenburg, Dieter Van Uytvanck, Thomas Zastrow, Pavel Strak, Daan Broeder, Florian Schiel, Volker Boehlke, Uwe Reichel, and Lene Offersgaard. 2018. CLARIN B Centre Checklist. Technical Report CE-2013-0095, CLARIN ERIC. http://hdl.handle.net/11372/DOC-78.

The Acorformed Corpus: Investigating Multimodality in Human-Human and Human-Virtual Patient Interactions

M. Ochs[1,2], P. Blache[1,3], G. Montcheuil[1,3,5], J.M. Pergandi[1,4], R. Bertrand[1,3], J. Saubesty[1,3], D. Francon[6], and D. Mestre[1,4]

[1]Aix Marseille Université, Université de Toulon, CNRS,
[2]LIS UMR 7020, [3]LPL UMR 7309, [4]ISM UMR 7287 ; [5]Boréal Innovation,
[6]Institut Paoli-Calmettes (IPC), Marseille, France
magalie.ochs@lis-lab.fr, blache@lpl-aix.fr

Abstract

The paper aims at presenting the Acorformed corpus composed of human-human and human-machine interactions in French in the specific context of training doctors to break bad news to patients. In the context of human-human interaction, an audiovisual corpus of interactions between doctors and actors playing the role of patients during real training sessions in French medical institutions have been collected and annotated. This corpus has been exploited to develop a platform to train doctors to break bad news with a virtual patient. The platform has been exploited to collect a corpus of human-virtual patient interactions annotated semi-automatically and collected in different virtual reality environments with different degree of immersion (PC, virtual reality headset and virtual reality room).

1 Introduction

For several years, there has been a growing interest in Embodied Conversational Agents (ECAs) to be used as a new type of human-machine interface. ECAs are autonomous entities, able to communicate verbally and nonverbally (Cassell, 2000). Indeed, several researches have shown that embodied conversational agents are perceived as social entities leading users to show behaviors that would be expected in human-human interactions (Krämer, 2008).

Moreover, recent research has shown that virtual agents could help human beings *improve their social skills* (Anderson et al., 2013; Finkelstein et al., 2013). For instance in (Anderson et al., 2013), an ECA endowed the role of a virtual recruiter is used to train young adults to job interview. In our project, we aim at developing a virtual patient to train doctors to break bad news. Many works have shown that doctors should be trained not only to perform medical or surgical acts but also to develop skills in communication with patients (Baile et al., 2000; Monden et al., 2016; Rosenbaum et al., 2004). Indeed, the way doctors deliver bad news has a significant impact on the therapeutic process: disease evolution, adherence with treatment recommendations, litigation possibilities (Andrade et al., 2010). However, both experienced clinicians and medical students consider this task as difficult, daunting, and stressful. Training health care professional to break bad news is now recommended by several national agencies (e.g. the French National Authority for Health, HAS)[1].

A key element to exploit embodied conversational agents for social training with users is their *believability* in terms of socio-emotional responses and global multimodal behavior. Several research works have shown that non-adapted behavior may significantly deteriorate the interaction and the learning (Beale and Creed, 2009). One methodology to construct believable virtual agent is to develop a model based on the analysis of a corpus of human-human interaction in the social training context (as for instance in (Chollet et al., 2017)). In our project, in order to create a virtual patient with believable multimodal reactions when the doctors break bad news, we have collected, annotated, and analyzed two multimodal corpora of interaction in French in this context. Both human-human and human-machine interaction are considered to investigate the effects of the virtual reality displays on the interaction. In this paper, we present the two corpora in the following sections.

[1]The French National Authority for Health is an independent public scientific authority with an overall mission of contributing to the regulation of the healthcare system by improving health quality and efficiency.

Magalie Ochs, Philippe Blache, Grégoire Montcheuil, Jean-Marie Pergandi, Roxane Bertrand, Jorane Saubesty, Daniel Francon and Daniel Mestre 2019. The Acorformed Coprus: Investigating Multimodality in Human-Human and Human-Virtual Patient Interactions. *Selected papers from the CLARIN Annual Conference 2018*. Linköping Electronic Conference Proceedings 159: 113–120.

2 Multimodal Human-Human Corpus Analysis to Model Virtual Patient's Behavior

The modeling of the virtual patient is based on an audiovisual corpus of interactions between doctors and actors playing the role of patients (called "Standardized patients") during real training sessions in French medical institutions (it is not possible, for ethical reasons, to record real breaking bad news situations). The use of "Standardized Patients" in medical training is a common practice. The actors are carefully trained (in our project, actors are also nurses) and follow pre-determined scenarios defined by experts to play the most frequently observed patients reactions. The recommendations of the experts, doctors specialized in breaking bad news situations, are global and related to the attitude of the patient ; the verbal and non-verbal behavior of the actor remains spontaneous. Note that the videos of the corpus have been selected by the experts as representative of real breaking bad news situations.

On average, a simulated consultation lasts 9 minutes. The collected corpus, in French, is composed of 13 videos of patient-doctor interaction (the doctor or the patient vary in the video), with different scenarios[2].

The initial corpus has been semi-manually annotated, leading to a total duration of 119 minutes. Different tools have been used in order to annotate the corpus. First, the corpus has been automatically segmented using SPPAS (Bigi, 2012) and manually transcribed using Praat (Boersma, 2002). The doctors' and patient's non-verbal behaviors have been manually annotated using ELAN (Sloetjes and Wittenburg, 2008). Different gestures of both doctors and patients have been annotated: head movements, posture changes, gaze direction, eyebrow expressions, hand gestures, and smiles. Three annotators coded the corpus. Each of them annotated a third of the corpus. The annotators were graduate students in linguistics and were paid to annotate. In order to insure homogeneity among the annotators, a guide was given describing every annotation steps. Moreover, the annotations' sessions were supervised, allowing the annotators to ask questions at any moment. In order to validate the annotation, 5% of the corpus has been annotated by one more annotator. The inter-annotator agreement, using Cohen's Kappa, was satisfying (k=0.63). More details on the corpus are presented in (Porhet et al., 2017).

2.1 Verbal cues annotations

Audio files were extracted from the video recordings. The speech signal was segmented into Inter-Pausal Units (IPUs), defined as speech blocks surrounded by at least 200 ms silent pauses. Due to its objective nature (Koiso et al., 1998), the IPU can be automatically segmented. However, due to poor audio quality, they were manually corrected. We manually transcribed each participant's speech on two different tiers using the TOE convention (Transcription Orthographique Enrichie / Enriched Orthographical Transcription, (Bertrand et al., 2008)). Note that we do not consider the acoustic features (e.g. prosody) since the audio quality of the videos does not enable us to study this aspect. The part-of-speech (POS) tags were automatically identified using MarsaTag (Rauzy et al., 2014). MarsaTag is a stochastic parser for written French which has been adapted to account for the specificities of spoken French. Among other outputs, it provides a morpho-syntactic category for each POS token.

2.2 Visual cues annotations

Different modalities of both the doctors and the patients have been annotated. The modalities as well as the corresponding values are described in Table 1[3] .

We summarize the annotation for each interlocutor in Table 2. The table reveals that the most frequent non-verbal signals are the doctor's and patient's head movements while few smiles appear. The number of words shows that the doctors speak more than the patient, as expected given the context of the interaction.

[2]The corpus is on Ortolang part of the CLARIN infrastructure

[3]As we are interested only in movements, we did not differentiated one movement from another. The hand annotation indicate the time interval from the moment the hands start moving until they return to the rest position.

Modality	Values
Head movements	nod, shake (negation), tilt, bottom, up, side
Posture change (movements of the bust)	forward, backwards, other change
Gaze direction	oneself, interlocutor, other direction, closed eyes
Eyebrow expression	frown, raise
Hand gesture	movement
Smile	smile, no smile

Table 1: Non-verbal modalities

Category	Doctors	Patients
Head	3649	1970
Hands	635	463
Gaze	1823	716
Smile	20	20
Eyebrows	225	189
Posture	239	257
Words	44816	727

Table 2: Total number of annotations per interlocutor

The annotated corpus has been analyzed for three different purposes:

- to build the *dialog model of the virtual patient*: the dialog model of the virtual patient is based on the notion of "*common ground*" (Garrod and Pickering, 2004; Stalnaker, 2002), *i.e.* a situation model represented through different variables that is updated depending on the information exchange between the interlocutors. The variables describing the situation model (e.g. the cause of the damage), specific to breaking bad news situations, have been defined based on the manual analysis of the transcribed corpus and in light of the pedagogical objective in terms of dialog. The dialog model is described in more detail in (Ochs et al., 2017) ;
- to design *non-verbal behaviors of the virtual patient*: the corpus has been used to enrich the non-verbal behavior library of the virtual patient with gestures specific to breaking bad news situations.
- to design *the feedback behavior of the virtual patient*: in order to identify the multimodal signals triggering feedback from the patients, we have applied sequences mining algorithms to extract rules to model the multimodal feedback behavior of the virtual patient (for more details (Porhet et al., 2017)).

3 Multimodal Human-Virtual Patient Corpus Analysis to Investigate the Users' experience with different virtual reality displays

Based on the corpus analysis presented in the previous section, we have implemented a virtual reality training system inhabited by a virtual patient and developed to give the capabilities to doctors to simulate breaking bad news situation. The system is *semi-autonomous* since it includes both automatic and manual modules, making it possible to simulate a fully automatized human-machine interaction (for more details on the semi-autonomous system (Ochs et al., 2018a)). Implemented on three different virtual environment displays (PC, virtual reality headset, and an immersive virtual reality room), the doctors can interact in natural language with a virtual patient that communicates through its verbal and non-verbal behavior (Figure 1). In order to collect the interaction and create the corpus of human-machine interaction in the context of breaking bad news, we have implemented a specific methodology. First, the doctor is filmed using a camera. His gestures and head movements are digitally recorded from the tracking data: his head (stereo glasses), elbows and wrists are equipped with tracked targets. A high-end microphone synchronously records the participant's verbal expression. As for the virtual agent, its gesture and verbal

Figure 1: Participants interacting with the virtual patient with different virtual environment displays (from left to right): virtual reality headset, virtual reality room, and PC.

expressions are recorded from the Unity Player. The visualization of the interaction, is done through a 3D video playback player we have developed (Figure 2). This player replays synchronously the animation and verbal expression of the virtual agent as well as the movements and video of the participant.

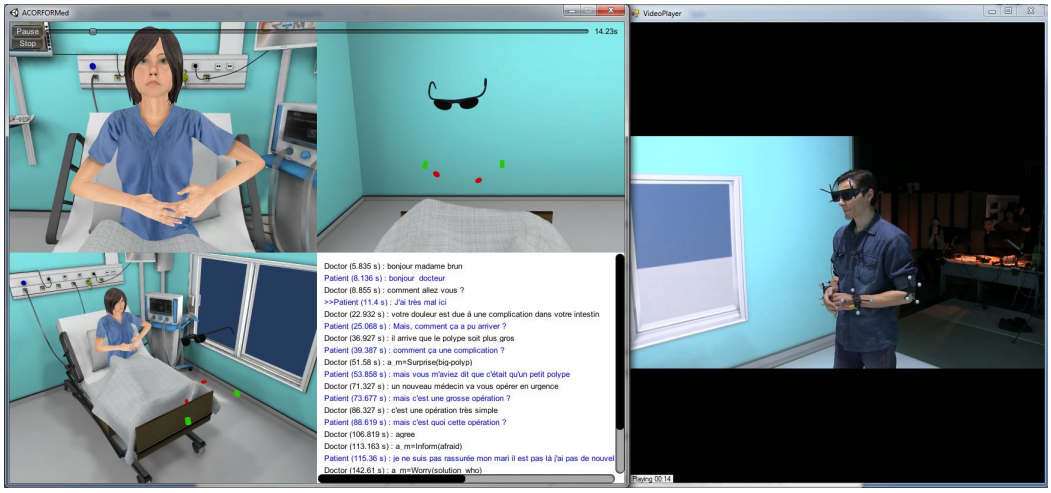


Figure 2: 3D video playback player

This environment facilitates the collection of corpora of doctor-virtual patient interaction in order to analyze the verbal and non-verbal behavior in different immersive environments.

Using the semi-autonomous system, we have collected 108 interactions in French of participants with the virtual patient. In total, 36 persons have participated in the experimentation. Ten of them are real doctors that already have experience in breaking bad news to real patients. Each participant has interacted with the systems 3 times with three different devices: PC, virtual reality headset, and virtual reality room. The task of the participants was to announce a digestive perforation after a gastroenterologic endoscopy in immediate post operative period[4]. The collected corpus is composed of 108 videos (36 per device). The total duration of the corpus is 5h34 (among which two hours with real doctors). In average, an interaction lasts 3mn16 (an example of interaction is presented on the following web page http://www2.lpl-aix.fr/ acorformed/videos.html).

3.1 Segmentation

The interaction between the participants and the virtual patient is split into 3 phases: the beginning, the central part, and the conclusion. Based on a previous analysis of human-human interaction in the same context (Saubesty and Tellier, 2016), we suppose that the verbal and nonverbal behavior may differ de-

[4]The scenario has been carefully chosen with the medical partners of the project for several reasons (e.g. the panel of resulting damages, the difficulty of the announcement, its standard characteristics of announce).

pending on the phases of the interaction. Keeping this in mind, we performed our analysis independently for each phase for all the data sources we have. We defined the size of each phase relative to the total duration of the interaction. As a first step, we define empirically the duration of each phase: 15% of the total conversation for the introduction, 70% for the central part of the interaction, and 15% for the conclusion. Note that a script in Python has been written with the percentage of each phase in parameter to automatically compute the verbal and non-verbal cues described in the following with different segmentation.

3.2 Verbal cues

In order to analyze the verbal behavior of the participants, we have defined high-level characteristics reflecting the *lexical richness* and the *linguistic complexity* of the user's verbal behavior based on the frequency of the part-of-speech tags for each participant and each phase of the interaction. Using a specific tool called SPPAS (Bigi, 2012), we performed a tokenization followed by a phonetization on the transcription file. The part-of-speech (POS) tags were automatically identified using MarsaTag (Rauzy et al., 2014). We consider 9 parts-of-speech tags: adjective, adverb, auxiliary, conjunction, determiner, noun, preposition, pronoun, verb.

Based on these POS tags, we computed the *lexical richness*, measured as the fraction of adjectives and adverbs out of the total number of tokens and the *linguistic complexity*, measured as the fraction of conjunctions, prepositions and pronouns out of the total number of token. The descriptive statistics are reported Table 3.

	Introduction		Central part		Conclusion	
	Average	SD	Average	SD	Average	SD
Lexical Richness	0.16	0.07	0.15	0.03	0.18	0.09
Linguistic Complexity	0.15	0.05	0.17	0.03	0.19	0.08
Length of the sentences	6.24	4.04	8.73	1.70	7.86	3.73
Length of IPUs	1.92	1.97	1.92	0.40	2.76	3.32

Table 3: Average and Standard Deviation (SD) of the verbal cues per phase.

Moreover, we have computed the *length of the sentences in terms of number of words* and the *lengths of inter-pausal units in terms of duration*. We compute the average length of sentences in each phase of the interaction for each participant. The length corresponds to the number of words of a sentence. The MarsaTag tool (Rauzy et al., 2014) has been used to define the sentences from the transcript text. The speech signal was segmented into Inter-Pausal Units (IPUs), defined as speech blocks surrounded by at least 200 ms silent pauses[5]. Due to its objective nature, the IPU has been automatically segmented using SPASS (Bigi, 2012). The descriptive statistics are reported Table 3.

3.3 Non-Verbal cues

Concerning the non-verbal cues, we have computed the *entropy* to characterize the movements of the participant in the virtual environment. The entropy is a common measure in virtual reality domain to assess the movements of the participants (Maïano et al., 2011). To obtain the entropy of the curve defined by the movement of each tracker on the participant, following the method described in (Dodson et al., 2013), we have computed the upper-bound on the Shannon entropy of curves of each plane (x, y and z) and each tracked point (head, left wrist, right wrist, left elbow, and right elbow). Finally, the different computed values of entropy are averaged to obtained two non-verbal cues: the average movements of the head, and the average movement of the arms. The descriptive statistics are reported Table 4.

[5]For French language, lowering this 200 ms threshold would lead to many more errors due to the confusion of pause with the closure part of unvoiced consonants, or with constrictives produced with a very low energy.

	Introduction		Central part		Conclusion	
	Average	SD	Average	SD	Average	SD
Head	1.61	0.43	2.94	0.55	1.55	0.45
Arms	1.31	0.42	2.47	0.44	1.31	0.49

Table 4: Average and Standard Deviation (SD) of the non-verbal cues per phase.

3.4 Sense of presence

In order to evaluate the global experience of the users, we asked the participants to fill different questionnaires on their subjective experience to measure their feeling of presence (with the *Igroup Presence Questionnaire*, IPQ (Schubert, 2003)), feeling of co-presence (Bailenson et al., 2005), and perception of the believability of the virtual patient (questions extracted from (Gerhard et al., 2001))[6]. These subjective evaluations enabled us to *tag* the video of the corpus with the results of these tests and then to correlate objective measures (e.g. verbal and non-verbal cues of the participants) to subjective measures (e.g. feeling of presence and perception of the virtual patient's believability) using machine learning methods (for more details see (Ochs et al., 2018b)).

4 Conclusion

In this article, we have presented two multimodal *comparable* corpora. The corpora have been collected in the same context of doctors' trainings to break bad news but in two different conditions of interaction: human-actor patient and human-virtual patient. They have been analyzed manually and using data mining methods in order to construct an autonomous virtual reality training platform inhabited by a virtual patient.

Given the different natures of the corpora, different annotations techniques - manual, semi-automatic and automatic - have been used, leading to different annotations schemes. Our next step is to harmonize the annotations of the two corpora in order to compare the verbal and non-verbal behaviors of the doctors depending on the type of the interaction. Our final goal is to identify objective verbal and non-verbal cues that could reflect the engagement of the user in the interaction with the virtual patient based on the verbal and the non-verbal cues identified in the human-human interaction.

Acknowledgements

This work has been funded by the French National Research Agency project ACORFORMED (ANR-14-CE24-0034-02) and supported by grants ANR-16-CONV-0002 (ILCB), ANR-11-LABX-0036 (BLRI) and ANR-11-IDEX-0001-02 (A*MIDEX).

5 Bibliographical References

References

K. Anderson, E. André, T. Baur, S. Bernardini, M. Chollet, E. Chryssafidou, I. Damian, C. Ennis, A. Egges, P. Gebhard, et al. 2013. The tardis framework: intelligent virtual agents for social coaching in job interviews. In *Advances in computer entertainment*, pages 476–491. Springer.

A. D. Andrade, A. Bagri, K. Zaw, B. A. Roos, and Ruiz J. G. 2010. Avatar-mediated training in the delivery of bad news in a virtual world. *Journal of palliative medicine*, 13(12):1415–1419.

W. Baile, R. Buckman, R. Lenzi, G. Glober, E. Beale, and A. Kudelka. 2000. Spikes—a six-step protocol for delivering bad news: application to the patient with cancer. *Oncologist*, 5(4):302–311.

J. N. Bailenson, C. Swinth, K.nd Hoyt, S. Persky, A. Dimov, and J. Blascovich. 2005. The independent and interactive effects of embodied-agent appearance and behavior on self-report, cognitive, and behavioral markers of copresence in immersive virtual environments. *Presence: Teleoperators and Virtual Environments*, 14(4):379–393.

[6]The analyze of the subjective experience of the participants is out of scope of this paper and is described in an other article (Ochs et al., 2018c)

R. Beale and C. Creed. 2009. Affective interaction: How emotional agents affect users. *International journal of human-computer studies*, 67(9):755–776.

Roxane Bertrand, Philippe Blache, Robert Espesser, Gaëlle Ferré, Christine Meunier, Béatrice Priego-Valverde, and Stéphane Rauzy. 2008. Le cid-corpus of interactional data-annotation et exploitation multimodale de parole conversationnelle. *Traitement automatique des langues*, 49(3):pp–105.

B. Bigi. 2012. Sppas: a tool for the phonetic segmentations of speech. In *The eighth international conference on Language Resources and Evaluation*.

P. Boersma. 2002. Praat, a system for doing phonetics by computer. *Glot international*, 13(341-345).

J. Cassell. 2000. More than just another pretty face: Embodied conversational interface agents. *Communications of the ACM*, 43:70–78.

M. Chollet, M. Ochs, and C. Pelachaud. 2017. A methodology for the automatic extraction and generation of non-verbal signals sequences conveying interpersonal attitudes. *IEEE Transactions on Affective Computing*.

Michael Maurice Dodson, Michel Mendes France, and Michel Mendes. 2013. On the entropy of curves. *Journal of Integer Sequences*, 16(2):3.

S. Finkelstein, S. Yarzebinski, C. Vaughn, A. Ogan, and J. Cassell. 2013. The effects of culturally congruent educational technologies on student achievement. In *International Conference on Artificial Intelligence in Education*, pages 493–502. Springer.

S. Garrod and M. Pickering. 2004. Why is conversation so easy? *Trends in cognitive sciences*, 8(1):8–11.

M. Gerhard, D. J Moore, and D. Hobbs. 2001. Continuous presence in collaborative virtual environments: Towards a hybrid avatar-agent model for user representation. In *International Workshop on Intelligent Virtual Agents*, pages 137–155. Springer.

Hanae Koiso, Yasuo Horiuchi, Syun Tutiya, Akira Ichikawa, and Yasuharu Den. 1998. An analysis of turn-taking and backchannels based on prosodic and syntactic features in japanese map task dialogs. *Language and speech*, 41(3-4):295–321.

N. Krämer. 2008. Social effects of virtual assistants. a review of empirical results with regard to communication. In *Proceedings of the international conference on Intelligent Virtual Agents (IVA)*, pages 507–508, Berlin, Heidelberg. Springer-Verlag.

Christophe Maïano, Pierre Therme, and Daniel Mestre. 2011. Affective, anxiety and behavioral effects of an aversive stimulation during a simulated navigation task within a virtual environment: A pilot study. *Computers in Human Behavior*, 27(1):169–175.

K. Monden, L. Gentry, and T. Cox. 2016. Delivering bad news to patients. *Proceedings (Baylor University. Medical Center)*, 29(1).

M. Ochs, G. Montcheuil, J-M Pergandi, J. Saubesty, B. Donval, C. Pelachaud, D. Mestre, and P. Blache. 2017. An architecture of virtual patient simulation platform to train doctor to break bad news. In *International Conference on Computer Animation and Social Agents (CASA)*.

M. Ochs, P. Blache, G. Montcheuil, J.-M. Pergandi, J. Saubesty, D. Francon, and D. Mestre. 2018a. A semi-autonomous system for creating a human-machine interaction corpus in virtual reality: Application to the acorformed system for training doctors to break bad news. In *Proceedings of LREC*.

Magalie Ochs, Sameer Jain, Jean-Marie Pergandi, and Philippe Blache. 2018b. Toward an automatic prediction of the sense of presence in virtual reality environment. In *Proceedings of the 6th International Conference on Human-Agent Interaction*, pages 161–166. ACM.

Magalie Ochs, Daniel Mestre, Grégoire De Montcheuil, Jean-Marie Pergandi, Jorane Saubesty, Evelyne Lombardo, Daniel Francon, and Philippe Blache. 2018c. Training doctors' social skills to break bad news: evaluation of the impact of virtual environment displays on the sense of presence. *Journal on Multimodal User Interfaces*, pages 1–11.

C. Porhet, M. Ochs, J. Saubesty, G. Montcheuil, and R. Bertrand. 2017. Mining a multimodal corpus of doctor's training for virtual patient's feedbacks. In *Proceedings of 19th ACM International Conference on Multimodal Interaction (ICMI), Glasgow, UK*.

Stéphane Rauzy, Grégoire Montcheuil, and Philippe Blache. 2014. Marsatag, a tagger for french written texts and speech transcriptions. In *Proceedings of Second Asian Pacific Corpus linguistics Conference*, page 220.

M. Rosenbaum, K. Ferguson, and J. Lobas. 2004. Teaching medical students and residents skills for delivering bad news: A review of strategies. *Acad Med*, 79.

J. Saubesty and M. Tellier. 2016. Multimodal analysis of hand gesture back-channel feedback. In *Gesture and Speech in Interaction, Nantes, France*.

T. Schubert. 2003. The sense of presence in virtual environments: A three-component scale measuring spatial presence, involvement, and realness. *Zeitschrift für Medienpsychologie*, 15(69-71).

H. Sloetjes and P. Wittenburg. 2008. Annotation by category: Elan and iso dcr. In *6th International Conference on Language Resources and Evaluation*.

R. Stalnaker. 2002. Common ground. *Linguistics and Philosophy*, 25(5):701–721.

Discovering software resources in CLARIN

Jan Odijk
UiL-OTS
Utrecht University, the Netherlands
j.odijk@uu.nl

Abstract

I present a CMDI profile for the description of software that enables discovery of the software and formal documentation of aspects of the software, and a proposal for faceted search in metadata for software. The profile has been tested by making metadata for over 80 pieces of software. The profile forms an excellent basis for formally describing properties of the software, and for a faceted search dedicated to software which enables better discoverability of software in the CLARIN infrastructure. A faceted search application for this purpose has been implemented. A curation procedure is proposed to ensure that descriptions of software made on the basis of other profiles contain the relevant information in the right form and use the right vocabularies, and we created an experimental faceted search that includes software descriptions based on the WebLichtWebService profile.

1 Introduction

Enabling the easy discovery of resources is an important goal of CLARIN. The Virtual Language Observatory (VLO) serves this purpose, but it is currently mostly suited for the discovery of *data*. Discovering *software* is not so easy in the current VLO. The (pretty complex) query Software Query[1] approximates finding all software descriptions in the VLO. It finds 1219 descriptions of software (on 2019-01-11). However, there are no facets dedicated to software to refine one's search. In order to address this issue I present a CMDI profile for the description of software (ClarinSoftwareDescription, CSD) that enables discovery of the software and formal documentation of aspects of the software (section 2). The profile has been tested by making metadata for over 80 pieces of software (section 3). I also describe how the quality of these metadata descriptions was ensured (section 4). I present a proposal for faceted search in metadata for software (section 5). An experimental version of the proposed faceted search has been implemented. I propose to add this faceted search to the VLO. It should then also cover descriptions of software created on the basis of other profiles. I show how metadata curation software, combined with provided metadata curation files, can curate existing metadata descriptions for software using other profiles to make them suited for this faceted search (section 6). An experiment with the WebLichtWebService profile was carried out, resulting in a faceted search covering not only CSD but also WebLichtWebService based descriptions of software. In section 7 I summarise the main findings, point out some problems encountered, indicate required future work, and make some recommendations.

2 Metadata Profile ClarinSoftwareDescription

The ClarinSoftwareDescription (CSD) profile[2] enables one to describe information about software in accordance with the CMDI metadata framework used in CLARIN (Broeder et al., 2010; Broeder et al.,

[1] https://vlo.clarin.eu/search?0&fq=resourceClass:R-code&fq=resourceClass:software&fq=resourceClass:Tools&fq=resourceClass:web+application&fq=resourceClass:webservice&fq=resourceClass:software,+webservice&fq=resourceClass:web+service&fq=resourceClass:tool+service&fq=resourceClass:software&fqType=resourceClass:or

[2] clarin.eu:cr1:p_1342181139640.

Jan Odijk 2019. Discovering software resources in CLARIN. *Selected papers from the CLARIN Annual Conference 2018*. Linköping Electronic Conference Proceedings 159: 121–132.

2012). The profile has been set up in such a way that it enables (1) the description of properties that support discovery of the resource, and (2) the description of properties for documenting the resource, in as formal a manner as possible.

I briefly describe the major components and elements of the profile. The elements crucial for finding the resource are dealt with in more detail in section 5.

The profile consists of the CMDI components *Generalinfo*, *SoftwareFunction*, *SoftwareImplementation*, *Access*, *ResourceDocumentation*, *SoftwareDevelopment*, *TechnicalInfo*, *Service* and *LRS*.

The component *Generalinfo* enables one to describe general information about the resource. It is an extension of the component *cmdi-generalinfo*.[3] It includes elements for the *name* and *title* of the resource, its *version*, *publication year*, *owner* and contact *email address*, a *URL* and/or a *PID* to the resource, its *time coverage*, *release status*, *CLARIN Centre* hosting the resource and the *national project(s)* in which it has been made part of the CLARIN infrastructure, any *alternative names* for the resource and a *description* of the software.

The *SoftwareFunction* component enables one to describe the function of the software in terms of the closed vocabulary elements *tool category, tool tasks*, *research phase(s)* (for which it is most relevant), *research domains* and, for the linguistics domain, relevant *linguistic subdisciplines* for which it was originally developed.[4] It also describes the *language variants* that the software applies to and offers facilities for documentation about its *performance*.

The *SoftwareImplementation* component enables one to describe information for users on the implementation and installation of the software. It describes how the software is *distributed*, what the *installation requirements* are, the nature of the *interface* with the user or other software, properties of the *package* that the software is delivered in (if any) and what the *input* and the *output* of the software is.

The input and output specification enable a quite detailed description of properties of the input required and output generated by a piece of software. The following is an example of the output specification for the Alpino parser:

outputType text
characterEncoding utf8
Schema LASSY DTD
MimeType text/xml
AnnotationType
    AnnotationType Morphosyntax/Inflection
    AnnotationType Morphosyntax/Lemma
    AnnotationType Morphosyntax/POS
    AnnotationType Morphosyntax/Word form
    AnnotationType Orthography/Token
    TagSet POSTags/DCOI Tagset
AnnotationType
    AnnotationType Syntax/Chunks
    AnnotationType Syntax/Dependency Relations
    AnnotationType Syntax/Grammatical Relations
    AnnotationType Syntax/Phrases
    AnnotationType Syntax/Syntactic Categories
    AnnotationType Syntax/Multiword Expressions
    TagSet Syntax/Alpino Tagset

It specifies that Alpino yields *text* as output, with character encoding *utf8*, in accordance with a schema called *LASSY DTD*, and with mimetype *text/xml*. Alpino generates multiple annotations, here grouped in two groups because the tag set used differs per group. The first group of annotations concerns *inflection*, *lemma*, *part of speech (POS)*, *word form* and *token*, encoded with the *DCOI Tagset*. The second group involves *chunks*, *dependency relations*, *grammatical relations*, *phrases*, *syntactic categories*, and

[3] clarin.eu:cr1:c_1342181139620.

[4] which, of course, does not preclude its use in other research domains that were not foreseen during development.

multiword expressions, encoded with the Alpino Tagset. This information can be used, together with the metadata of the input data, to automatically generate metadata for the output data generated by Alpino, provided of course that metadata for (textual or other) data use the same annotation labels. The ability to generate such rich metadata for output of tools is very important for data provenance in general, and for applications such as the CLARIN SwitchBoard (see (Zinn, 2016a) and below), which forms a great aid for users for finding applications and services that they can apply to a particular data set.

Note that all values of the metadata elements *AnnotationType* and *TagSet* come from a closed vocabulary. Since the number of possible values is already large (currently 61 different possible values for *AnnotationType*, 10 different possible values for *TagSet*), and since one can certainly expect these numbers to grow, I grouped the values in classes, indicated here by the label before the slash. In this way, closely related values can be inspected together, and one can concentrate on those finegrained distinctions that one is interested in without being bothered by finegrained distinctions that one is not interested in. The importance of such grouping was already pointed out by (Odijk, 2009, 12-13). No support for such a feature is currently available in CLARIN. This is why I opted for the poor man's option of specifying a superordinate category in each value before the actual value separated from it by a slash, but this is clearly to be seen as an ad hoc and temporary solution. For a more principled solution, see section 2.1.

The *Access* component enables one to describe information about the availability and accessibility of the resource. It is an extension of the *cmdi-access* component.[5] It contains a reference to a *catalogue link*, information about the *license* for the resource, information about *copyright* and *copyright holder(s)* and a *contact* organisation and/or person.

The *ResourceDocumentation* component enables one to describe the documentation of the resource. It offers facilities to describe the *documentation* and *publications* on the resource, a *description* of the resource and *pictures* (including *logos*) related to the resource. The *SoftwareDevelopment* component is intended for information on the history and development of the software. It offers facilities to describe the *source*(s) that the software is based on or from which it has been derived, the *project* in which the software was created, the *creator(s)* of the software, and any planned *software updates*. The *TechnicalInfo* component enables one to describe technical information on a resource and is mainly aimed at developers. It provides facilities to describe the *run time environment*, any *access protocols*, as well as the *programming language(s)* that have been used to implement the software.

The *Service* component (CLARIN-NL Web Service description) is intended for describing properties of web services. It is compatible with the CLARIN CMDI core model for Web Service description version 1.0.2.[6]

The *LRS* component is intended for the description of the properties of a particular task for the CLARIN Language Resource SwitchBoard (CLRS, (Zinn, 2016b; Zinn, 2016a; Zinn, 2017)). Multiple LRS components can be present. It is our viewpoint that specifications for an application for inclusion in the CLRS registry[7] should be derivable from the metadata for this application. This was not the case for the CSD profile when the CLRS came into existence, so I added a component to offer facilities for supplying the missing information. I devised a script to turn a CSD-compatible metadata record that contains an LRS component into the format required for the CLRS and tested it successfully with the *Frog* web service and application (van den Bosch et al., 2007). See `https://languagemachines.github.io/frog/` for Frog's source code and `http://portal.clarin.nl/node/8516` for its entry in the faceted search described in Section 5.

2.1 Semantics

Many of the profile's components, elements and their possible values have a semantic definition by a link to an entry in the CLARIN Concept Registry (CCR, (Schuurman et al., 2016)).[8] For the ones that were lacking I created definitions and provided other relevant information required for inclusion into the CCR.

[5] clarin.eu:cr1:c_1311927752326.

[6] This component was created by Menzo Windhouwer, and adapted to the requirements of CMDI version 1.2.

[7] `https://github.com/clarin-eric/LRSwitchboard-rest/blob/master/Registry.js`

[8] `https://concepts.clarin.eu/ccr/browser/`.

I submitted this file (2017-09-08), in the format required, to the maintainers of the CCR.[9] However, the CCR coordinators'[10] mill runs slowly, and, though there has been some activity on the proposed concepts, so far none of them have been incorporated in the CCR. This constitutes a real bottleneck, which should be addressed in CLARIN. After our submission to the CCR, I made some new modifications to the profile, so there are new elements and values for which the semantics does not exist yet.

I also specified relations between concepts in the input, but I was immediately told that that was not supported yet by CCR. It could be implemented in CLARIN by specifying *isa* relations in the CCR. By making use of small taxonomies of concepts derived from this information, the CMDI Component Registry Editor, dedicated CMDI metadata editors, and faceted search facilities can make the work of people who edit profiles and components, create or adapt metadata, or use faceted search considerably more pleasant and more effective. Unfortunately, no such options are currently offered in the CCR.

2.2 Comparison with other profiles for software

There are about 20 profiles for the description of software in the CLARIN Component Registry (as determined on 2017-09-29), but most are not in use or in use for a single description only. The only profiles that are used for multiple software resources (measured on 2019-01-15) are *ToolProfile*[11] (69 resources), *WebLichtWebService*[12] (419 resources), *resourceInfo*[13] (189 software resources), *OLAC-DcmiTerms*[14] (175 software resources), and *LINDAT_CLARIN*[15] (83 software resources). The instances describing software can be identified on the basis of the VLO facet *resource type*, using the query given in Section 1 and repeated here for convenience: Software Query. It finds (on 2019-01-11) 535 descriptions with *resource type= software,web service*[16], 419 descriptions with *resource type = web service*[17], 162 with *resource type = webservice*[18], 181 with *resource type = software*[19], 72 with *resource type = tool service*[20], and a small number of descriptions with other values for *resource type*.

I summarise the most important differences between the CSD profile and the most used profiles: (1) the CSD profile is fully dedicated to the description of software (v. the *OLAC-DcmiTerms* and *LINDAT_CLARIN* profiles); (2) the CSD profile can be used to describe any type of software (v. *WebLicht-Webservice*, which is intended for web services only); (3) CSD offers more elements, and more formalised elements than the other profiles, not only elements useful for discovery but also for (formalised) documentation; (4) CSD offers more and/or more extensive closed vocabularies for many metadata properties, e.g. for *toolTask*, *applicationType*, *ResearchDomain*, *LinguisticsSubject*, etc. Corresponding metadata elements in other profiles usually allow any string as value.[21]

[9] It can be found here: `https://surfdrive.surf.nl/files/index.php/s/oWUgl1664VraCMo`.

[10] https://www.clarin.eu/content/concept-registry-coordinators

[11] clarin.eu:cr1:p_1290431694581.

[12] clarin.eu:cr1:p_1320657629644.

[13] clarin.eu:cr1:p_1360931019836.

[14] clarin.eu:cr1:p_1288172614026.

[15] clarin.eu:cr1:p_1403526079380

[16] `https://vlo.clarin.eu/search;jsessionid=07859474E8155F6D595632FD827F38F4?0&fq=resourceClass:software,+webservice&fqType=resourceClass:or`.

[17] `https://vlo.clarin.eu/search;jsessionid=07859474E8155F6D595632FD827F38F4?0&fq=resourceClass:web+service&fqType=resourceClass:or`

[18] `https://vlo.clarin.eu/search;jsessionid=07859474E8155F6D595632FD827F38F4?0&fq=resourceClass:webservice&fqType=resourceClass:or`

[19] `https://vlo.clarin.eu/search;jsessionid=07859474E8155F6D595632FD827F38F4?0&fq=resourceClass:software&fqType=resourceClass:or`.

[20] `https://vlo.clarin.eu/search;jsessionid=07859474E8155F6D595632FD827F38F4?0&fq=resourceClass:tool+service&fqType=resourceClass:or`.

[21] There are also other desiderata: in particular one would like to avoid what (from the perspective of successful communication) I would like to call the 'horrors of natural language'. One can achieve this by not using existing natural language words as values in a closed vocabulary. Successful communication is seriously hampered by natural language, even in as simple a domain as words or terms: natural language words have associations, have a (common sense) meaning, are often ambiguous, are specific to one language, and have variations (abbreviations, acronyms etc.). These properties make successful communication difficult if not impossible. Good closed vocabulary definitions use arbitrary labels that at best resemble existing words for mnenomical reasons but that are no natural language words. Many standards adhere to this guideline, e.g., the ISO-codes for languages and countries. Unfortunately, they have not been adhered to for all metadata elements in the CSD profile, and here surely improvement is possible.

The *WebLichtWebService* profile is intended for web services only, and it has many relevant properties not represented in a formal way (e.g. there is no formal specification of the language(s) that a web service can apply to, not for Mimetype.

The *ToolProfile* profile is a very good profile, which offers many facilities for specifying properties of software. However, the metadata elements *ToolType* and *ClassificationType* allow any string as value and are not restricted to a closed vocabulary. The same holds for *FieldOfResearch*.

The *resourceInfo* profile for tools, which did not yet exist when I started creating the CSD profile is actually a specific instantiation of a profile for all kinds of resources. It is a profile that has been used mainly to convert META-SHARE descriptions into CMDI descriptions. It offers special elements for the description of software in the component toolServiceInfo[22], with closed vocabulary elements *resourceType* and *toolServiceType*, an open vocabulary element *toolServiceSubtype*, and components *inputInfo*, *outputInfo*, *toolServiceOperationInfo*, *toolServiceEvaluationInfo*, and *toolServiceCreationInfo*, which are all highly relevant to the description of software.

The *LINDAT_CLARIN* profile is intended for all kinds of resources, including software, but it has only very general metadata elements and nothing dedicated to the description of software.

3 Metadata Descriptions using the CSD profile

I have described more than 80 software resources with the CSD profile, and describing these software resources resulted in various improvements of earlier versions of the profile.[23] These software resources mainly concern resources from the Netherlands. Most descriptions started from the information contained in the CLARIN-NL Portal, Services part.[24] The information there was semi-automatically converted to CMDI metadata in accordance with the CSD profile. The resulting descriptions were further extended and then submitted to the original developers and CLARIN Centres that host the resources for corrections and/or additions. The CMDI descriptions can be found here: `https://surfdrive.surf.nl/files/index.php/s/VEJOEkfbFtWR6Y6`. our team is in the process of making them available to the Centres that host the software so that they can be harvested by the VLO.

4 Metadata Quality

All metadata descriptions have been validated against the profile definition. I created several schematron[25] files for issuing errors or warnings for phenomena that are syntactically correct but incorrect or potentially incorrect in other ways. These schematron files check for the presence or absence of important (but optional) elements, for dependencies between elements or their values, e.g. an element to specify the language that the software can apply to must be present unless the value for the element *languageIndependent* is *yes*. I also made a schematron file to check for the presence of elements that are crucial for the faceted search described in section 5. A script has been provided for validation and for applying the schematron files. Additionally, a script was made to identify all URLs in the metadata descriptions and check for their resolution.[26]

The quality checks offered by the CLARIN Curation Module[27] (Ostojic et al., 2017) have also been used but are less useful because they can be applied only to a single metadata description at a time, check for the presence of metadata relevant for faceted search for data in the VLO, many of which are not so relevant for software, and because the profiles are cached so that modifications of the profiles are not immediately taken into account.

[22] clarin.eu:cr1:c_1360931019834

[23] A first version of the profile was presented by (Westerhout and Odijk, 2013), and at that time the profile was tested only on 5 software resources.

[24] `http://portal.clarin.nl/clarin-resource-list-fs`.

[25] `http://schematron.com/`

[26] On 2018-10-01, 969 URLs were correctly found, 11 URLs were found but no access was granted, 35 URLs were not found, and 4 exceptions were raised.

[27] `https://clarin.oeaw.ac.at/curate/`

5 Faceted Search

A major purpose of metadata is to facilitate the discovery of resources. An important instrument for this purpose in CLARIN is the Virtual Language Observatory (VLO, (Van Uytvanck, 2014)). The VLO offers faceted search for resources through their metadata, but its faceted search is fully tuned to the discovery of *data*. For this reason, our team defined a new faceted search, specifically tuned to discovery of *software*. This faceted search offers *search* facets and *display* facets:

Search Facets LifeCycleStatus, ResearchPhase, toolTask, ResearchDomain, LinguisticsSubject, inputLanguage, applicationType, NationalProject, CLARINCentre,

Display Facets name, title, version, inputMimetype, outputMimetype, outputLanguage, Country, Description, ResourceProxy, AccessContact, ProjectContact, CreatorContact, Documentation, Publications, sourcecodeURI, Licence, CMDI File Link, Project, logo or picture, OriginalLocation, and all search facets.

I will discuss search facets in section 5.1, display facets in section 5.2, and end in section 5.3 with a description of the implementation of the faceted search.

5.1 Search Facets

I submit that many of the facets under search facets are very useful for a researcher who is trying to find a piece of software that might be relevant to his/her research. The *ResearchDomain* facet enables the researcher to select the tools that (according to the developers of the software) are relevant to a particular research domain (linguistics, philosophy, literary studies, etc.). For the research domain *Linguistics* further subdivisions can be made using the facet *LinguisticsSubject* (e.g. syntax, phonology, morphology, etc.). The *ResearchPhase* facet enables the researcher to restrict the tools to those tools that are suited for the actual research phase: is the researcher looking for data, does the researcher want to enrich existing data, does the researcher want to search in data, etc. etc. An extensive description of the meaning of this facet and its values (i.e., which research phases are distinguished) is provided here: `http://dev.clarin.nl/node/4723`.

The *toolTask* facet specifies the function(s) of a piece of software, i.e. what does it do? For example, is it a tool for searching in data, for enriching words in text with part of speech tags, for enriching words in text with lemma's, etc. The *applicationType* facet indicates whether the software is a web application, a desktop tool, or a web service, etc. The *inputLanguage* facet is also important, because often a researcher is only interested in a specific language or a small number of languages. The type of input that the tool can work on is also very important, but there currently is no search facet for it. There is a facet for *inputMimeType* but I believe that the large amount of possible values and the fine-grained distinctions made by it make it less suited as a search facet. In the future, I plan to add a facet that can be derived from the *inputMimeType* but has only a limited number of values, basically corresponding to the major modalities and a small number of subtypes (text, audio, audio/speech, video). I also hope to add a search facet for licence class in the future, but for that I first will have to define a limited number of values for licence classes (which I want to be a bit richer than the *Availability* facet in the VLO).

All the facets have values from (what I would like to call) half-open vocabularies. These are basically closed vocabularies, with one special value *other*. These closed vocabularies can be extended, yielding an *updated* closed vocabulary, but they can only be extended with new values with a semantics that does not overlap with the existing values (except for *other*). In the updated vocabulary, all previously existing values retain their original semantics, except for the value *other*, the scope of which is reduced. Such updates will be required regularly, especially in early phases, because no one has a full overview of all the different types of tools, and no one can foresee what new types of tools will come into existence in the future. For this reason, many people use open vocabularies, which of course provides the necessary flexibility, but results in a complete mess and impedes effective search seriously. This has been observed by many (e.g. (Odijk, 2014)) and a curation task force has been set up in CLARIN to reduce the mess resulting from this freedom as much as possible.[28] I try to avoid changes of such vocabularies in which

[28] So far such efforts have only been partially successful, e.g. the situation for the VLO facet *resource type* has been improved significantly recently, but, restricting attention to values for software, the values *software* and *software,webservice*, *Tools* and

the semantics of existing values change, though this may occasionally be necessary (in such a case we speak of an *upgrade* of the vocabulary).

It is crucial for effective search (i.e. easy queries with optimal recall and precision) to have closed vocabularies as much as possible. Values occurring in actual metadata descriptions may have different forms, but it is crucial to map these to values from the closed vocabulary. Regular monitoring of newly occurring values and adapting the curation tables is therefore required, and each national CLARIN project should reserve some effort and money to contribute to this task.

5.2 Display Facets

The display facets form a subset of the full metadata, and contain some additional elements.

The meaning of most display facets is obvious from their names (name, title, version, inputMimetype, outputMimetype, outputLanguage, Country, Description, AccessContact, ProjectContact, CreatorContact, Documentation, Publications, Licence, and Project).

The facet *ResourceProxy* contains one or more links to the actual application(s). The facet *sourcecodeURI* provides a link to the source code of the resource. The facet *CMDI File Link* contains a URL to the full metadata. If the metadata contain a logo or a picture, it is displayed in the faceted search.

The facet *OriginalLocation* contains the URL of the description in the CLARIN-NL Portal, Services part.[29] that the metadata record is based on. It is mainly maintained for future redirection purposes.

5.3 Implementation

Of course, for a faceted search application to work on the metadata offered by the VLO, first of all a distinction must be made between the metadata that describe data and the metadata that describe software. Currently, no such distinction is made, but it can be largely added automatically on the basis of the CMDI profiles used and some existing facets (in particular *resource type*), e.g. by using the query described in section 1.

Furthermore, all metadata profiles for the description of software must be able to provide the values for the facets. That is the case to a large extent, though some metadata curation is needed (in some cases, quite a lot) and existing values must be mapped to the closed vocabulary for use in the faceted search. This is the topic of the next section.

An initial, experimental, implementation of this faceted search has been made available.[30]. It enables one to test the faceted search with many users and to identify errors and omissions in the metadata descriptions. It can thus be tested extensively before it or an improved version of it is included in the VLO. Initial results of this test already led to the suggestion for a new facet, i.e. a facet that indicates what skills a researcher must have in order to be able to use the software, e.g. must the researcher be able to program (and in which language), must the researcher know a particular query language, is extensive knowledge of the structure of a dataset required, etc. etc.

6 Curation of existing metadata for software

I followed the metadata curation strategy sketched by (Odijk, 2015). The basic idea is as follows: a new standardised metadata record is automatically created for all software descriptions, in principle each time a record is harvested. This metadata record contains the components and elements that are required for the faceted search as defined above. The record is constructed from the original CMDI record for the resource, combined with the data for this resource contained in a curation file, by a script. The curation file contains a sequence of conditions on each relevant element, and a specification of which values for which elements should be included in the new record if all the conditions are met. In general, the conditions simply test for identity with a value. The curation file basically consists of two *XSV* files, one specifying the conditions, and the other to specify the changes that must be made (mostly: set an element to a particular value). An *XSV* (eXtended Separated Value) file is a CSV file in which each value can itself

tool service, *Web service*, *web service* and *webservice* exist next to one another. Values for the *resource type* facet for data are a much greater mess.

[29] `http://portal.clarin.nl/clarin-resource-list-fs`.

[30] `http://portal.clarin.nl/clariah-tools-fs`

consist of multiple values separated by a separator. Working with XSV files is very easy, but imposes some limitations, which probably can be overcome by using XSLT. The curation file can be used to add information that was lacking or only present in an unformalised way, and it can be used to map existing values to other values from a specific closed vocabulary. I report on experiments with such a curation file for the *WebLichtWebService* profile, since curation was most needed and most complex for this profile.

The *WebLichtWebService* profile lacks many elements that are necessary for faceted search, e.g. *toolTask*, *researchDomain*, *linguisticsSubject*, *inputLanguage*, *outputLanguage*, *Country*, *CLARINCentre*, *Documentation*, *Publications* and *license*. I made a curation file for many of these properties, which can be used to add the relevant information in a new metadata record for a WeblichtWebService description: this is the case for the facets *toolTask*, *researchDomain*, *linguisticsSubject*, *inputLanguage*, *outputLanguage*, and *Country*.

It may be surprising that the WebLichtWebService lacks formal representations for input and output language, because many of these web services function properly in the WebLicht environment (Hinrichs et al., 2010). The descriptions indeed contain information about the input and output languages, but it is hidden in parameters which have a parameter name (e.g. *lang*), without an explicit meaning, and a parameter value (e.g. *de*), also without an explicit meaning. Therefore, this information is insufficiently formally encoded, and only an intelligent human being can perhaps interpret this. The same holds for input and output Mimetype specifications. Two web services may still interact correctly if the same parameter name and values are used for language and Mimetype in all WebLicht web services. This appears to be the case for Mimetype (parameter name *type*), but not for language (mostly *lang* is used as the parameter name, but occasionally *language* also occurs. For values, both *de* and *Deutsch* occur as values to specify, I assume, the German language, and both *en* and *English* occur as values to specify, I assume, the English language.

An initial, experimental, and still incomplete version of faceted search that includes 286 (partially) curated software descriptions that are based on the WebLichtWebService profile has been made available.[31].

I still have to make curation files for the *ToolProfile*, *resourceInfo* and the *OLACDcmiTerms* profiles. I already inventoried the problems for the first two profiles, and curation files for these will be much simpler than the one for the *WebLichtWebService* profile.

The *ToolProfile* profile has elements for most facets. All query facets can be derived from existing fields except for *ResearchPhase*. Some elements use open vocabularies and require a mapping to standardized values (e.g. *FieldOfResearch* from which *researchDomain* and *linguisticsSubject* can be derived). Elements for the display facets *NationalProject*, *Publication*, *SourceCodeURI*, *CLARINCentre*, and *picture* are lacking.

The *resourceInfo* profile also has elements for most facets. It lacks elements for the query facets *LifeCycleStatus*, *ResearchPhase*, *researchDomain*, *linguisticsSubject*, *NationalProject*, and *CLARINCentre*. It lacks elements to derive the display facets *sourcecodeURI* and *picture*.

I still have to investigate the *LINDAT_CLARIN* profile.

7 Concluding Remarks

7.1 Summary

I presented a CMDI profile for the description of software that enables discovery of the software and formal documentation of aspects of the software, and a proposal for faceted search in metadata for software. The profile has been tested by making metadata for over 80 pieces of software. The profile forms an excellent basis for formally describing properties of the software, and for a faceted search dedicated to software which enables better discoverability of software in the CLARIN infrastructure. A faceted search application for this purpose has been implemented. A curation procedure has been proposed to ensure that descriptions of software made on the basis of other profiles contain the relevant information in the right form and use the right vocabularies, and our team created an experimental faceted search that includes software descriptions based on the WebLichtWebService profile.

[31] `http://portal.clarin.nl/clariah-tools-fs-global`

I encountered some problems or less desirable features of the CMDI infrastructure, of which I will briefly mention some:

1. Closed vocabularies are defined with an element, not as a separate and reusable enumerated type. I believe that this is a very unfortunate design decision, which has many negative effects, in particular, it is not possible to reuse this closed vocabulary. This problem has been partially solved by CLAVAS, but the maintainers of CLAVAS only want to accept widely accepted and used and rather stable vocabularies. An additional (minor) problem is that copying the vocabulary is only possible in the Component Registry by editing the element.
2. CMDI offers no possibility to reuse metadata *elements*: one can only reuse components, not elements, which (especially in combination with the previous point) creates many problems (e.g. I cannot reuse the closed vocabulary for *license* from the *resourceInfo* profile, which is the most extensive list of license types in the whole CMDI infrastructure). If one wants to stimulate reuse of an element in one's profile, one has to put an (otherwise unnecessary) component on top of the element.
3. There is a lot of variety in the contents of the CMDI envelope element *MdSelfLink*, resulting in several unresolved or syntactically incorrect results. One special case is that the *MDSelfLink* refers to an OAI-PMH description containing the metadata rather than to the metadata itself.
4. Lack of good CMDI metadata editors. Though there are some CMDI editors (e.g. Arbil, CMDI Maker, COMEDI), all have severe limitations, e.g. none supports CMDI 1.2. Arbil is a desktop application (which is OK) but requires a steep learning curve and is not really supported any more. *CMDI Maker*[32], despite its name, only supports the IMDI profile. ProForma[33] has been discontinued. COMEDI[34] (Lyse et al., 2015) is a web-based editor, and it suffers from most of the problems that most web interfaces have (Odijk, 2018), which makes it not easy to use for metadata entry. It remains to be seen whether the editor based on the CLARIAH CMDI Forms based approach proposed by (Zeeman and Windhouwer, 2018) will be any better in this respect, but I am not optimistic.

7.2 Future work

The work on the profile and the faceted search has not finished yet. In particular,

- The CSD profile must still be published in the CMDI registry. I did this in an earlier phase, but because of a bug in the CMDI registry for published profiles that are still under development, it was impossible for a team member to edit components originally created by another member of the development team. Therefore, the publishing was partially undone.
- The semantics of the metadata elements has to be finished (cf. section 2.1).
- The documentation of the profile has to be finalised.
- In the metadata descriptions I did not systematically distinguish between input and input parameters. This distinction should be drawn more sharply, and it will probably require an improvement of the facilities for describing parameters. In addition, the profile should enable descriptions of triples of parameters, input and output. This will reduce the need for the (somewhat ad-hocly added) LRS component.
- Some details must still be harmonised. This involve mainly adapting the systematic naming conventions that were adopted but could not be maintained because I reuse components developed by others who follow different naming conventions.
- Due to the long development time of the profile by multiple persons some redundancies have occurred in the profile, which should be removed.
- The faceted search should be extended for other profiles that describe software.
- I would like to derive metadata information that is created or generated in other initiatives as much as possible in an automatic manner, with options for regular (automated) updates. Specifically, parts of

[32] http://cmdi-maker.uni-koeln.de/
[33] http://www.sfs.uni-tuebingen.de/nalida/proforma/
[34] http://clarino.uib.no/comedi/page

the metadata description should be derived automatically from CLAM[35] and WADL[36] descriptions for web services, and from descriptions originating from the codemeta[37] initiative.

I hope to work on these issues in the CLARIAH-PLUS project.

7.3 Recommendations

I end with some recommendations. Some of these follow directly from issues raised earlier, others were not mentioned before but result from our experiences in working with metadata:

- (to CLARIN ERIC) Set up a faceted search in the VLO dedicated to the discovery of software. The proposal sketched here can form a basis to start from.
- (to national coordinators) Coordinate metadata creation nationally. If every individual researcher or data centre manager creates metadata in isolation, the resulting metadata will be very diverse, use mutually incompatible vocabularies, vary enormously in quality and fine-grainedness, and will often lack important metadata information.
- (to national coordinators) Every national consortium must reserve effort (hence money) for active participation in the metadata curation task force. This is necessary because real work will only be done if people have been assigned an explicit task and are paid for the work they do.
- (to CLARIN ERIC) CLARIN should define a minimum set of metadata elements (defined semantically):
  - separately for data and for software
  - separately for faceted search and for a minimal proper description of the data or software

  Procedures and supporting software should be set up for testing compliance to these requirements, and deviations should only be allowed in exceptional cases. This is an extension of the work already started by the Austrian national consortium ((Ostojic et al., 2017). The metadata curation task force should coordinate this.
- (to profile and component developers) Use closed ('half-open') vocabularies whenever possible, but be prepared to update them regularly and to upgrade them occasionally
- (to the developers of CMDI) Enable the definition of closed vocabularies outside of a CMDI metadata element. Ensure that such vocabularies can be reused by others in multiple elements. Ensure that viewing and copying the values should be possible in the Component Registry without having to edit.
- (to CLARIN ERIC) There is a real need for a good CMDI editor, which is preferably not web-based, and enables editing of multiple files at once (both 'horizontally', i.e. all properties of one entry at a time, and 'vertically', i.e. to fill a property for a range of entries).
- (to the developers of the VLO and the CCR) It should be possible to use the 'isa' relation in the CCR to define small taxonomies of concepts, which can then be used in the faceted search to present the possible values of a facet in a hierarchical way, so that users see only a small list to select from and are only confronted with fine-grained distinctions when they are relevant to them. The CLARIN-NL Portal, CLARIN Services part[38] illustrates such hierarchical facet values. Such taxonomies will also be beneficial for profile and component editors, and for dedicated CMDI metadata editors.

Acknowledgements

The work on metadata for tools described here started already in 2012 but has been interrupted several times. Many people have worked with me on the profile and the metadata descriptions, in particular Eline Westerhout and Rogier Kraf. Eric Renckens wrote many of the descriptions on the CLARIN-NL Portal pages that formed the basis for these metadata descriptions. Daan Broeder created the faceted search in the *CLARIN in the Netherlands* Portal. I am indebted to Menzo Windhouwer and Twan Goosen for their excellent support. The developers of the software and the CLARIN Centre managers hosting the software and their metadata provided and/or corrected the information contained in the metadata descriptions.

[35] https://proycon.github.io/clam/
[36] https://javaee.github.io/wadl/
[37] https://codemeta.github.io/
[38] http://portal.clarin.nl/clarin-resource-list-fs.

References

[Broeder et al.2010] D. Broeder, M. Kemps-Snijders, D. Van Uytvanck, M. Windhouwer, P. Withers, P. Wittenburg, and C. Zinn. 2010. A data category registry- and component-based metadata framework. In N. Calzolari, B. Maegaard, J. Mariani, J. Odijk, K. Choukri, S. Piperidis, M. Rosner, and D. Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, pages 43–47, Valetta, Malta. European Language Resources Association (ELRA).

[Broeder et al.2012] Daan Broeder, Menzo Windhouwer, Dieter Van Uytvanck, Twan Goosen, and Thorsten Trippel. 2012. CMDI: A component metadata infrastructure. In *Proceedings of the LREC workshop 'Describing LRs with Metadata: Towards Flexibility and Interoperability in the Documentation of LR'.*, pages 1–4, Istanbul, Tyrkey. European Language Resources Association (ELRA).

[Hinrichs et al.2010] Erhard W. Hinrichs, Marie Hinrichs, and Thomas Zastrow. 2010. WebLicht: Web-Based LRT Services for German. In *Proceedings of the ACL 2010 System Demonstrations*, pages 25–29.

[Lyse et al.2015] Gunn Inger Lyse, Paul Meurer, and Koenraad De Smedt. 2015. COMEDI: A component metadata editor. In Jan Odijk, editor, *Selected Papers from the CLARIN 2014 Conference*, volume 28 of *NEALT Proceedings Series*, pages 82–98, Linköping, Sweden. Linköping Electronic Conference Proceedings. http://www.ep.liu.se/ecp/116/008/ecp15116008.pdf.

[Odijk2009] Jan Odijk. 2009. Data categories and ISOCAT: some remarks from a simple linguist. Presentation given at FLaReNet/CLARIN Standards Workshop, Helsinki, 30 September.

[Odijk2014] Jan Odijk. 2014. Discovering resources in CLARIN: Problems and suggestions for solutions. unpublished article, Utrecht University. http://dspace.library.uu.nl/handle/1874/303788, August.

[Odijk2015] Jan Odijk. 2015. Metadata curation strategy. manuscript, Utrecht, http://www.clarin.nl/sites/default/files/Metadata%20curation%20strategy%202015-06-29.pdf. Appendixes: http://www.clarin.nl/sites/default/files/Resource%20Type%20Curation%202015-6-29.xlsx and http://www.clarin.nl/sites/default/files/modality%20cleanup.xlsx, June 29.

[Odijk2018] Jan Odijk. 2018. Why I do not like web interfaces for data entry. Working paper, Utrecht University, October 11.

[Ostojic et al.2017] Davor Ostojic, Go Sugimoto, and Matej Ďurčo. 2017. The curation module and statistical analysis on VLO metadata quality. In *Selected papers from the CLARIN Annual Conference 2016, Aix-en-Provence, 26–28 October 2016*, number 136 in Linköping Electronic Conference Proceedings, pages 90–101. Linköping University Electronic Press, Linköpings Universitet.

[Schuurman et al.2016] Ineke Schuurman, Menzo Windhouwer, Oddrun Ohren, and Daniel Zeman. 2016. CLARIN Concept Registry: The New Semantic Registry. In Koenraad De Smedt, editor, *Selected Papers from the CLARIN Annual Conference 2015, October 14–16, 2015, Wroclaw, Poland*, number 123 in Linköping Electronic Conference Proceedings, pages 62–70, Linköping, Sweden. CLARIN, Linköping University Electronic Press. http://www.ep.liu.se/ecp/article.asp?issue=123&article=004.

[van den Bosch et al.2007] A. van den Bosch, G.J. Busser, W. Daelemans, and S. Canisius. 2007. An efficient memory-based morphosyntactic tagger and parser for Dutch. In F. Van Eynde, P. Dirix, I. Schuurman, and V. Vandeghinste, editors, *Selected Papers of the 17th Computational Linguistics in the Netherlands Meeting*, pages 99–114. Leuven, Belgium.

[Van Uytvanck2014] Dieter Van Uytvanck. 2014. How can I find resources using CLARIN? Presentation held at the *Using CLARIN for Digital Research* tutorial workshop at the *2014 Digital Humanities Conference*, Lausanne, Switzerland. https://www.clarin.eu/sites/default/files/CLARIN-dvu-dh2014_VLO.pdf, July.

[Westerhout and Odijk2013] Eline Westerhout and Jan Odijk. 2013. Metadata for tools: creating a CMDI profile for tools. Presentation held at CLIN 2013, Enschede, the Netherlands. http://www.clarin.nl/sites/default/files/13CLIN.pdf, 18January.

[Zeeman and Windhouwer2018] Rob Zeeman and Menzo Windhouwer. 2018. Tweak your CMDI forms to the max. Presentation at the CLARIN Annual Conference, Pisa, Italy. https://www.clarin.eu/sites/default/files/CLARIN2018_Session-4-5_Paper-22_Zeeman-Windhouwer.pdf, October10.

[Zinn2016a] Claus Zinn. 2016a. The CLARIN language resource switchboard. https://www.clarin.eu/sites/default/files/08%20-%20ZINN-Lg-Sw-Board.pdf. Presentation at the CLARIN 2016 Annual Conference.

[Zinn2016b] Claus Zinn. 2016b. The CLARIN language resource switchboard. https://www.clarin.eu/sites/default/files/zinn-CLARIN2016_paper_26.pdf. Abstract for the CLARIN 2016 Annual Conference.

[Zinn2017] Claus Zinn. 2017. A bridge from EUDAT's B2DROP cloud service to CLARIN's language resource switchboard. https://www.clarin.eu/sites/default/files/Zinn-CLARIN2017_paper_17.pdf. Abstract for the CLARIN 2017 Annual Conference.

Media Suite: Unlocking Audiovisual Archives for Mixed Media Scholarly Research

Roeland Ordelman
Netherlands Institute for Sound and Vision
University of Twente
The Netherlands
rordelman@beeldengeluid.nl

Liliana Melgar
Department of Media Studies
University of Amsterdam
The Netherlands
melgar@uva.nl

Jasmijn Van Gorp
Department of Media and Culture Studies
Utrecht University
The Netherlands
j.vangorp@uu.nl

Julia Noordegraaf
Department of Media Studies
University of Amsterdam
The Netherlands
J.J.noordegraaf@uva.nl

Abstract

This paper discusses the rationale behind and approach towards the development of a research environment –the *Media Suite*– in a sustainable, dynamic, multi-institutional infrastructure that supports mixed media scholarly research with large audiovisual data collections and available multimedia context collections, serving media scholars and digital humanists in general.

1 Introduction

In some domains of scholarly research, the focus is on the creation of new data collections. In other domains, for example, in Media Studies (e.g., film and television studies) research often focuses on data collections maintained at cultural heritage institutions, such as archives, libraries, and other knowledge institutions. However, especially when audiovisual media are concerned, access to, and use of these collections is often restricted due to intellectual property rights (IPR) or privacy issues (e.g., with respect to recorded interviews). Moreover, individual institutions often do not have the technical infrastructure in place to serve basic scholarly needs with respect to search, exploration and inspection of individual items (i.e., play-out or viewing). Therefore, scholars either fall back on collections that are openly available or spend considerable amounts of time in *on-site* visits to archives for consulting data collections (Bron et al., 2016). Data collections at these institutes can be regarded as "locked", or at least hard to use for scholarly research.

To unlock these "institutional" collections and let scholars take advantage of the sheer quantity and richness of these data sets, we are developing an infrastructure for *online* scholarly exploration of collections that are distributed across various institutional content owners. Specifically, we focus on *audiovisual* data collections and related *multimedia* sources, such as radio and television broadcasts, films, oral history interviews, but also (news)paper archives, film posters and eyewitness reports. An online application, named *Media Suite*[1], serves as the interface to this underlying infrastructure, where content and metadata can be explored, browsed, compared, and where personal virtual collections can be compiled and stored in a personal workspace. In this workspace, scholars have additional tools for working with these *mixed media* collections, such as tools for automatic annotation, visualisation, analysis, and sharing.

The ultimate goal of developing the *Media Suite* and its infrastructure, is to (i) enable distant reading (Moretti, 2013), that is, identifying patterns or new research questions in and across aggregated collections, (ii) facilitate close reading: the detailed examination of individual items (e.g., videos) in a

[1] http://mediasuite.clariah.nl

Roeland Ordelman, Liliana Melgar, Carlos Martinez-Ortiz, Julia Noordegraaf and Jaap Blom 2019. Media Suite: Unlocking Archives for Mixed Media Scholarly Research. *Selected papers from the CLARIN Annual Conference 2018*. Linköping Electronic Conference Proceedings 159: 133–143.

collection or specific sections of these items (e.g., video segments) during search and scholarly interpretation, and (iii) make sure that the "scholarly primitives" (Unsworth, 2000; Blanke and Hedges, 2013), basic activities such as "discovering", "annotating", "comparing" and "storing", common to research across humanities disciplines, are well supported.

In pursuit of these goals, we face challenges on various levels, broadly identified as appropriate access to *data* and *tools*. Appropriate data access entails the ability to view and browse individual media objects for close reading, thus requiring solutions for accessing copyrighted materials or materials for which access is restricted due to privacy reasons. Second, appropriate data access is about "searchability", the availability of fine-grained metadata for retrieval, and about the required insight into the quality of the sources and the metadata needed for data analytics. Metadata, traditionally created manually for the data sets scholars are interested in, is typically sparse, quite diverse, and often incomplete. Apart from indexing this metadata properly, providing insight into this diversity is crucial for scholars to assess the quality of a search result, its significance to a research question, or validity of an analysis. This is traditionally referred to as *source criticism*, and currently referred to as *digital source criticism* (Hoekstra et al., 2018).

Emerging methods to generate metadata automatically, using for example automatic speech recognition (ASR) or computer vision technology, may bridge the gap between metadata sparsity and distant reading requirements of scholars, but they also bring up technological and methodological challenges. For example, questions arise on how can we efficiently generate high quality metadata for large amounts of "locked" content using automatic metadata extraction technology, and how the use of this type of metadata –that may still have classification errors or may be sensitive for biases– have to be accounted for in the interpretation of results and thus impact the methodology of scholars. Raising awareness about the operation of computational instruments for data extraction and processing and their impact on the heuristics and results of data-driven research is referred to as *digital tool criticism* (Koolen et al., 2018).

Eventually, enabling scholarly research that supports source and tool criticism should be reflected in the design of a user-interface that balances ease of use with the need to provide transparency regarding the scope and quality of the underlying data and their processing. As scholars have a wide variety of research interests, and also, have different levels of computer literacy –hence skills, or lack of skills to apply specific data processing tools themselves, for example for creating visualisations or applying content analysis tools–, the *interaction* with data and tools should be balanced accordingly: allowing for specific, specialised functions from individual scholars, without impeding the generic functions that apply to a wider community.

To solve the locked data problem and still allow for a flexible interaction with data and tools, the central approach of the *Media Suite* is to "bring the tools to the data" –as opposed to "bringing the data to the tools" that is custom in many other research areas– and to provide mechanisms that enable researchers to work with data and tools *within* the closed environment of the infrastructure, sealed with a federated authentication mechanism. In the past, substantial effort was undertaken to develop specific tooling that eventually could not be connected properly to work with the data collections they were intended for, due to access restrictions. In that sense, the *Media Suite* functions as a "virtual research environment" (VRE) that facilitates the proper functioning of the tools in the context of research and cultural heritage institutions. As a consequence, this research environment has a special liability towards the data and tools it provides in terms of transparency (source criticism), credibility (tool criticism) and flexibility.

Figure 1 shows the main elements that constitute the *Media Suite* research environment. Below we discuss shortly each of these elements.

2 Data Sources – Data Governance

The *Media Suite* currently provides access to audiovisual collections and multimedia context collections[2] from the following institutions, among others: (a) The Netherlands Institute for Sound and Vision (NISV), offering about a million hours of radio and television, film and oral history collections, including photos and digitised program guides and audience ratings), (b) Eye film institute, initially providing

[2]http://mediasuitedata.clariah.nl/dataset

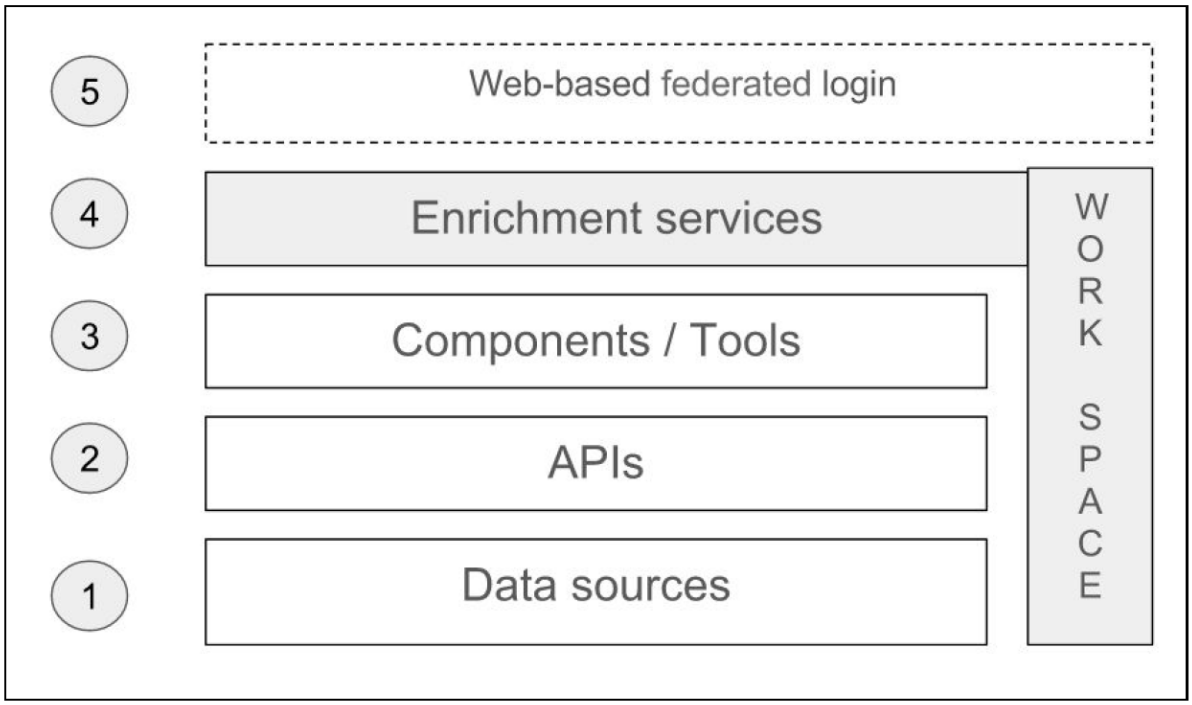


Figure 1: The building blocks of the CLARIAH Media Suite

access to the UNESCO world heritage Jean Desmet collection, including films, paper and poster collections, and (c) Oral History collections from various organisations in the Netherlands deposited at DANS. Also, although not an audiovisual collection, the large Dutch newspapers collection (more than 100 million pages) from the Dutch National Library is an important part of the *Media Suite*, as it allows scholars to make comparisons across media types.

To make these collections available in the *Media Suite*, we adhere to the general principle that collection owners provide access to collection metadata via the OAI-PMH protocol that enables the *Media Suite* to harvest the metadata and index it. It is assumed that the link to the source data (e.g., a video, scan or transcription) is incorporated in the metadata and points to a (streaming) sever hosted by the collection owner. Access restrictions (i.e., who is allowed to access what), is then organised at a broader –currently only national but ultimately international– research infrastructure level (CLARIAH[3], CLARIN[4]), via authentication and authorization protocols. In an ideal scenario, collection owners register and update their collections in a collection registry (we currently use CKAN[5]), that is "read" by the *Media Suite* for harvesting[6]. In practice however, we often have had to adapt this approach to the reality of sub-optimal situation with respect tot data governance at institutions. Institutional collection maintainers have internal data governance processes to ensure that data assets are formally managed. Data governance with respect to external processes – loosely defined as being part of an 'infrastructure' – is typically not accounted for at the institutions involved. This means that key data governance areas such as availability (e.g., metadata can be harvested), usability (e.g., source data can be viewed), integrity (e.g., protocols are in place to handle duplication and enrichment), and security (e.g., provenance information is maintained), need to be (re)organised or (re)considered, formalised and supported by the *Media Suite* and the emerging infrastructure in which it is embedded. From the practical point of view of making collections available

[3] https://www.clariah.nl/
[4] https://www.clarin.eu/
[5] https://ckan.org/
[6] http://mediasuitedata.clariah.nl/

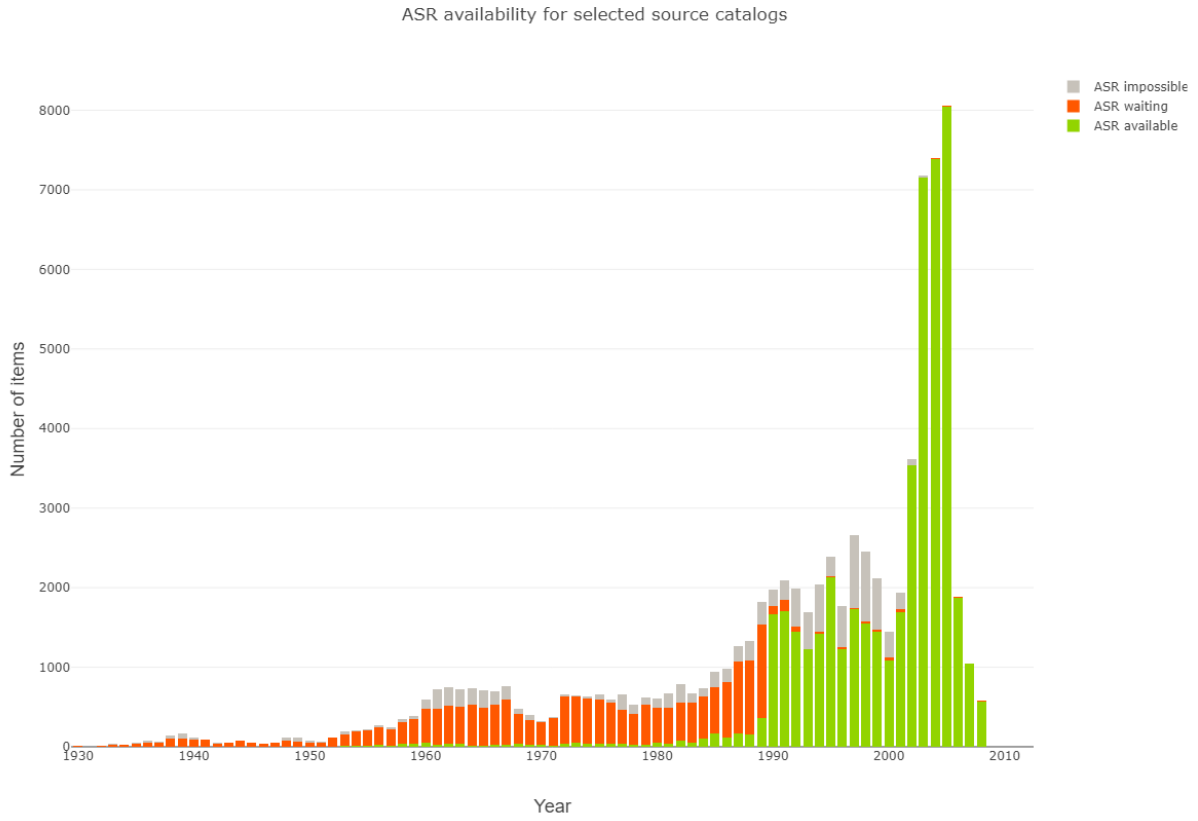


Figure 2: This chart shows the availability of automatic speech recognition (ASR) transcripts for the source catalogs per year. The green bars show the material that has ASR, the orange bars show the material that does not have ASR yet, the grey bars show the material for which ASR is currently impossible (no digital version). Screenshot date: January 2019

in the *Media Suite*, this drills down to manually mapping the various metadata models to a coherent search index and unravel technical issues for each collection individually.

The wish to use automatic metadata extraction technology to improve fine-grained, time-based access to audiovisual content is an additional complication in the provisioned distributed data chain. Here, the model is to provide services such as automatic speech recognition (Ordelman and van Hessen, 2018) for collections (”bulk” processing, up to 100 hours or more) via the infrastructure. However, making such a service available in a robust and collection-owner-friendly manner –a rather complex endeavour in itself– is only part of the work. Collection owners also need to arrange and manage internal data workflows: feeding the service with content and incorporating the output of the service (e.g, time-coded speech recognition transcripts) in the existing metadata model (including provenance information). The current status is that we have a speech recognition service available in the infrastructure tat operates faster then real-time and capable of processing approximately 1.000 hours per day. It is currently processing the NISV archive going backwards in time but also taking priority requests from scholars (e.g., to process news and actualities first). See Figure 2 that gives an impression of the progress for the NISV catalogs.

For the upcoming years, the goal is to connect more data collections from individual collection owners in the Netherlands, and increase the quantity and quality of metadata, focusing on both internal data governance processes and the use of automatic metadata extraction technology. Also, the incorporation of social media data, in particular data that are related to the ”traditional” media collections (e.g., hashtags related to television programming), is targeted. Finally, we want to make it possible for individual

scholars to upload their own data sets. Although the focus of the infrastructure is especially on opening up the ”locked” institutional collections, we noticed that scholars may also want to include in their analysis data sets coming from elsewhere or created by an individual scholar for a specific purpose, such as social media data on a specific topic, or recorded interviews.

3 Sustainable development (APIs)

The development of a digital infrastructure that is ”sustainable”, to make sure that it will remain available and maintained in the long run, including support, updates and upgrades, is central to the CLARIAH project as a whole, and specifically to the CLARIAH centres appointed by the project to support their domain-specific parts of the infrastructure [7] and to foster interoperability between these parts. For the *Media Suite*, this part of the infrastructure covers Dutch audiovisual collections augmented with available multimedia context collections. Examples of interoperability, are the connection with the CLARIAH infrastructure that focuses on textual data, containing the newspaper collections of the Dutch National Library that were mentioned before, and initial steps to link collections via Linked Open Data.

To foster sustainability we have to find a middle ground between the wishes of scholars and institutional ICT development and maintenance frameworks. Another critical requirement in this context is that the research infrastructure should also comply with other types of infrastructures that are being developed, such as in The Netherlands the infrastructure for digital heritage (Network Digital Heritage - NDE[8]) and, in a European context, the infrastructure components developed in the CLARIN and DARIAH ERICs.

Practically, this means that the infrastructure adheres to existing protocols, conventions, and standards. Moreover, to warrant interoperability and avoid proliferation of functions and processes (resulting in what is sometimes called a cauliflower architecture), a –from a research project point of view– rather strict development regime is followed, enforced by sprint plannings, focusing on a modular organisation of *Media Suite* components via application programming interfaces (APIs) that can be shared within the infrastructure. Examples of these APIs are a Collection API that provides high-level information (metadata) about the collections, such as which collections are available, data format, and volume, a Search API that allows searching the available collection indices, and the Annotation API that provides functionality for data annotation using the W3C Web Annotation data model (Sanderson et al., 2017).

4 Tools and user-friendly interaction design

The APIs discussed above are the corner stones for the development of the tools needed by scholars for doing their research. The development of these tools is to a large extent driven by requirements that were articulated in prototype applications built in earlier projects, such as video search and comparative analysis of media in *AVResearcherXL* (Van Gorp et al., 2015), search and visualization of results in *TROVe*[9], multi-collection search in *CoMeRDa* (Bron et al., 2013), exploratory search in *DIVE* (De Boer et al., 2015) and Oral History research in *Verteld Verleden* (Ordelman and de Jong, 2011). With a few exceptions and some ongoing work, the methods and functions underlying these prototypes have been extracted and re-implemented in the *Media Suite*.

The digital humanities community incorporates a wide diversity of scholars with different research questions, methods, and levels of expertise in working with information processing techniques and technologies. To address the challenges this imposes on requirements elicitation, development and evaluation of both re-implementations and new tools, the *Media Suite* team follows the principles of co-development where programmers and researchers work closely together, involving also the research community immediately via component testing, hackathons, datathons, public fora, and workshops. Because the use scenarios of scholars are diverse, it is even more important to focus on the similarities in research methods from different disciplines (de Jong et al., 2011; Melgar Estrada and Koolen, 2018), and to take

[7] https://clariah.nl/over/organisatie/centra
[8] https://www.netwerkdigitaalerfgoed.nl/en/
[9] https://www.clariah.nl/en/projects/finished/seed-money/trove

Figure 3: Collection Inspector: metadata information and completeness per field

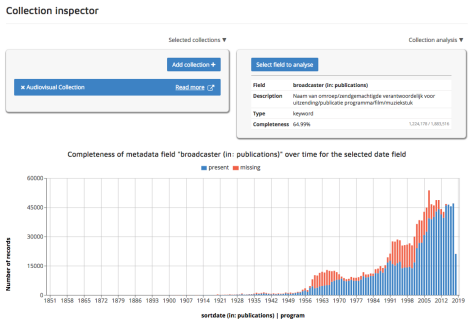


Figure 4: Collection Inspector: metadata completeness per field over time

these similarities as the baseline for tool development. Additionally, the concept of the scholarly primitives (Unsworth, 2000; Blanke and Hedges, 2013) serves as valuable guidance for identifying must-have functions in the *Media Suite*, and a model for a coherent and user-friendly design of the interface that fits in with the daily practice of scholars. Finally, as developing new tools "from scratch" for every research question would be a very inefficient (and costly!) endeavour, the analysis of tools that are "out there" has been taken up, resulting for example in a comparative study of qualitative data analysis software (Melgar et al., 2017), that provides the clues for deciding which tools we will or will not build ourselves and what type of data export functions and formats to support, of course within the boundaries of copyright and privacy restrictions.

A tool that was not directly based on previous work but actually emerged from working with "real" data has been coined as the Collection Inspector tool. As referred to in section 2, the metadata of collections from various collection owners is very heterogeneous, may not be complete, and may require some "metadata archaeology" to find out the proper meaning of fields, a meaning that may have changed in the course of an archive's history due to protocol and vendor changes. From a search perspective –the *Media Suite* allows scholars to design their own facets or filters based on available metadata fields– incompleteness and meaning of these fields is highly significant, and may lead to misinterpretations, for example with respect to recall, the search equivalent for completeness. The Collection Inspector enables scholars to assess the collection metadata, providing field descriptions, type, overall completeness, and completeness over time. Figures 3 and 4 show screen-shots of the Collection Inspector, on the left the descriptions and overall completeness data per field, on the right completeness of a single field over time. Together with the before-mentioned collection registry tool (CKAN), which contains information and visualisations that provide aggregate views on the content, scope and quality of the collections as well as their digital processing, and the options for scholars to define their own metadata filters, the collection inspector tool brings a valuable facility to the *Media Suite* for conducting digital source criticism.

Working with real data and the possibility to access (viewing/listening) the content itself was often very limited in the earlier prototypes due to the "locked data" problem. This underlined the importance of a well-thought-out design of content viewing/listening functions in relation to other functions that are associated with content-level, or in retrieval terms, document-level access, such as annotation, document level browsing, and within/cross-collection linking and recommendation. We grouped such functions in what we call the "Resource viewer" tool that currently incorporates playing video (also full-screen) and audio, annotation (see Figure 5), within-document browsing based on time-coded metadata such as speech transcripts (see Figure 6), and browsing all available metadata for the resource. However, while working with the Resource Viewer, scholars immediately suggest several opportunities to enhance "distant" reading on the document level. Note that audiovisual documents can be long and lack structure

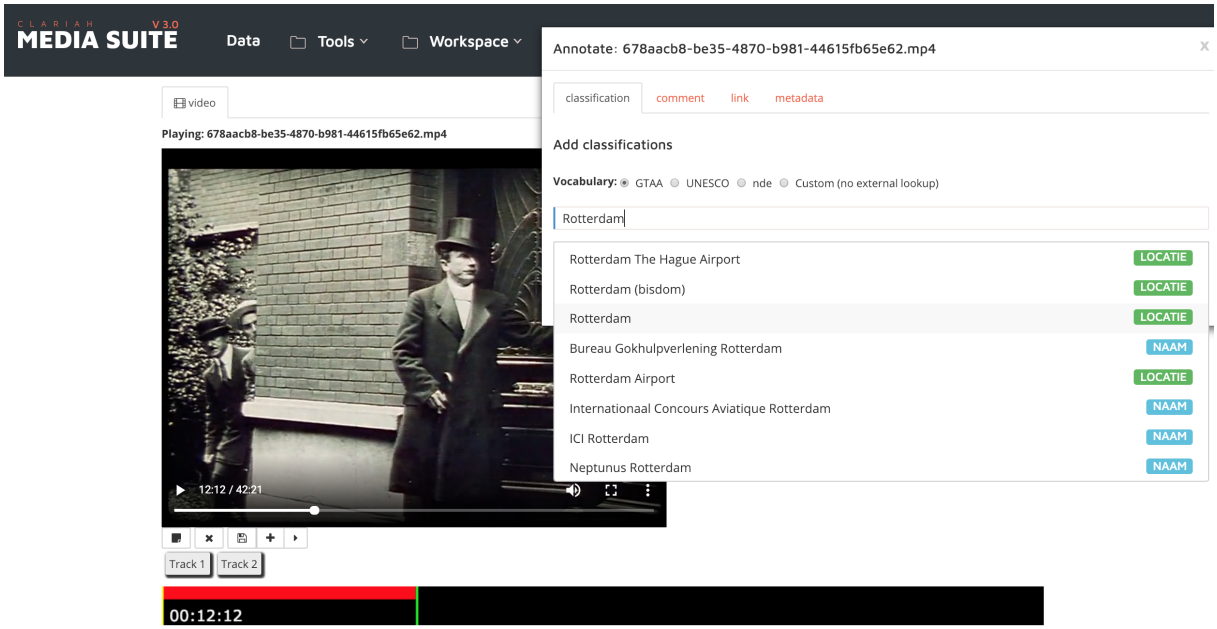


Figure 5: Annotating a video item in the Resource Viewer of the Media Suite using the NISV Audiovisual Thesaurus (GTAA) to label a segment, in this case with the location ”Rotterdam”.

such as paragraphs and headings in text. Word clouds, segmentation (e.g., based on shots) and segment-labelling (e.g., based on topic detection), or summaries based on speaker or face recognition, could alleviate this lack of structure and smooth the analysis of individual resources.

Zooming in from tools that put more emphasis on the distant reading part, as has been the focus until recently (data registry, collection inspection, search end exploration facilities as shown in Table 1) to tools that operate on levels of resource analysis and close reading, the further development of the Resource viewer requires special attention in the forthcoming period.

5 Workspace – working with personal virtual collections

In addition to copyright and privacy restrictions, access to the audiovisual content in the *Media Suite* is also limited due to its nature; consisting of pixels (video) and samples (audio) and hopefully some manually generated metadata or subtitles (text). Typically, scholars want to search audiovisual data using (key)words that may be ‘hidden’ (encoded) in the pixels or the samples. This is called the semantic gap (Smeulders et al., 2000) that needs to be “bridged” by decoding the information in the pixels and the samples to semantic representations, e.g., a verbatim transcription of the speech or labels of visual concepts in the video (a car, a face, the Eiffel Tower), that can be matched with the keywords from the scholars. These semantic representations can be generated manually or, especially when data collections are large, automatically using automatic speech recognition (ASR) or computer vision technology.

The generation of semantic representations is addressed in different ways. One the one hand, tools such as ASR are regarded as ‘must have’ components in an infrastructure focusing on fine-grained access and ’distant reading’ of large data sets. We are implementing an automatic speech recognition service that resides within the CLARIAH infrastructure and that can handle requests from the infrastructure itself (e.g., bulk processing of collections, possibly activated by a scholar with an interest in a specific data set), but also requests from individual scholars that want to process their private collections. On the other hand, supporting manual annotation is key for interpretation in scholarly contexts.

The *Media Suite* aims to support the generation of both ways of semantic representations in complementary ways via information work-flows centred around a workspace. More in general, the workspace

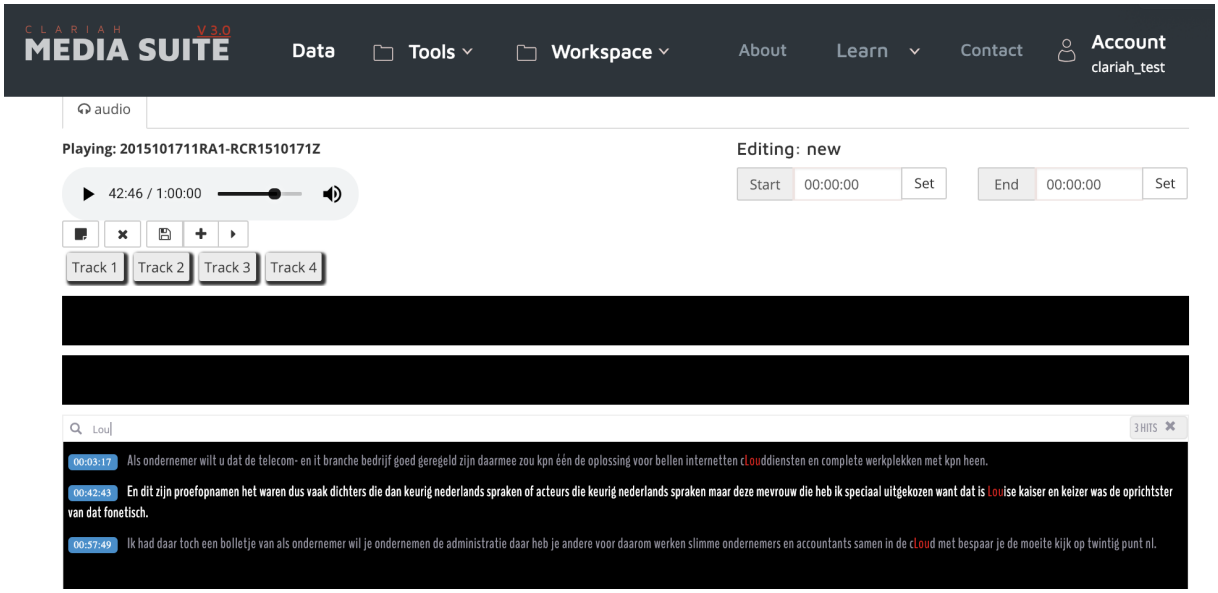


Figure 6: Browsing a radio item via speech transcripts in the Resource Viewer of the Media Suite. Via a within-document search function users are able to jump to specific parts of the radio broadcast, in this case where ”Louise Kaiser” is mentioned.

serves as the foundation for a scholar's research projects from which s/he can create and describe projects, organise collaborations (e.g., with peers or students), and keep track of all data –bookmarked selections from collections, annotations, visualisations– related to the projects. Figure 7 shows a screen-shot from the workspace, in this example on a project that is about media presentations of the 1953 flood disaster in The Netherlands. From within the workspace, a scholar can directly access the resource viewer for a stored resource, access saved session data such as queries and filtering options, and upload external data that are relevant for a project.

A special facility in the workspace is the option to generate your own visualisations using Jupyter Notebooks[10] and the (protected) APIs developed in the infrastructure. Jupyter Notebooks serve as a programming interface that allows scholars with programming skills to write their own code for creating overviews of the data, investigating a section of interest, performing advanced data analysis, and generating complex visualisations. In this way, we bring programming facilities to the data and to use third party code such as visualisation libraries and language processing toolkits (Wigham et al., 2018) to extend and complement their use of the *Media Suite*'s graphical user interface (Melgar et al., 2019).

6 Conclusion and future work

We described the *Media Suite* and its underlying infrastructure, and the challenges in building such an infrastructure that satisfies the needs of humanities scholars working with audio-visual media and contextual collections. We chose the approach of building a research environment that adheres to infrastructural requirements while at the same time being flexible, transparent, and user-friendly. In order to develop this environment in a sustainable way, that can be used and developed further after the project's lifetime, we need to carefully align the requirements of scholars with the context of the ecosystem the *Media Suite* needs to live in: an ICT infrastructure hosted and maintained by multiple institutions that in turn, adheres to a diverse set of institutional requirements with respect to, for instance, data access permissions and software development and maintenance. In order to have this infrastructure it is required that it is generic enough to cater for the general needs of every group that we have identified, while at the same time it

[10] `https://jupyter.org/`

Discovering	Overview of available collections via the collection registry (CKAN); Advanced search with options for filtering; Segment-level search on the basis of time-coded speech transcripts; Exploratory search via linked open data; A resource viewer for viewing and analysis of individual media items; Automatic metadata extraction technology
Annotating	Time/space-based multimedia annotation including segmenting, commenting, adding user metadata, links to other information sources, and use of code-book/thesaurus labels.
Comparing	Cross-media and cross-collection comparisons via saved queries
Sampling	Create personal virtual collections based on selections (bookmarks) stored in a personal workspace (see also section 5 below)
Illustrating	Generic visualisations of search results, flexible creation of ad-hoc visualisations using Jupyter Notebooks (see also section 5 below)
Representing	Understood as the need to support the ”presentation” phase of research, for example via enhanced publications with links to *Media Suite* content on the segment level (Van Den Heuvel et al., 2010). Support by the *Media Suite* is currently limited, as the infrastructure still lacks options for generating persistent identifiers on the segment-level

Table 1: Media Suite tools categorised via scholarly primitives

incorporates flexible functionality capable of addressing very specialised research questions. The *Media Suite* is currently functional and used by scholars doing actual research projects. Further development will focus on improving the current implementation of functions (e.g., development of a CLARIAH-wide annotation client[11], various interface improvements), adding collections, including new types such as social media data, increasing metadata granularity using automatic metadata extraction (e.g., speaker labelling, face recognition), and in particular, enhancing the functionality of the Resource viewer and Workspace. Also, we intend to setup a large system evaluation by a group of users outside the project to benchmark the current version of the system.

7 Acknowledgements

The research for this paper was made possible by the CLARIAH-CORE project (`www.clariah.nl`) financed by NWO. This paper is the result of a joint effort, specifically a close collaboration between all scholars, software developers, and domain specialists, listed here: `http://mediasuite.clariah.nl/documentation/faq/who-develops`.

[11]For work in progress see: `https://clariah.github.io/scholarly-web-annotation-client/`

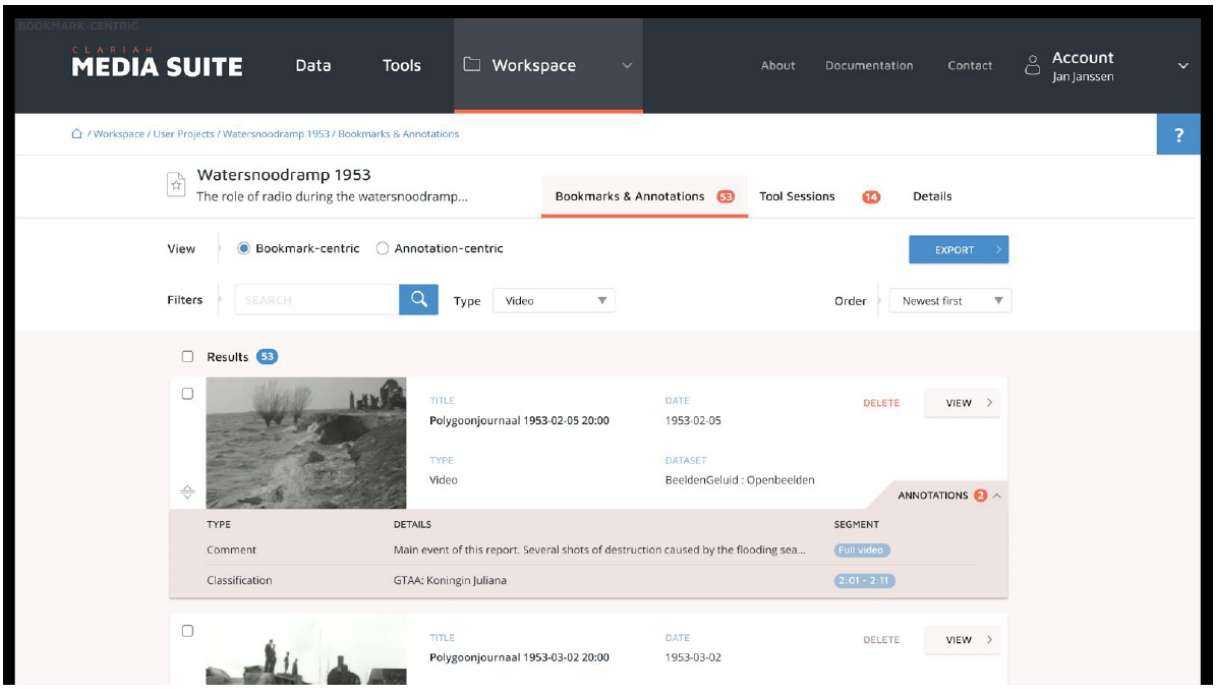


Figure 7: The CLARIAH Media Suite's Workspace

References

Tobias Blanke and Mark Hedges. 2013. Scholarly primitives: Building institutional infrastructure for humanities e-science. *Future Generation Computer Systems*, 29(2):654–661.

Marc Bron, Jasmijn Van Gorp, Frank Nack, Lotte Belice Baltussen, and Maarten de Rijke. 2013. Aggregated search interface preferences in multi-session search tasks. In *SIGIR '13: 36th international ACM SIGIR conference on Research and development in information retrieval*. ACM, July.

Marc Bron, Jasmijn Van Gorp, and Maarten de Rijke. 2016. Media studies research in the data-driven age: How research questions evolve. *Journal of the Association for Information Science and Technology*, 67(7):1535–1554.

Victor De Boer, Johan Oomen, Oana Inel, Lora Aroyo, Elco Van Staveren, Werner Helmich, and Dennis De Beurs. 2015. Dive into the event-based browsing of linked historical media. *Web Semantics: Science, Services and Agents on the World Wide Web*, 35:152–158.

Franciska de Jong, Roeland Ordelman, and Stef Scagliola. 2011. Audio-visual collections and the user needs of scholars in the humanities: a case for co-development. In *Proceedings of the 2nd Conference on Supporting Digital Humanities*, November.

FG Hoekstra, Marijn Koolen, and Marijke van Faassen. 2018. Data scopes: towards transparent data research in digital humanities. *Digital Humanities 2018 Puentes-Bridges*.

Marijn Koolen, Jasmijn van Gorp, and Jacco van Ossenbruggen. 2018. Toward a model for digital tool criticism: Reflection as integrative practice. *Digital Scholarship in the Humanities*.

Liliana Melgar, Marijn Koolen, Hugo Huurdeman, and Jaap Blom. 2017. A process model of scholarly media annotation. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, CHIIR '17, pages 305–308, New York, NY, USA. ACM.

Liliana Melgar, Marijn Koolen, Kaspar van Beelen, Hugo Huurdeman, Mari Wigham, Carlos Martínez Ortíz, and Roeland Ordelman. 2019. The CLARIAH Media Suite: a hybrid approach to system design in the humanities. In *CHIIR 2019: ACM SIGIR Conference on Human Information Interaction and Retrieval*, Glasgow, Scotland, UK.

Liliana Melgar Estrada and Marijn Koolen. 2018. Audiovisual media annotation using qualitative data analysis software: A comparative analysis. *The Qualitative Report*, 23(13):40–60.

Franco Moretti. 2013. *Distant reading*. Verso Books, London.

Roeland J.F. Ordelman and Franciska M.G. de Jong. 2011. Distributed access to oral history collections: Fitting access technology to the needs of collection owners and researchers. In *Digital Humanities 2011: Conference Abstracts*, pages 347–349. Stanford University Library.

Roeland Ordelman and Arjan van Hessen. 2018. Speech recognition and scholarly research: Usability and sustainability. In *CLARIN 2018 Annual Conference*, pages 163–168.

Robert Sanderson, Paolo Ciccarese, and Benjamin Young. 2017. Web annotation data model. *W3C Candidate Recommendation*.

Arnold WM Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. 2000. Content-based image retrieval at the end of the early years. *IEEE Transactions on pattern analysis and machine intelligence*, 22(12):1349–1380.

John Unsworth. 2000. Scholarly primitives: What methods do humanities researchers have in common, and how might our tools reflect this. In *Symposium on Humanities Computing: Formal Methods, Experimental Practice. King's College, London*, volume 13, pages 5–00.

Henk Van Den Heuvel, René Van Horik, Stef Scagliola, Eric Sanders, and Paula Witkamp. 2010. The veterantapes: Research corpus, fragment processing tool, and enhanced publications for the e-humanities. In *LREC*.

Jasmijn Van Gorp, Sonja de Leeuw, Justin van Wees, and Bouke Huurnink. 2015. Digital media archaeology: Digging into the digital tool avresearcherxl. *VIEW. Journal of European Television History and Culture/E-journal*, 4(7):38–53.

Mari Wigham, Liliana Melgar Estrada, and Roeland Ordelman. 2018. Jupyter Notebooks for generous archive interfaces. In *IEEE Big Data 2018: 3rd Computational Archival Science (CAS) Workshop*, Seattle, WA.

Curating and Analyzing Oral History Collections

Cord Pagenstecher
University Library, Center for Digital Systems
Freie Universität Berlin, Germany
cord.pagenstecher@cedis.fu-berlin.de

Abstract

This paper presents the digital interview collections available at Freie Universität Berlin, focusing on the online archive *Forced Labor 1939-1945*, and discusses the digital perspectives of curating and analyzing oral history collections. The digital curation of interview collections faces a number of problems like standards interoperability or privacy protection, but also chances built on progress in automatic speech recognition. Digital archives enhance the possibility of comparative studies. A pilot study based on two interviews from the Berlin collections, highlights differences in narrative performativity, in dialogical interaction, and in multilingualism. Finally, the paper looks at perspectives of interdisciplinary cooperation with CLARIN projects and at the challenges of cross-collection search and de-contextualization.

1 Interview collections at Freie Universität Berlin

Since 2006, the Center for Digital Systems (CeDiS) of Freie Universität Berlin has been creating or giving access to several major collections with testimonies focusing on the Second World War and Nazi atrocities. The *Visual History Archive* of the USC Shoah Foundation (www.vha.fu-berlin.de), the *Fortunoff Video Archive* of Yale University, the online interview archive *Forced Labor 1939–1945* (www.zwangsarbeit-archiv.de/en), the British-Jewish collection *Refugee Voices* (www.refugeevoices.fu-berlin.de), the *Archiv Deutsches Gedächtnis* of FernUniversität Hagen (deutsches-gedaechtnis.fernuni-hagen.de) and the new interview archive *Memories of the Occupation in Greece* (www.occupation-memories.org) contain thousands of audio-visual life-story interviews.

Some of these collections are only accessible in the library or the campus network of Freie Universität Berlin, others are presented online in new working environments. Contrary to other Oral History collections where much research still relies on written transcriptions, some of these platforms come with a time-coded alignment of transcriptions, media files, and metadata, and allow for thematically focused searches and annotations throughout the video-recordings. To make the recordings accessible for research, teaching, education and the general public, CeDiS has created translations, maps and learning applications giving didactical support for teachers and students. Additionally, its team is engaged in academic debates through publications and conferences on oral history and digital humanities (Apostolopoulos and Pagenstecher, 2013; Pagenstecher and Tausendfreund, 2015; Nägel, 2016; Apostolopoulos et al., 2016; Pagenstecher and Pfänder, 2017; Pagenstecher, 2018).

The oral history projects started when Freie Universität Berlin became the first full-access-site to the Shoah Foundation's *Visual History Archive* outside the United States. Numerous German research projects (Bothe and Brüning, 2015; Michaelis, 2013) and university courses are using the collection; large educational programs were developed and implemented in German schools (Pagenstecher and Wein, 2017). Whereas the Shoah Foundation initially had not transcribed its 53,000 interviews, CeDiS created 908 German-language (plus 50 foreign-language) transcriptions (http://transcripts.vha.fu-berlin.de). These transcripts are time-coded every minute enabling full text search over all 958 interviews (Abenhausen et al., 2012). The Shoah Foundation offers the German transcripts as a kind of subtitles within their online archive – if your university has subscribed with the *Visual History Ar-*

Cord Pagenstecher 2019. Curating and Analyzing Oral History Collections. *Selected papers from the CLARIN Annual Conference 2018*. Linköping Electronic Conference Proceedings 159: 144–151.

chive's new commercial provider ProQuest (ProQuest, 2017). In 2017, the Shoah Foundation provided another 984 transcripts in English in their online archive (USC Shoah Foundation, 2017).

2 The Online Archive *Forced Labor 1939-1945*

In a second step, Freie Universität Berlin created a sophisticated online platform for a new interview collection on Nazi forced labor. The interview archive *Forced Labor 1939-1945: Memory and History* commemorates more than 20 million people who were forced to work for the Reich.

590 former forced laborers tell their life stories in detailed audio and video interviews. Most of the interviews were conducted in the Ukraine, Poland, and Russia. About a third of the interviewees were prisoners of concentration camps – many of them Jews or Roma. The biographical interviews do not only relate to Nazi forced labor; they also touch upon various other historical aspects of the Century of Camps, from Holodomor to Perestroika, from the Spanish Civil War to the Yugoslav Wars.

The collection was initiated and financed by the foundation *Remembrance, Responsibility and Future*. The testimonies were recorded in 2005 and 2006 by 32 partner institutions in 25 countries (Plato et al., 2010). Most of them were transcribed, translated into German, indexed and made available in an online archive together with accompanying photos and documents (Apostolopoulos and Pagenstecher, 2013). The user interface is available in English, German and Russian. Users are required to register before they can access the full interviews online. Since 2009, almost 10,000 archive users – students, researchers, teachers, and other interested persons – have been granted access to the collection.

Faceted search options allow the user to filter the interviews for victims' groups, areas of deployment, places, camps and companies or language of interview. The time-coded alignment of transcriptions, translations and media files supports full-text search through the audio or video recordings. Thus, the user can jump directly to interview sequences concerning a specific topic or compare national or gender-specific narrations about different topics, for example sabotage in the camps.

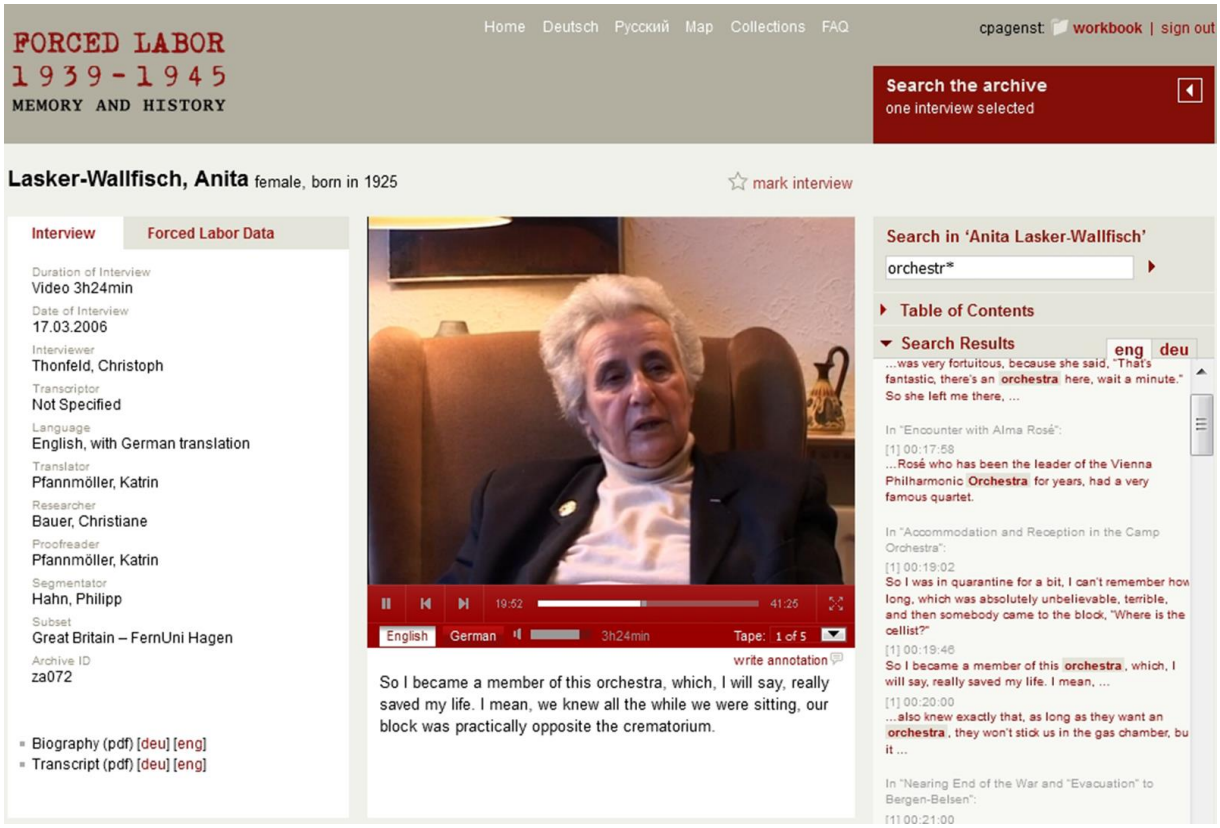


Figure 1: Interview with Holocaust survivor Anita Lasker-Wallfisch with metadata, subtitles and full-text search in the *Forced Labor* archive, www.zwangsarbeit-archiv.de, 30 Jan 2019

A map visualizes the interviewees' birthplaces and deployment locations and demonstrates the European dimensions of Nazi forced labor – and of post-war migration patterns. Using satellite imagery, the user can move from the geographical macro level to the topographical micro level by zooming in onto – vanished or preserved – barracks and factories. Through this form of visualization, digital mapping contextualizes the survivors' testimonies within current local cultures or memory – or forgetting.

In 2019, the archive got a new user interface supporting mobile devices and additional research options, including a register of persons, camps and factories linked to specific interview segments. Recent CeDiS projects with other collections use the same technology as the *Forced Labor 1939–1945* project, adding project-specific functionalities. With about 2,500 audio and video interviews, the *Archiv Deutsches Gedächtnis* at FernUniversität Hagen is the largest collection of oral history interviews in Germany (Gref et al., 2018). Containing narrative language data from many different research and documentation projects over the last four decades, it is being digitized and made available step-by-step by FernUniversität Hagen in cooperation with CeDiS. In the *Memories of the Occupation in Greece* project, however, over 90 video interviews with 91 witnesses of the German occupation of Greece during World War II were recorded, transcribed, translated and annotated between 2016 and 2018. Several other collections on different historical topics will use similar platforms in the coming years.

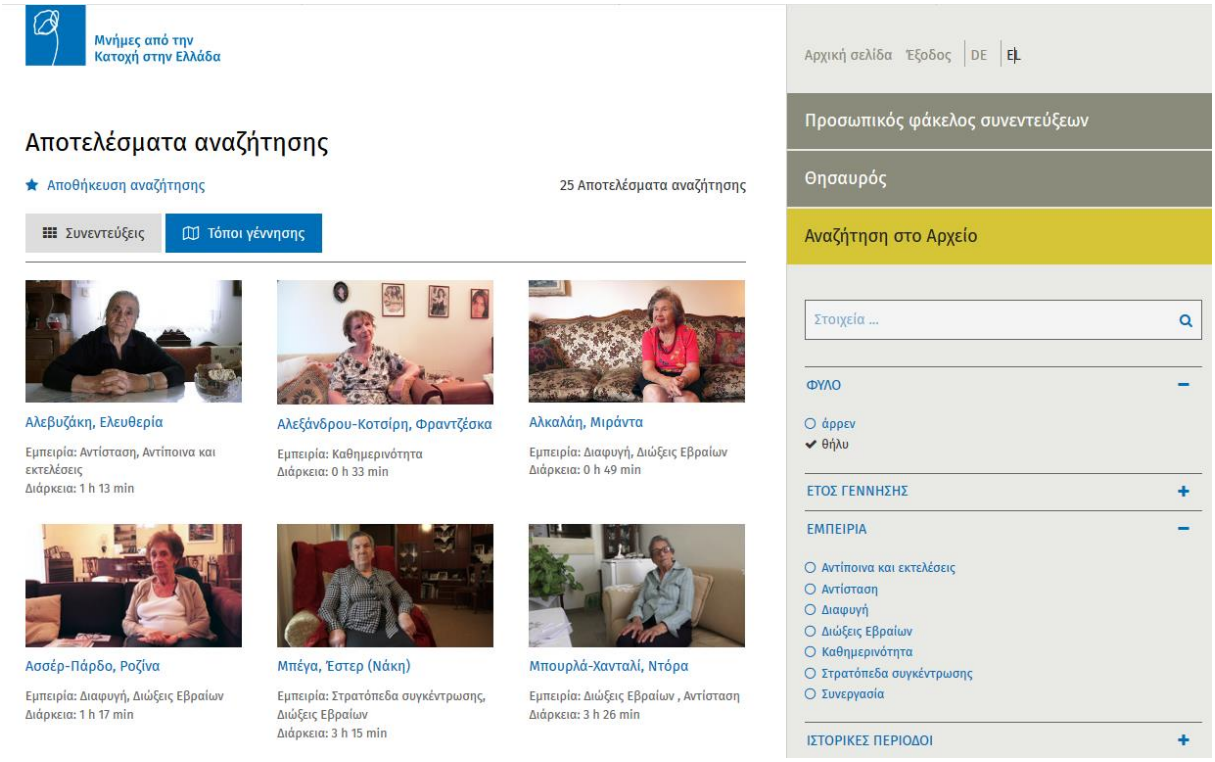


Figure 2: Searching for female narrators in the responsive Greek-language user interface of the *Memories of the Occupation in Greece* archive, www.occupation-memories.org, 30 Jan 2019

3 Digital Perspectives

The digital curation of interview collections faces a number of problems. Digital preservation strategies must deal with constantly changing technologies, standards and file formats in order to pursue an affordable sustainability. Online archives enhance the accessibility of testimonies but have to respect the narrators' privacy rights when dealing with sensitive biographical narrations. Every collection has different – and often not well-defined – ethic and legal restrictions. Increasing digital availability and growing data protection standards make these varieties a difficult issue which must be tackled systematically collection by collection.

Future goals will include the discussion and dissemination of interoperable metadata standards, long-term preservation strategies, comparable transcription and indexing guidelines, together with cor-

responding software tools to support labor-intensive curation processes. If digital interview archives can gradually achieve some of these goals, well known in the wider context of Digital Humanities, they hold a rich potential for new and interdisciplinary research approaches.

Indexation and full-text search make the long recordings accessible but require the huge effort of manual transcriptions. Even though automatic speech recognition technology has made considerable progress in recent years, the poor recording quality of many oral history testimonies limits the usability of automatically generated transcriptions – given also high standard expectations of the research community. Considerable progress is expected in this field, however, through a cooperation with CLARIN partners like WebMAUS/BAS Munich, Malach/LINDAT Prague or the *oralhistory.eu* group. CLARIN workshops in Oxford 2016 (CLARIN-PLUS, 2016) and Munich 2018 (CLARIN workshop, 2018) discussed standards, explored requests and tested tools. Important steps would be the creation of dirty transcripts (for search instead of display) and the forced alignment of existing transcriptions without time codes. An implementation of phonetical search through Czech and English oral history interviews, developed by researchers from Pilsen, is accessible at the Malach Center for Visual History in Prague (Stanislav et al., 2016). The Fraunhofer IAIS and the *Archiv Deutsches Gedächtnis* at FernUniversität Hagen are working on improvements of the IAIS audio mining system for transcribing and indexing oral history interviews (Gref et al., 2018). Dutch scholars are combining a sequence different tools for various curation steps into a "transcription chain" (Hessen et al., 2018).

The digital interview archives created by CeDiS have been aimed at historians, educators, and the general public, supporting the qualitative and hermeneutic study of individual testimonies. Therefore, no tools for corpus-linguistic, data-driven or other quantitative analyses had been integrated. Given the growing importance of Digital Humanities approaches, however, such tools can provide a future perspective for oral historians and their collections. Searching for keywords in context over large interview corpora could detect patterns of experience, memory and narration, and might be used for a wide array of research questions. Gender studies could ask: Are women narrating their life-story in a different way than men? Social gerontologists could look at how elderly people speak about childhood experiences. Different comparative studies on various aspects of history and memory, but also in the fields of language acquisition, can make use of the digital search facilities in these interview archives.

4 Comparative Studies

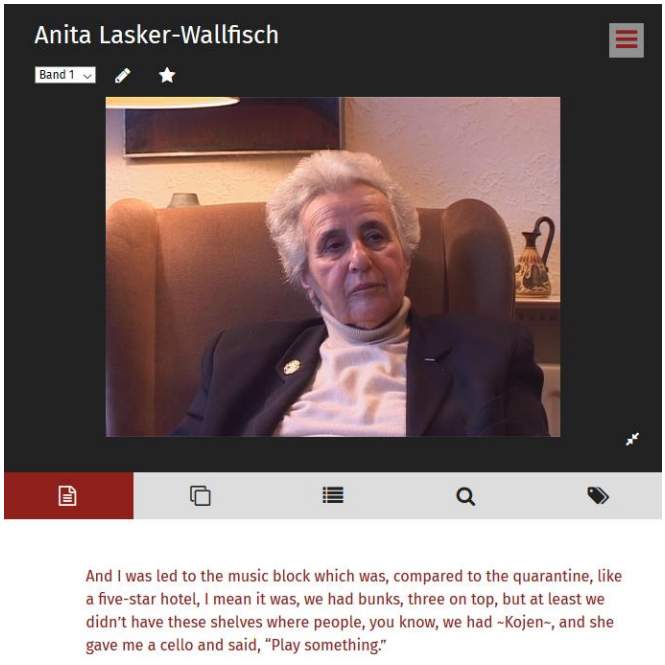


Figure 3: Interview with Anita Lasker-Wallfisch, www.zwangsarbeit-archiv.de, 30 Jan 2019

Some preliminary studies have proved that the interview archives can be very useful for comparative approaches – even without applying quantitative methods (Michaelis, 2013; Plato et al., 2010; Thonfeld, 2014). As a very limited example, I have compared two interviews with Holocaust survivor Anita Lasker-Wallfisch, who had been cellist in the women's orchestra at Auschwitz and later became co-founder of the English Chamber Orchestra. The first interview was recorded in 1998 as part of the Shoah Foundation's *Visual History Archive*, the second in 2006 as part of the *Forced Labor* archive. The analysis highlights differences in narrative performativity, in dialogical interaction, and in multilingualism (Pagenstecher, 2018).

The comparison of the two interviews shows a greatly increased narrative experience. Lasker-Wallfisch's performative effort became more elaborate and successful in her later narration when she directly quoted other people more often. In 1998, there are about

100 instances of direct speech, in 2006 about 320 instances. The transcript of the later interview, which is just over 50% longer, contains over 300% more quotation marks. In 1998, Lasker-Wallfisch described her introduction to the orchestra at Birkenau using indirect speech: "So, she asked me to play something." In 2006, however, she used a direct quotation: "And she gave me a cello and said: 'Play something'." This seems to be a general tendency in narrating: When studying re-tellings in other contexts, linguists found a move towards performativity, marked by an increase in direct speech. More experienced narrators give their testimony with more performative elements and an enhanced narrative authority.

Comparing the testimonies from an interactional point of view, the focus is on the dialogue with the interviewer. This dialogue is somewhat hidden, because the camera focuses on the narrator. Often, it is overlooked by historians who are more interested in fact-finding than in the co-construction of the testimony. Oral history beginners are even being told "An interview is not a dialogue" in an introductory text (Oral History Center). As a first step to studying the interaction, the interviewer's interventions within in the interviewee's main narration were counted. Both interviewers – Scottish BBC journalist Joanna Buchan in 1998, German Historian Christoph Thonfeld in 2006 – intervened roughly once per minute throughout the interview, which apparently is an average value (Michaelis, 2013, p. 288). But half of Thonfeld's interventions were just supporting incentives to continue ("hm", "yes"), whereas Buchan asked many where, when and how questions, sometimes interrupting Lasker-Wallfisch's narrative flow. These results point to different professional backgrounds of the individual interviewer (oral historian vs. journalist), but also to different methodical guidelines in the interviewing projects (*Forced Labor* archive vs. *Shoah Foundation*). Digital interview collections can support such a comparative analysis of transcripts on a larger scale, helping us to better understand the working alliance between narrator and interviewer which lies at the heart of each oral history interview.

Comparative studies on historical topics like the Holocaust or Nazi Forced Labor need to deal with different languages. In the *Forced Labor* archive, "Auschwitz" is mentioned in 188 interviews in 19 different languages, while the *Visual History Archive* contains almost 14,000 interviews in around 30 languages with this index term. Due to the deportation and forced migration experiences, many testimonies contain language-mixes or are almost bilingual documents. Words or sentences from another language are not yet searchable systematically, but at least are marked in italics in the *Forced Labor* archive. Comparing Anita Lasker-Wallfisch's testimonies in English, we can study an increased use of her German mother tongue in the later narration. In 1998, she used only a handful of German words, apparently taken over from the SS, such as "Zählappell". In all survivors' testimonies, the German perpetrators' camp language has entered the victims' memories narrated in another language. In 2006, however, her German mother-tongue keeps surfacing continuously for different topics from the pre-war and the post-war period. The main reason for this could be the German interviewer, but also her own cautious re-opening towards her country of birth and persecution over the years.

In general, a linguistic approach, supported by digital tools, can help the historian in listening more closely to the details of narrating, focusing on specific words rather than on general content. It might be interesting, for example, in which contexts Anita Lasker-Wallfisch – and other survivors – talk about themselves as individuals, using the "I", or as members of a group, saying "we". We might want to understand when narrators use active verbs, remembering or reclaiming their agency, or passive constructions, signaling powerlessness and victimhood.

In such future research projects, an increased cooperation between oral historians and corpus and interactional linguists can be very productive; the CLARIN network could be a helpful framework.

5 Cross-collection Search

Now, these comparisons between two single interviews are quite limited in their scope. For a more systematic approach, however, towards different narrator group subsets, the interview transcripts would have to be more standardized in a really machine-readable form. But the oral history community has been slow to accept transcription or annotation standards. Therefore, CeDiS is working on a TEI schema for oral history interviews, building on the TEI guidelines for transcribed speech (Text Encoding Initiative, 2018) and the ISO standard 24624:2016 prepared by CLARIN Center IdS Mannheim (Schmidt, 2017). Such a schema will include relevant metadata about the narrators' and the interview-

ers' age, gender, mother-tongue, or social group, but also different annotation layers including speaker changes, pauses, non-verbal utterances, words in other languages etc.

Existing digital archives allow comparative studies within a single collection. A cross-collection search is difficult, however, since different collections are not linked through a meta-catalogue. Especially in Germany, the interview collections, often run by under-funded non-governmental initiatives, have very different cataloguing systems and metadata schemas; many interviews have not even been digitized. But even for the digital archives developed at CeDiS, the application of long-term open linked data strategies proved to be difficult, because of very limited time frames, different thematic contexts or restrictive access conditions in the various projects. In the future, however, CeDiS will use more interoperable platforms and standards, and also assign a Digital Object Identifier (DOI) to each interview and make some basic, anonymized metadata harvestable. It is also planned to enhance the visibility of the collections in generic archival portals like Archivportal-D, language resource registries like the Virtual Language Observatory or cultural heritage catalogues like the Europeana.

The different domains of archives, language and heritage – not to mention film history or Holocaust research – are working with diverse metadata standards. Large library-based oral history collections in the US have created MARC21 records for their interviews; some catalogues like the European Holocaust Research Infrastructure (EHRI) have adapted the Encoded Archival Description (EAD) schema. Most of these standards are not very adequate for oral history interviews. The rather flexible Component Metadata Initiative (CMDI) framework with its Oral History profile might provide an interoperable solution, however (CLARIN-PLUS 2016).

In a separate project, the CeDiS team explores the chances of linking interview data by creating a cross-collection catalogue of audio- or video-recorded testimonies. This pilot is being developed within the HERA-funded project *Accessing Campscapes*, which studies the contested transformation of former Nazi and Stalinist camps into sites of remembrance with approaches from contemporary archaeology, oral history and memory research (www.campscapes.org). Various projects have interviewed survivors of these camps at different times; some narrators have given several testimonies. Such a cross-collection database can support comparative studies, point the researcher to prominent as well as forgotten survivor narratives, and help in researching the contested pasts of these places.

Creating such a catalogue, however, faces various challenges – like different curation strategies, heterogeneous metadata and restricted access to various collections. The pilot of the *Accessing Campscapes* project will only collect metadata of some selected institutions at a certain point in time. A central directory of oral history sources, which harvests the growing number of databases at individual institutions automatically, remains a future goal.

6 Reflections

To summarize, digital oral history collections can be a valuable source for interdisciplinary research, specifically in a cooperation between linguists and historians. The collections created or hosted at CeDiS of Freie Universität Berlin are already digitized and accessible. Their data need to become more machine-readable, however, to allow cross-collection searching and digital analysis.

New research perspectives can open up, when "oral history meets linguistics" – the title of a 2015 workshop in Freiburg (Pagenstecher and Pfänder, 2017). Cooperative projects with corpus linguists and conversation analysts can yield interesting results, since they can combine data-driven research with qualitative-hermeneutic approaches. The narrative patterns detected with a digitally supported analysis – or distant reading – will have to be interpreted through a careful listening to individual testimonies – or close reading.

While moving forward with technology and standards, some precaution and reflection will be necessary, however, when we treat recordings of personal life-stories as a corpus of audiovisual data. For an oral historian, perhaps the de-contextualization of the individual narration is the most worrying aspect, specifically when working with testimonies from Holocaust survivors. With reference to Walter Benjamin, Andree Michaelis (2013, p. 247) has written about the "testimony in the age of mechanical reproduction". So, what happens to the testimony and its aura – or respectful understanding – in the digital age?

In general, the digitized perception of historical sources – papers, videos or artifacts – usually implies a higher degree of abstraction on an intellectual and sensual level, because the material and embodied dimensions of the past are lost. When researchers watch survivors' recordings on the screen, instead of listening to them in person, they obviously miss a lot of context – what was said before the recording, how the apartment looked or smelled like etc. While interview protocols and set photos are available for many interviews, every secondary analysis will have to cope with a loss of contextual knowledge. Obviously, the meaning of "context" differs between disciplines: While linguists are used to working with data recorded by others, many qualitative social researchers would reject such an approach because the study-level metadata often is not giving enough contextual information.

While this larger distance seems to be inherent to digital research, digital environments for oral history allow working much closer to the audio-visual historical source. In the age of the tape recorder, most oral historians worked with a textual representation of the recording in the form of a transcript. Nowadays, digital technology helps us to study the audio-visual sources themselves, including the multiple modalities of text, speech, silence, gestures and facial expressions captured in the video images and the audio track. Given their text-oriented research tradition, historians now have to take up new approaches in analyzing these multimodal sources of memory. Any cooperation with linguists or other disciplines will be extremely helpful in that endeavor.

Acknowledgements

I am very grateful to Anita Lasker-Wallfisch and all other survivors for their engagement in retelling their experiences in the most inspiring way. I thank my colleagues at CeDiS, mainly Verena Nägel, Rico Simke, Doris Tausendfreund and Dorothee Wein, as well as the editors and reviewers for their support and feedback.

References

[Abenhausen et al.2012] Abenhausen Sigrid, Apostolopoulos, Nicolas, Körte-Braun, Bernd, Nägel, Verena (Eds.). 2012. *Zeugen der Shoah: Die didaktische und wissenschaftliche Arbeit mit Video-Interviews des USC Shoah Foundation Institute*, Berlin: Freie Universität

[Apostolopoulos et al.2016] Apostolopoulos, Nicolas, Barricelli, Michele, Koch, Gertrud (Eds.). 2016. *Preserving Survivors' Memories. Digital Testimony Collections about Nazi Persecution: History, Education and Media (Education with Testimonies, Vol. 3)*, Berlin: Stiftung EVZ, http://www.stiftung-evz.de/index.php?id=1655, 8 Sep 2018

[Apostolopoulos and Pagenstecher2013] Apostolopoulos, Nicolas, Pagenstecher, Cord (Eds.). 2013. *Erinnern an Zwangsarbeit: Zeitzeugen-Interviews in der digitalen Welt*, Berlin: Metropol

[Bothe and Brüning2015]. Bothe, Alina, Brüning, Christina Isabel (Eds.). 2015. *Geschlecht und Erinnern im digitalen Zeitalter. Neue Perspektiven auf ZeitzeugInnenarchive*, Berlin: Lit

[CLARIN-PLUS2016] *CLARIN-PLUS workshop "Exploring Spoken Word Data in Oral History Archives", Oxford*, 18./19.4.2016, https://www.clarin.eu/event/2016/clarin-plus-workshop-exploring-spoken-word-data-oral-history-archives, 8 Sep 2018

[CLARIN workshop2018] *CLARIN workshop: "Oral History: Users and their scholarly practices in a multidisciplinary world", Munich*, 19.-21.9.2018, http://oralhistory.eu/workshops/munich, 8 Oct 2018

[Gref et al.2018] Gref, Michael, Köhler, Joachim, Leh, Almut. 2018. Improved transcription and indexing of oral history interviews for digital humanities research. *International Conference on Language Resources and Evaluation (LREC) 2018*, http://publica.fraunhofer.de/dokumente/N-494202.html, 30 Jan 2019

[Hessen et al.]. Hessen, Arjan van et al. *Oral History & Technology: Transcription Chain*, http://oralhistory.eu/workshops/transcription-chain, 8 Sep 2018

[Michaelis2013]. Michaelis, Andree. 2013. *Erzählräume nach Auschwitz: Literarische und videographierte Zeugnisse von Überlebenden der Shoah*, Berlin: Akademie

[Nägel2016]. Nägel, Verena. 2016. Zeugnis – Artefakt – Digitalisat. Zur Bedeutung der Entstehungs- und Aufbereitungsprozesse von Oral History-Interviews. *Videographierte Zeugenschaft. Ein interdisziplinärer Dialog*. Ed. vy Eusterschulte, Anne, Knopp, Sonja, Schulze, Sebastian. Weilerswist: Velbrück Wissenschaft, 347-368

[Oral History Center] Oral History Center of Berkeley University, *Oral History Tips*, http://www.lib.berkeley.edu/libraries/bancroft-library/oral-history-center/oral-history-tips, 8 Sep 2018

[Pagenstecher and Pfänder2017]. Pagenstecher, Cord, Pfänder, Stefan. 2017. Hidden Dialogues: Towards an Interactional Understanding of Oral History Interviews. *Oral History Meets Linguistics*. Kasten, Erich, Roller, Katja, Wilbur, Joshua. Fürstenberg/Havel: SEC Publications, 185-207

[Pagenstecher2018] Pagenstecher, Cord. 2018. Testimonies in Digital Environments. Comparing and (De-)Contextualizing Interviews with Holocaust Survivor Anita Lasker-Wallfisch. *Oral History*, 46 (2), 109-118

[Pagenstecher and Tausendfreund2015] Pagenstecher, Cord, Tausendfreund, Doris. 2015. Interviews als Quellen der Geschlechtergeschichte. Das Online-Archiv "Zwangsarbeit 1939-1945" und das "Visual History Archive" der USC Shoah Foundation. *Geschlecht und Erinnerung im digitalen Zeitalter. Neue Perspektiven auf ZeitzeugInnenarchive*. Ed. by Bothe, Alina, Brüning, Christina Isabel. Berlin/Münster: Lit, 41-67

[Pagenstecher and Wein2017] Pagenstecher, Cord, Wein, Dorothee. 2017. Learning with Digital Testimonies in Germany. Educational Material on Nazi Forced Labor and the Holocaust. *Oral History and Education. Theories, Dilemmas, and Practices*. Ed. by Llewellyn, Kristina R., Ng-A-Fook, Nicholas. New York 2017, 361-378

[Plato et al.2010] Plato, Alexander von, Leh, Almut, Thonfeld, Christoph (Eds.). 2010. *Hitler's Slaves: Life Stories of Forced Labourers in Nazi-Occupied Europe*, New York/Oxford: Berghahn

[ProQuest2017] ProQuest. 2017. *USC Shoah Foundation Announces Partnership with ProQuest to Increase Access to Visual History Archive*, 30 June 2017, http://www.proquest.com/about/news/2016/USC-Shoah-Foundation-Partnership-with-ProQuest.html, 8 Sep 2018

[Schmidt et al.2017] Schmidt, Thomas, Hedeland, Hanna, Jettka, Daniel. 2017. Conversion and annotation web services for spoken language data in CLARIN. *Selected papers from the CLARIN Annual Conference 2016*. Linköping Electronic Conference Proceedings 136: 113-130

[Stanislav et al.2016] Stanislav, Petr, Svec, Jan, Ircing, Pavel. 2016. An Engine for Online Video Search in Large Archives of the Holocaust Testimonies. *Interspeech 2016: Show & Tell Contribution, September 8–12, 2016*. San Francisco, USA, https://www.isca-speech.org/archive/Interspeech_2016/pdfs/2016.PDF, 8 Sep 2018

[Thonfeld2014] Thonfeld, Christoph. 2014. *Rehabilitierte Erinnerungen? Individuelle Erfahrungsverarbeitungen und kollektive Repräsentationen von NS-Zwangsarbeit im internationalen Vergleich*, Essen: Klartext

[Text Encoding Initiative2018] Text Encoding Initiative, P5. 2018. *Guidelines for Electronic Text Encoding and Interchange*, Version 3.4.0. Last updated on 23rd July 2018, Chapter 8 Transcriptions of Speech, http://www.tei-c.org/release/doc/tei-p5-doc/en/html/TS.html, 8 Sep 2018

[USC Shoah Foundation2017] USC Shoah Foundation. 2017. *Nearly 1,000 English Transcripts Added to Visual History Archive*, 11 Sep 2017, https://sfi.usc.edu/news/2017/09/17961-nearly-1000-english-transcripts-added-visual-history-archive, 8 Sep 2018

Lexical Modeling for Natural Language Processing

Alexander Popov
Institute of Information and Communication Technologies
Bulgarian Academy of Sciences, Bulgaria
alex.popov@bultreebank.org

Abstract

This paper describes a multi-component research project on the computational lexicon, the results of which will be used and built upon in work within the CLARIN infrastructure to be developed by the Bulgarian national consortium. Princeton WordNet is used as the primary lexicographic resource for producing machine-oriented models of meaning. Its dictionary and semantic network are used to build knowledge graphs, which are then enriched with additional semantic and syntactic relations extracted from various other sources. Experimental results demonstrate that this enrichment leads to more accurate lexical analysis. The same graph models are used to create distributed semantic models (or "embeddings"), which perform very competitively on standard word similarity and relatedness tasks. The paper discusses how such vector models of the lexicon can be used as input features to neural network systems for word sense disambiguation. Several neural architectures are discussed, including two multi-task architectures, which are trained to reflect more accurately the polyvalent nature of lexical items. Thus, the paper provides a faceted view of the computational lexicon, in which separate aspects of it are modeled in different ways, relying on different theoretical and data sources, and are used to different purposes.

1 Introduction

In many theories, both within theoretical and computational linguistics, the lexicon plays the role of bridging language structures. Projects like VerbNet (Schuler, 2005), which extends Beth Levin's work on verb classes (1993), aim to bring together various levels of linguistic analysis – syntactic, semantic, predicate logic – and make them cohere via the lexical interface. Other projects like SemLink (Palmer, 2009) bind together even more heterogeneous data: VerbNet, FrameNet, PropBank, and recently the WordNet sense groupings used in the OntoNotes corpus (Bonial et. al., 2013). The Predicate Matrix project in turn extends SemLink to obtain an even wider coverage (De Lacalle et. al., 2014). Such a model of the lexicon moves toward a more realistic representation that reflects the interdependence of linguistic structures.

Combining heterogeneous data in computational lexicons is a challenge that has been taken up with renewed vigour in recent years, due to the increased popularity of *vector space models* (VSMs), also known in the literature as *embeddings*. New, more efficient approaches to embedding (Mikolov et. al., 2013b; Pennington et. al., 2014) allow for the learning of latent information from huge volumes of text and encoding that in real-valued vectors of varying dimensionality. In this way semantico-syntactic features can be extracted and compressed in powerful and compact models that are more easily interpretable by machine learning (ML) systems. This excitement around embedding methods has led to growing research in *sense embedding*, i.e. methods to derive embeddings for other linguistic items, namely word senses. Since training data annotated with word senses is much more difficult to obtain, various approaches have been tried, with most of them hingeing on somehow translating the structure of lexical resources (such as the knowlege graph of WordNet) into instructions for training (Faruqui et. al., 2014; Johansson and Pina, 2015; Rothe and Schütze, 2015; Goikoetxea et. al., 2015).

Alexander Popov 2019. Lexical Modeling for Natural Language Processing. *Selected papers from the CLARIN Annual Conference 2018*. Linköping Electronic Conference Proceedings 159: 152–165.

The present paper describes the first phase of a hybrid approach for the construction of such multi-aspectual representations of the lexicon that are suitable for use by automatic systems. The representations take different forms: knowledge graphs, vector space models, neural networks trained on lexical analysis tasks. This line of research assumes WordNet (Fellbaum, 2010) as its main knowledge resource and tests a number of hypotheses about how this computational lexicon can be enriched and adapted to various natural language processing (NLP) tasks, such as word sense disambiguation (WSD), word similarity/relatedness, context representation, part-of-speech (POS) tagging. The enriched resources and ML models will be integrated in and developed further within the CLaDA-BG project[1], in order to reflect a rich multi-level view of the lexicon. Though this paper focuses on English data, Bulgarian models will be produced as well.

This paper presents current results and perspectives. More specifically, it describes:

- the effects of knowledge graph enrichment on knowledge-based WSD accuracy;
- graph embedding on enriched knowledge graphs;
- using graph embedding models for supervised WSD;
- benefits from multi-task supervised lexical learning with recurrent neural networks.

2 Related Work

Knowledge-based WSD Navigli (2009) provides a good overview of knowledge-based WSD (KB-WSD). The current work deals exclusively with KBWSD via graphs, building on Agirre and Soroa (2009) and using the UKB software[2] for PageRank-based WSD. The latter work uses WordNet's original semantic network, relations between the manual annotations of the glosses, and the automatic annotations of glosses from eXtended WordNet (Harabagiu and Moldovan, 2000) to construct its knowledge graph (KG). Agirre et. al. (2018) describe new settings for that software that make the approach state-of-the-art within KBWSD and very competitive even against supervised systems.

Lexical learning with RNNs Various aspects of lexical learning have been tackled successfully with deep neural networks. Recurrent neural networks (RNNs) in particular are a powerful tool for handling sequences of language tokens. With respect to WSD, Kågebäck and Salomonsson (2016) have obtained state-of-the-art results on the *lexical sample task*, using bidirectional Long Short-term Memory (Bi-LSTM) networks, and Raganato et. al. (2017b) have done so on the *all-words task*. With respect to POS tagging, various papers have demonstrated the viability or RNNs (Wang et. al., 2015a; Wang et. al., 2015b; Huang et. al., 2015). Context embedding has become a popular method in NLP, especially within the deep learning paradigm. Kiros et. al. (2015) describe a model for learning a generic sentence encoder that provides distributed representations which can be used in different settings. Melamud et. al. (2016) present another neural architecture for context representation – called *context2vec* – which can be used for WSD.

Multi-task learning Raganato et. al. (2017b) combine WSD, POS tagging and coarse-grained semantic labeling (CGSL) in an RNN setup. The empirical comparison shows that with a number of data sets training on the coarse-grained task does help in improving the accuracy of WSD. The addition of POS tagging to the combination WSD+CGSL is not universally beneficial – it increases the accuracy of WSD on several data sets, but brings it down on others. Plank et. al. (2016) present a Bi-LSTM tagger tested on data for 22 languages; the addition of a frequency classification auxiliary task to the RNN is shown to improve results on many of the languages. Alonso and Plank (2016) present a study of different combinations of NLP tasks – primary (frame detection and classification, supersense tagging, NER, etc.) and auxiliary (chunking, dependency label classification, POS tagging, frequency estimation). Nguyen et. al. (2017) present a state-of-the-art joint Bi-LSTM model for POS tagging and graph-based dependency parsing. Ruder (2017) is a good comprehensive overview of various multi-task learning setups, while

[1]National interdisciplinary Research e-Infrastructure for Resources and Technologies in favor of the Bulgarian Language and Cultural Heritage, as part of the EU infrastructures CLARIN and DARIAH. Grant number D01-164/28.08.2018 by Bulgarian Ministry of Education and Science.

[2]`http://ixa2.si.ehu.es/ukb/`

Bingel and Søgaard (2017) evaluate all possible combinations of two NLP tasks out of a pool of ten tasks.

Vector space models for word senses (synsets) One approach for training sense embeddings is through *retrofitting methods*, i.e. adapting an already existing resource to reflect the structure of a knowledge base, typically WordNet (Faruqui et. al., 2014; Johansson and Pina, 2015; Rothe and Schütze, 2015). Camacho-Collados et. al. (2015) put forward NASARI – a system for automatically constructing vector representations for WordNet synsets and Wikipedia pages. Another way to learn distributed representations of senses is to automatically sense-annotate huge amounts of natural text and then train a neural network language model on the processed data. Iacobacci et. al. (2015) use the BabelNet sense inventory and the Babelfy knowledge-based WSD tool to tag a dump of Wikipedia on which to train the VSMs. Mancini et. al. (2016) propose an interesting solution called *SW2V*; they use a *shallow word-sense connectivity algorithm* in order to annotate the open class words of a large corpus with potential word senses, then train a modified CBOW architecture (Mikolov et. al., 2013b) on word and sense prediction simultaneously. A somewhat different way of creating distributed representations of concepts is via generating artificial sequences from knowledge graphs, on which neural network language models can be trained (Goikoetxea et. al., 2015). This is the approach adopted in this work. It allows to easily modify the training data and to produce hybrid VSMs (e.g. combinations of words and word senses). The article provides empirical evidence that it is very competitive compared to the other methods introduced above.

3 Graph Enrichment for Knowledge-based Word Sense Disambiguation

The experiments reported in this section present a number of strategies for obtaining a fuller lexical representation of words and senses, with WordNet (WN) as a starting point. The main idea is to add new relations to the original WN semantic network and to evaluate the effect of this modification on a KBWSD algorithm. The section first describes several strategies for adding new relations and then provides results on English sense-tagged data.

Preliminary experiments on Bulgarian data have shown that using a syntactically and semantically annotated corpus in conjunction with the hypernym-hyponym hierarchy from WN to generate new syntagmatic relations between concepts can significantly improve KBWSD accuracy (Simov et. al., 2015). Additional strategies have been explored for the extraction of new relations from sense-annotated data for English. The various approaches, including the baseline relations from the original WN and eXtended WordNet (XWN) networks, are summarized and abbreviated below:

- The original WN semantic network, comprising of lexical semantic relations, such as hypernymy, antonymy, meronymy, etc. – **WN**
- The additional relations between open class words extracted from the manually annotated WN glosses – **WNG**
- Logic form representations and parse trees from eXtended WordNet (XWN). XWN contains first-order logic representations of the glosses, from which ”event” relations can be extracted – **WNGL**
- Context representation of XWN annotations. Word order is used to connect word sense annotations in the glosses. Each sense annotation is indirectly connected to the one preceding it via an artificial node that indicates contextual dependency. The artificial nodes too are added to the graph – **WN30glCon**
- Syntactic relations from the SemCor corpus (Miller et. al., 1993). The sense-annotated corpus is analyzed with a dependency parser. The sense annotations are attached to the nodes in the dependency trees, so that the full syntactic information about context usage is inserted in the graph (including functional words). Sentences are also connected to the ones preceding them via an artificial node that connects their root nodes – **GraphRelSC**

For a more detailed description of the relation sets, see Simov et. al. (2016).

Table 1 shows experimental results with a selection of the best KG combinations. Two data sets are used for evaluation. The first one is WSDEval – a concatenation of the Senseval-2, Seneval-3, SemEval-07, SemEval-13, SemEval-15 data sets, as described in Raganato et. al. (2017a)[3]. The second one is a subset of SemCor (49 files in total) which was not used for the extraction of new relations. The annotation was performed with the UKB system, using the settings outlined in Agirre et. al. (2018). The addition of new relations improves accuracy scores in all cases of evaluation on the SemCor data, while it actually hurts scores on the WSDEval data set in two of the three setups with enriched graphs. However, the combination of the baseline relations and the syntactic enrichment extracted from SemCor provides a big boost to the KBWSD algorithm[4]. The system is thus able not just to achieve the best reported results for KBWSD (at least in Raganato et. al. (2017a)), but to come close even to the top supervised WSD models.

KG	*WSDEval*	*SemCor*
WN + WNG	67.30%	74.2%
WN + WNG + WNGL	66.9%	74.9%
WN + WNG + WN30glCon	67.1%	**75.1%**
WN + WNG + GraphRelSC	**68.2%**	**75.1%**

Table 1: KBWSD accuracy on WSDEval and SemCor.

4 Graph Embedding

Building on the observation that enriching the KG does indeed help with lexical analysis tasks such as KBWSD, the line of research presented in this section attempts to translate this type of structured information (KG) into a numerical representation (vector space model) and evaluate the latter on standardly used tasks. To do that, the method in Goikoetxea et. al. (2015) is used. It is outlined below, in combination with the preliminary step of graph enrichment:

1. The knowledge graph is assembled from various sets of relations.
2. The UKB tool is used in its "walkandprint" regime in order to produce random walks of variable lengths along the structure of the KG. The tool can be configured to output, with certain probability, synset IDs or a lemma from the corresponding synset. When training word or lemma embeddings, the probability for outputting a lemma is set to 1. Each random walk constitutes one training sentence.
3. The Word2Vec tool[5] is used to train a VSM on the artificially produced corpus, in particular the Skip-Gram architecture. The following parametrization is employed: 7 iterations over the data; number of negative examples set to 5; frequency cut sampling set to 7; context windows of size 5 or 15.
4. The resulting VSM is evaluated on the task of calculating relatedness and similarity scores between pairs of words. The Word-353 Similarity, WordSim-353 Relatedness (Agirre et. al. (2009) describe the two subsets of the WordSim-353 data set) and SimLex-999 (Hill et. al., 2015) data sets are used for the evaluation. The VSM is used to get the embeddings for the word pairs, then those are used to calculate their cosine distance. The distance metrics for all pairs of words are then used to calculate Spearman's rank correlation with respect to the gold corpus numbers provided by human annotators.

Table 2 shows the results with a selection of the best performing VSMs on the three evaluation data sets. The first three results are obtained with the baseline models: the **GoogleNews** model trained on a large corpora of natural language text; vectors trained on contexts generated from dependency parses

[3]This is part of the Unified Evaluation Framework (UEF) resource, assessible at `http://lcl.uniroma1.it/wsdeval/`

[4]When used in combination, WNGL, WN30glCon, GraphRelSC do not yield results that are better than the baseline. This indicates that with such parametrization of the algorithm two of the three new relation sets actually introduce noise rather than useful knowledge.

[5]`https://code.google.com/archive/p/word2vec/`

of natural language text (**Dependency**); the WordNet-generated pseudo-corpus (**WN + WNG**), which is produced from the original semantic network and the gloss annotations. The first two new models (**WN + WNG + HypInf** *C5/C15*) add the transitive closure over the hypernym-hyponym relations in WN. This means that the direct taxonomic relations are made explicit between all elements in a hypernymy chain: e.g. if "surgeon" is marked as a hyponym of "doctor" and "doctor" is marked as a hyponym of "medical professional", an explicit link is added between "surgeon" and "medical professional"; this is done for all implicit taxonomic relations of this nature. The table shows that this addition significantly improves the performance on the *SimLex999* data set, even though it reduces the performance on the rest of the data. The second group of models based on the extended graphs (**WN + WNglConOne** *C5/C15*) combines the original WN semantic network with the **WNglConOne** set of relations introduced in the previous section. Already these combinations achieve scores that are better than or at worst on par with all three baselines (with the VSM trained on a context window of 15 words being significantly better on two out of three data sets). The third group of models (**WN + WNG + WNGL + GrRelSC** + *C5/C15*) is based on data generated via the combination of WN, the gloss annotations, the relations extracted from the logic forms in XWN and of the syntactic relations from SemCor. These models achieve the best results on the *WordSim353* similarity data set.

Vector Space Model	*WordSim353 Similarity*	*WordSim353 Relatedness*	*SimLex999*
GoogleNews(Mikolov et. al., 2013a)	0.77145	0.61988	0.44196
Dependency(Levy et. al., 2014)	0.76699	0.46764	0.44730
WN + WNG(Goikoetxea et. al., 2015)	0.78670	0.61316	0.52479
WN + WNG + HypInf *C5*	0.77730	0.54419	0.55192
WN + WNG + HypInf *C15*	0.77205	0.55955	**0.55868**
WN + WNglConOne *C5*	0.77761	0.64747	0.53242
WN + WNglConOne *C15*	0.79659	**0.65548**	0.52632
WN + WNG + WNGL + GrRelSC *C5*	0.79847	0.63587	0.51974
WN + WNG + WNGL + GrRelSC *C15*	**0.81862**	0.61455	0.52350
WN + WNglConOne *C15* + **GoogleNews**	**0.82684**	**0.70972**	0.54675
WN + WNglConOne *C15* + **Dependency**	0.80428	0.66570	0.54041

Table 2: Comparing results from different VSMs on the similarity and relatedness tasks. *C5* and *C15* are used to indicate the size of the context window for the Skip-Gram model. The best results on the different data sets, using a single VSM as a source, are marked in bold. The final two experiments give the correlation scores for combinations of VSMs: a graph-based one and the GoogleNews/Dependency vectors; the first combination achieves the best overall results on two of the data sets and comes close to the best result on the third one.

It is evident that different types of new relations contribute to performance differently on the different data sets. For instance, hypernymy-derived relations lead to improvements on the similarity rather than on the relatedness task, possibly because paradigmatic relations in general are more informative with regards to semantic similarity. The experiments also show that combining different kinds of relation sets is beneficial, i.e. there is a significant degree of complementarity between them. This result provides justification for integrating the various sources of information in a common lexicon model. The last two experiments take one of the best models (**WN + WNglConOne** *C15*) and concatenate its vector representations with the vectors from the first two baselines, in order to test how well models trained on real text and on graphs complement each other. The results are very encouraging, especially the combination with the **GoogleNews** model, which achieves by far the highest scores on the *WordSim*

data sets and the closest results on *SimLex999* to those achieved via the hypernymy-inferred relations. While certainly there is a degree of redundancy in the vectors obtained via concatenation, the correlation improvements signal that there is also complementarity.

5 Neural Lexical Learning Using Graph Embeddings

This section presents two deep learning architectures for performing lexical analysis. The first one is an instance of now-standard recurrent neural network classifiers for sequence-to-sequence tasks, such as WSD and POS tagging (here called *Architecture A*), while the second does not optimize on a classification objective directly. Rather it tries to learn to represent context information in the same embedding space used to transform the input and gold data into numerical representations (here called *Architecture B*). The two architectures are described in greater detail below. They are both trained on the SemCor corpus and evaluated on the WSDEval data. Possible combinations of the two architectures and different tasks performed with them are discussed in connection with multi-task learning.

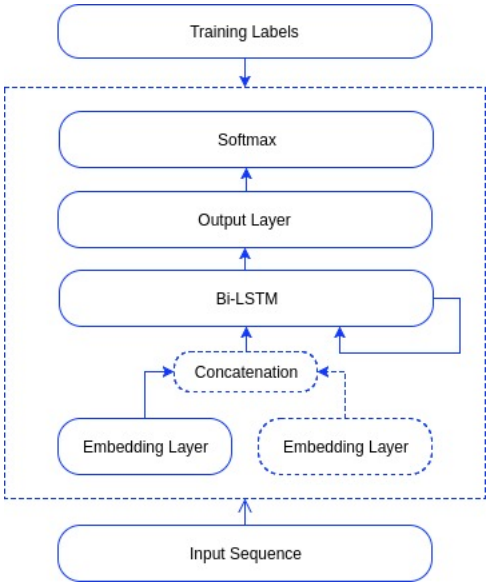


Figure 1: *Architecture A* – a recurrent neural network for sequence-to-sequence tagging. The dotted lines mean that a component or a connection is optional (in the case of concatenating embeddings from two different sources). Comparison between the training labels and the output is done via cross entropy.

Classification *Architecture A* is a Bi-LSTM network with an embedding layer at the input step and a softmax layer at the output (see fig. 1). When used for WSD, *Architecture A* maps the representation of a word context from its hidden layer(s) to a vector representing the synsets of all lemmas seen during training (WordNet 3.0 is used as dictionary). Words are disambiguated after consulting the dictionary to determine the possible synsets on the basis of lemma and POS information. Those not seen in training are disambiguated using WordNet's fist sense heuristic. Table 3 summarizes experimental work done with the architecture. Using the GloVe word embeddings (Pennington et. al., 2014), models trained with *Architecture A* achieve results within 1-2% accuracy from the state of the art, well above the most frequent sense baseline (typically difficult to beat in WSD). Experiments with additional VSMs whose embeddings are concatenated to the GloVe vectors do yield better results on separate data sets within WSDEval (Senseval-2, SemEval-13, SemEval-15), but they do not improve the purely GloVe-based result for the concatenation of all data sets, so they are not reported here (for more details and results see Popov (2017)). The model reported in the table has one hidden Bi-LSTM layer; the LSTM layers have sublayers of 200 neurons, each initialized from a random uniform distribution ([-1;1]); dropout of 20% is applied to the LSTMs; the learning rate is set to 0.2; training is done via stochastic gradient descent.

The same architecture can be parametrized for the POS tagging task – the only adaptation that needs to

be done is changing the output and softmax layers, so that the network learns to classify according to the smaller set of possible tags. A multi-task learning setting that combines WSD and POS tagging requires merely two separate output layers, so that different probability distributions per tag set can be obtained. The cross-entropy losses from the two training signals are then summed and passed to the optimizer.

System	SNE-2	SNE-3	SME-07	SME-13	SME-15	ALL
IMS-s+emb	**72.2**	**70.4**	**62.6**	65.9	71.5	**69.6**
Context2Vec	71.8	69.1	61.3	65.6	**71.9**	69.0
Architecture A	69.6	69.4	59.3	65.0	69.4	67.8
UKB-g*	68.8	66.1	53.0	**68.8**	70.3	67.3
IMS-2010	68.2	67.6	59.1	-	-	-
MFS	65.6	66.0	54.5	63.8	67.1	64.8
IMS-2016	63.4	68.2	57.8	-	-	-
Architecture B	64.7	57.9	47.9	61.9	64.8	61.3
UKB-g	60.6	54.1	42.0	59.0	61.2	57.5

Table 3: Comparison of the models trained with *Architectures A & B* with other systems trained on SemCor and evaluated on several data sets ("SNE" stands for "Senseval", "SME" stands for "SemEval"). *IMS-s+emb*, *Context2Vec*, *UKB-g*. UKB-g* and *MFS* are reported in Raganato et. al. (2017a); IMS-2010 is reported in Zhong and Ng (2010); IMS-2016 (this is the configuration IMS + Word2Vec) is reported in Iacobacci et. al. (2016). The results from the UEF stand for the F-1 score, but since all systems there either use a back-off strategy or are knowledge-based, this is equivalent to accuracy, just as in the present work.

Regression *Architecture B* is similar, but instead of using a softmax classifier, it compares the context representation (per word) to an embedding of the relevant gold label synset (see fig. 2). To that purpose, the hidden layer representation of the RNN is resized to the dimensionality of the VSM used to embed the inputs and to supply the synset embeddings. Least squares is used to compare the context vector and synset embeddings, the result of which serves for training. *Architecture A* tries to pick just one position in the large lexicon vector at its end and to depress probability mass in all other dimensions. This means that it is exploring a single source of information from the gold data – "this is the correct answer and everything else is wrong". *Architecture B* meanwhile aims at learning from a richer representation – by having access to the embedding of the gold synset in a VSM, at least hypothetically it should have access to a range of semantic features that describe the correct answer. Therefore it is a much more detailed model of the lexicon, which provides an interpretation of the meaning of word senses. Combining *Architectures A* and *B* is done analogously to the previous setup, only in this case one of the computational pathways ends with a softmax layer, while the other – with a least squares comparison. *Architecture B* can also be used for WSD – via calculating the cosine distance between the context embeddings and potential synsets. It achieves results below the most frequent sense (MFS) baseline, but above most of the popular knowledge-based solutions (see table 3 and Raganato et. al. (2017a) for more KBWSD results).

As outlined above, the models trained by *Architecture B* are essentially doing the following:

1. A sequence of vectors, one per word/lemma in the input sequence, are fed into the hidden layer.
2. The Bi-LSTM layers produce a context representation per each word/lemma.
3. The context representations that correspond to open class words are selected.
4. Each context representation is resized to match the dimensionality of the input vectors.
5. The resized context representation is compared with a distributed representation of the corresponding gold label synset.
6. The network is optimized on the mismatch between the two vectors.

In order to obtain a VSM that combines words/lemmas and synsets in a shared space, the methodology from Goikoetxea et. al. (2015) is employed once again. The following parameters are used with the

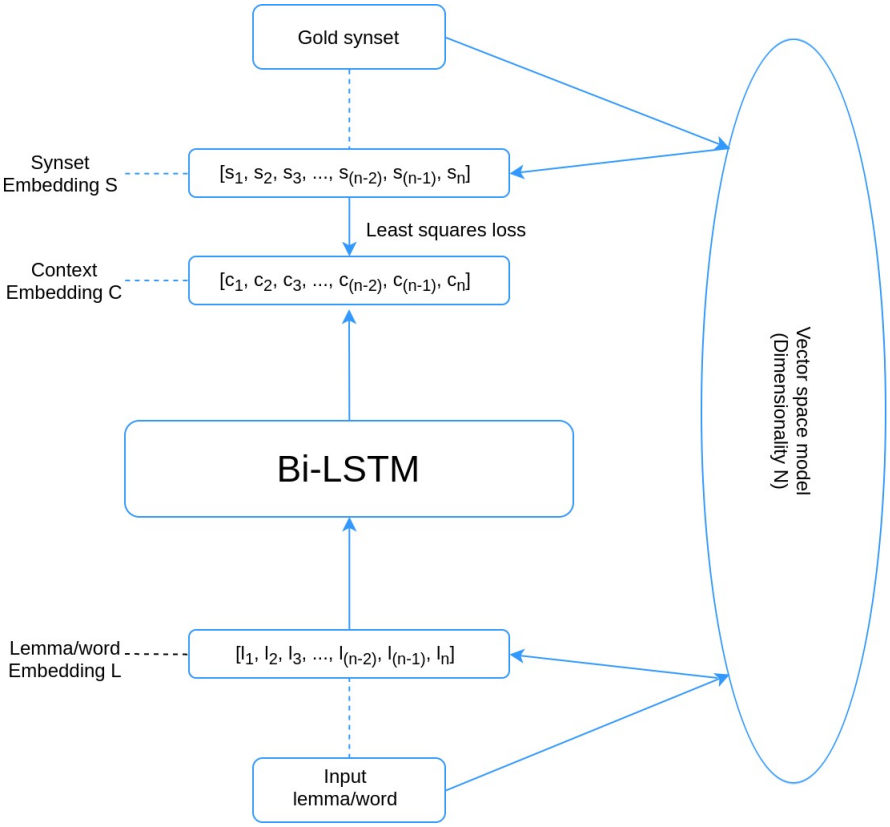


Figure 2: Diagrammatic representation of *Architecture B*. The same principles apply as with *Architecture A*, but the output layer produces a vector of the size of the VSM, which is then compared to the embedding vector for the gold synset; a mean of the least squares error is backpropagated as a learning signal. Crucial to the architecture is the availability of a VSM where both lemmas/words and synsets are represented as vectors of the same dimensionality.

Word2Vec tool: number of training iterations = 7; number of negative samples = 5; window size = 15; threshold for the occurrence of words = 1e-7. The graph used for the generation of the pseudo-corpus – **WN30WN30glConOne** – is based on the original WordNet relations and is also further enriched with syntagmatic relations extracted from the gloss annotations in eXtended WordNet (see section 3 above). This experimental setup also concatenates a Wikipedia dump to the pseudo-corpus training data, so that functional words can be represented as well.

Next I present some results from experiments with different VSMs as sources of input features to *Architecture B*. The evaluation is done only with respect to the development set, Senseval-2, and the goal is to demonstrate how different approaches to distributed representation lead to different quality of the embedding. This is particularly important in the case of *Architecture B*, as the network effectively learns how to navigate within the same VSM – it takes as input word/lemma embeddings, calculates via recurrences the contexts for the input tokens and tries to match those to the corresponding synset representations *within* the same space. This means that the more meaningfully words/lemmas and synsets are related spatially (and therefore semantically), the easier it is for the network to establish how to map actual input to expected output. The following parametrization of the network is used: 1 hidden Bi-LSTM layer; 400 units per sub-layer in the LSTMs; no dropout; optimization with stochastic gradient descent; learning rate of 0.2; random uniform initialization within [-1;1].

Table 4 shows that simply concatenating the lemmatized Wikipedia dump to the pseudo corpus leads to a big improvement – probably due both to the ability of the model to represent words missing in WN (mostly functional ones) and to having access to syntagmatic knowledge from natural language text. There is also difference in the performance of the models depending on how the pseudo-corpus was constructed. Corpus1 (C1) was built by generating 100 million random walks from the graph and

Vector Space Model	Accuracy (SNE-2)
WN30WN30glConOne-C3 + WikiLemmatized	**64.7**
WN30WN30glConOne-C3	63.1
SW2V (600 hidden units)	62.1
WN30WN30glConOne-C1	61.7
SW2V (400 hidden units)	60.2
WN30WN30glConOne-C2	57.4
AutoExtend	53.2

Table 4: Comparison of the models trained with *Architecture B* on the Senseval-2 data and using different VSMs. All models share the same parameters, except for one of the SW2V models, which has more hidden layer neurons. The SW2V embeddings are associated with mixed case strings of word forms as described in Mancini et. al. (2016); the AutoExtend vectors are described in Rothe and Schütze (2015).

then adding next to each synset ID in the random walks a randomly chosen lemma from that synset; Corpus2 (C2) was built by generating 200 million random walks and directly substituting synset IDs with representative lemmas. Thus, C1 and C2 are roughly of the same size, but C1 is much more effective in this evaluation. Corpus3 (C3) was built in the same way as C1, but the number of random walks in it is 200 million, i.e. twice bigger; it is the best-performing model based on pseudo-sentences only. VSMs that also represent words/lemmas and synsets in a shared space, but are constructed differently, like SW2V and AutoExtend, do not seem to fare better than the approach proposed here. The SW2V embeddings, which are directly trained on natural language text (non-lemmatized at that), do perform a little better than some of the pseudo-corpus-only-based embeddings (C1 and C2, bot not C3), if the hidden layer of the RNN is enlarged (which is not the case with the pseudo-text-based vectors). However, the combination of WN and Wikipedia beats all other VSMs, indicating that the KG is crucial for representing the relation between lemmas and synsets.

6 Multi-task Lexical Learning

This section presents results on multi-task learning, where the experimental work is based on *Architectures A* and *B*. The first subsection discusses the combined training of a WSD classifier and a POS tagger, and the second one – of a WSD classifier and a context embedding model.

6.1 WSD and POS tagging within *Architecture A*

SemCor is used as training data, but this time the files with the original POS annotations are used[6], as opposed to the data from the Universal evaluation framework (UEF) by Raganato et. al. (2017a), used in the rest of the experiments. For evaluation, the Senseval-2 all-words data from the UEF set is used; the POS tagger run over it has been manually corrected for the open class words and therefore it is used here as a kind of silver resource. The final WSD still uses the gold POS tags at the post-processing step outside of the network, but the network is influenced by the training of the POS tagging branch. The GloVe vectors are used as input features. The multi-task solution achieves 0.5% higher accuracy on WSD and 1.2% higher accuracy on POS tagging, compared to the analogous single-task models (see table 5; note that the result on Senseval-2 is somewhat higher in this setup, but that is most probably due to some difference in the data set representations used for training).

6.2 WSD and context embedding. Combining *Architectures A and B*

SemCor is used for training and the all-words task data from Senseval and SemEval – for evaluation (all data is downloaded from the UEF). The mixed lemma&synset best-performing VSM from the previous section is used for the input and the gold synset representations. Table 6 summarizes the results. Both the classifier-based disambiguation and the context embedding modules make gains under multi-task learning. This is especially evident with respect to the context embedding system, which overcomes the

[6]Downloaded from `https://github.com/rubenIzquierdo/wsd_corpora/tree/master/semcor3.0`

System	WSD (SNE-2)	POS (SNE-2)
Architecture A (single-WSD)	70.6	-
Architecture A (single-POS)	-	90.9
Architecture A (multi-task)	**71.1**	**92.1**

Table 5: Comparison of single-task models that learn to solve only either WSD or POS tagging, and a multi-task model that learns to solve both in parallel. "SNE-2" stands for "Senseval-2".

MFS baseline on some data sets. When the multi-task-trained classifier is regularized using dropout, it comes very near to the best result obtained with *Architecture A*. That best model uses the GloVe vectors, i.e. it has access to a VSM with representations of the input words obtained from a much larger data set. It is noteworthy that the GloVe vectors represent word forms, whereas the VSM used by the default models here encodes representations of lemmas only, i.e. it doesn't make any use of morphological information.

System	SNE-2	SNE-3	SME-07	SME-13	SME-15	ALL
Model A (single)	68.5	67.1	58.2	63.6	67.0	66.2
Model A (single, GloVe)	69.6	69.4	59.3	65.0	69.4	67.8
Model A (multi)	68.9	67.8	58.0	63.7	68.4	66.7
Model A (multi+dropout)	69.6	68.0	59.1	64.5	70.2	67.5
Model B (single)	64.7	57.9	47.9	61.9	64.8	61.3
Model B (multi)	66.8	60.1	49.2	63.4	67.7	63.3
MFS	65.6	66.0	54.5	63.8	67.1	64.8

Table 6: *Models A* perform *Architecure A*-style WSD, while *Models B* use the method described for *Architecture B*. The label "multi" in parenthesis indicates that this is one of the two computational pathways of the multi-task model. All models are trained using the same parameters. The amount of dropout applied to the last of the *A-models* is 0.5. "SNE" stands for "Senseval"; "SME" – for "SemEval".

These results are encouraging because they suggest that: 1) there is a significant amount of mutual support between the two tasks; 2) the poverty of the graph-induced vectors (compared to the GloVe vectors) can be somewhat mitigated in such multi-task learning settings.

Analysis of the Results Here I offer a preliminary analysis of the behavior of the two subsystems in the multi-task learning model, in order to demonstrate that the A and B branches learn different types of information and it is not the case that they give the same answers, with one of them being just a little more accurate. To this purpose, three subsets of the gold annotations have been excerpted from the *ALL* evaluation data set, together with the corresponding answers given by: the classification module (here called *A*), the context embedding module (here called *B*) and the WordNet 1st sense heuristic (here called C). The excerpted annotations all correspond to three types of situations. For obvious reasons, we are not interested in the cases where A and B provide the same answer, so this leaves the following: 1) A, B and C all give different answers; 2) B and C give the same answer, which is different from that provided by A; 3) A and C give the same answer, which is different from that provided by B. Table 7 provides an overview of how often one or the other model is correct.

If it were the case that the type B modules (context embedding) are merely learning the same information as type A modules (classification), one would expect to find almost no examples where the context embedding module, and not the classification one, provides a correct answer. This would be especially true when its answers deviate from the WN 1st sense heuristic, which in a way corresponds to the MFS heuristic and is something that can be learned from the training data fairly well. On the contrary, the context embedding architecture knows better than the classifier in many cases. In fact, it is more often correct in its predictions, by a wide margin. Note that this does not contradict the higher accuracy score of the classifier approach in general, as the latter uses a backoff heuristic (WN 1st sense) whenever it en-

Combination	A $\neq C \neq B$	B = C $\neq A$	A = C $\neq B$	Total
A correct	46	256	452	754
B correct	79	598	257	934
C correct	78	598	452	1128
Neither correct	82	229	241	552
Both (A&B) correct	3	12	15	30

Table 7: Comparison of different models. The first column gives information about cases where neither of the three models agrees with any of the rest; in the second column the context embedding module picks the same answer as the WN 1st sense heuristic; and in the third one the classification module conforms to the WN 1st sense heuristic. ”A” stands for ”classification module”; ”B” – for ”context embedding module”; ”C” – for ”WN 1st sense”. The ”Both correct” line means that the two modules (A and B) chose different synsets which are both listed in the gold annotation.

counters a word it has not trained on. But if such cases are counted as errors on the part of the A models, then the context embedding module is clearly more powerful than the softmax-based part of the architecture. Model B is also leading the board in the cases where all three models give different answers (albeit it is in practice tied with model C). And when the classifier and the 1st sense heuristic are in agreement, the context embedding module is correct in about a quarter of all cases. This short analysis is of course far from sufficient for any final conclusions, but it nevertheless strongly suggests that the two pathways in the multi-task learning architecture indeed pick on different types of data. Therefore figuring out how to integrate them even better and how to build ensemble models for combining their answers might lead to further improvements with respect to WSD.

7 Conclusion and Further Work

This article has outlined a view of the lexicon as a model that combines different kinds of information pertaining to various NLP tasks and methods. It has demonstrated the usefulness of combining heterogeneous data and task objectives via a number of experimental setups. Experiments with knowledge-based WSD have shown that enriching a knowledge graph with syntagmatic information from corpora can result in significant advantage over baseline structured resources. Such enriched graphs can be beneficially exploited for the creation of vector space models that are able to beat popular corpus-based VSMs on standard evaluation tasks like word similarity and relatedness calculation. The VSMs can also be used as a source of features and training data for supervised deep learning architectures – for classification or regression tasks. And finally, multi-task learning is shown to provide significant gains when multiple objectives are combined, which depend on a common lexical representation.

The described models and algorithms will be integrated in the infrastructure developed within the CLaDA-BG project, and these promising results will be used as justification to investigate more complex models based on the interaction of different kinds of lexical interfaces. Some of the potential applications include: WSD (knowledge-based and supervised); text similarity (via context embedding); improved POS tagging.

Further explorations of multi-task learning setups are certainly necessary in order to determine which tasks benefit most from co-training with WSD models, and which tasks help WSD in turn. POS tagging, for instance, does not seem like an ideal candidate from this point of view, as morphological patterning seems to be much more co-dependent with syntactic structure. Syntactic and semantic valency analysis should, however, be very good sources of complementary data that is nevertheless crucially dependent on knowledge of the lexicon. The only reason POS tagging was selected for this demonstration is that the implementation of the system is much easier. More experimental work is necessary in order to determine which auxiliary tasks do and do not help with WSD (or other problems). This should be combined with integrative work to establish links and interfaces between existing lexical models (structured, symbolic, numerical). A unified solution that is able to model language in many different ways, while sharing most of its parameters amongst the kinds of analyses it produces, would constitute a serious step towards

building multi-purpose and complexly structured linguistic and conceptual representations that resemble to a greater degree human cognition, rather than task-specific machinery.

Acknowledgements

This research has received partial support by the grant 02/12 – *Deep Models of Semantic Knowledge (DemoSem)*, funded by the Bulgarian National Science Fund in 20172019. In addition, some of the work has been done within the Bulgarian National Interdisciplinary Research e-Infrastructure for Resources and Technologies in favor of the Bulgarian Language and Cultural Heritage, part of the EU infrastructures CLARIN and DARIAH – *CLaDA-BG*, Grant DO01-164/28.08.2018. I am grateful to the anonymous reviewers for their comments and suggestions.

References

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Paşca, M., & Soroa, A. 2009. A Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches. In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics: 19-27.

Agirre, E. and Soroa, A. 2009. Personalizing PageRank for Word Sense Disambiguation. In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics: 33-41.

Agirre, E., de Lacalle, O.L. and Soroa, A. 2018. The Risk of Sub-optimal Use of Open Source NLP Software: UKB is Inadvertently State-of-the-art in Knowledge-based WSD. arXiv preprint arXiv:1805.04277.

Alonso, H.M. and Plank, B. 2016. When is Multitask Learning Effective? Semantic Sequence Prediction Under Varying Data Conditions. arXiv preprint arXiv:1612.02251.

Bingel, J. and Søgaard, A. 2017. Identifying Beneficial Task Relations for Multi-task Learning in Deep Neural Networks. arXiv preprint arXiv:1702.08303.

Bonial, C., Stowe, K. and Palmer, M. 2013. Renewing and Revising SemLink. In Proceedings of the 2nd Workshop on Linked Data in Linguistics (LDL-2013): Representing and linking lexicons, Terminologies and Other Language Data: 9-17.

Camacho-Collados, J., Pilehvar, M.T. and Navigli, R. 2015. NASARI: a Novel Approach to a Semantically-aware Representation of Items. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: 567-577.

De Lacalle, M.L., Laparra, E. and Rigau, G. 2014. Predicate Matrix: Extending SemLink through WordNet Mappings. In LREC: 903-909.

Faruqui, M., Dodge, J., Jauhar, S.K., Dyer, C., Hovy, E. and Smith, N.A. 2014. Retrofitting Word Vectors to Semantic Lexicons. arXiv preprint arXiv:1411.4166.

Fellbaum, Christiane. 1998. Wordnet. Wiley Online Library.

Goikoetxea, J., Soroa, A. and Agirre, E. 2015. Random Walks and Neural Network Language Models on Knowledge Bases. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: 1434-1439.

Harabagiu, S. M. and Moldovan, D. I. 2000. Enriching the WordNet Taxonomy with Contextual Knowledge Acquired from Text. Natural Language Processing and Knowledge Representation: Language for Knowledge and Knowledge for Language, 301-333.

Hill, F., Reichart, R., Korhonen, A. (2015). Simlex-999: Evaluating Semantic Models with (Genuine) Similarity Estimation. Computational Linguistics, 41 (4): 665-695.

Huang, Z., Xu, W. and Yu, K.. 2015. Bidirectional LSTM-CRF Models for Sequence Tagging. arXiv preprint arXiv:1508.01991.

Iacobacci, I., Pilehvar, M.T. and Navigli, R. 2015. Sensembed: Learning Sense Embeddings for Word and Relational Similarity. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers): 95-105.

Iacobacci, I., Pilehvar, M.T. and Navigli, R. 2016. Embeddings for Word Sense Disambiguation: An Evaluation Study. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers): 897-907.

Johansson, R. and Pina, L.N. 2015. Embedding a Semantic Network in a Word Space. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: 1428-1433.

Kågebäck, M. and Salomonsson, H. 2016. Word Sense Disambiguation Using a Bidirectional lstm. arXiv preprint arXiv:1606.03568.

Kiros, R., Zhu, Y., Salakhutdinov, R.R., Zemel, R., Urtasun, R., Torralba, A. and Fidler, S. 2015. Skip-thought Vectors. In Advances in Neural Information Processing Systems: 3294-3302.

Levin, Beth. 1993. English Verb Classes and Alternations: A Preliminary Investigation. University of Chicago Press.

Levy, O. and Goldberg, Y. 2014. Dependency-based Word Embeddings. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers): 302-308.

Mancini, M., Camacho-Collados, J., Iacobacci, I. and Navigli, R. 2016. Embedding Words and Senses Together via Joint Knowledge-enhanced Training. arXiv preprint arXiv:1612.02703.

Melamud, O., Goldberger, J. and Dagan, I. 2016. context2vec: Learning Generic Context Embedding with Bidirectional LSTM. In Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning: 51-61.

Mihalcea, R. and Moldovan, D. I. 2001. eXtended WordNet: Progress Report. In Proceedings of NAACL Workshop on WordNet and Other Lexical Resources: 95100.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. and Dean, J. 2013. Distributed Representations of Words and Phrases and Their Compositionality. In Advances in Neural Information Processing Systems: 3111-3119.

Mikolov, T., Chen, K., Corrado, G. and Dean, J. 2013. Efficient Estimation of Word Representations in Vector Space. arXiv preprint arXiv:1301.3781.

Miller, G.A., Leacock, C., Tengi, R. and Bunker, R.T. 1993. A Semantic Concordance. In Proceedings of the Workshop on Human Language Technology, Association for Computational Linguistics: 303-308.

Navigli, R. 2009. Word Sense Disambiguation: A Survey. ACM Computing Surveys (CSUR), 41(2): p.10.

Nguyen, D.Q., Dras, M. and Johnson, M. 2017. A Novel Neural Network Model for Joint POS Tagging and Graph-based Dependency Parsing. arXiv preprint arXiv:1705.05952.

Palmer, M. 2009. Semlink: Linking Propbank, Verbnet and FrameNet. In Proceedings of the Generative Lexicon Conference, Pisa, Italy: GenLex-09: 9-15.

Pennington, J., Socher, R. and Manning, C. 2014. Glove: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP): 1532-1543.

Plank, B., Sgaard, A. and Goldberg, Y. 2016. Multilingual Part-of-speech Tagging with Bidirectional Long Short-term Memory Models and Auxiliary Loss. arXiv preprint arXiv:1604.05529.

Popov, A. 2017. Word Sense Disambiguation with Recurrent Neural Networks. In Proceedings of the Student Research Workshop associated with RANLP: 25-34.

Raganato, A., Camacho-Collados, J. and Navigli, R. 2017. Word Sense Disambiguation: A Unified Evaluation Framework and Empirical Comparison. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers: 99-110.

Raganato, A., Bovi, C.D. and Navigli, R. 2017. Neural Sequence Learning Models for Word Sense Disambiguation. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: 1156-1167.

Rothe, S. and Schütze, H. 2015. Autoextend: Extending Word Embeddings to Embeddings for Synsets and Lexemes. arXiv preprint arXiv:1507.01127.

Ruder, S., 2017. An Overview of Multi-task Learning in Deep neural Networks. arXiv preprint arXiv:1706.05098.

Schuler, K. K. 2005. VerbNet: A Broad-coverage, Comprehensive Verb Lexicon.

Simov, K., Osenova, P., & Popov, A. 2016. Using Context Information for Knowledge-based Word Sense Disambiguation. In International Conference on Artificial Intelligence: Methodology, Systems, and Applications: 130-139. Springer, Cham.

Simov, K., Popov, A., & Osenova, P. 2015. Improving Word Sense Disambiguation with Linguistic Knowledge from a Sense Annotated Treebank. In Proceedings of the International Conference Recent Advances in Natural Language Processing: 596-603.

Wang, P., Qian, Y., Soong, F.K., He, L. and Zhao, H. 2015. Part-of-speech Tagging with Bidirectional Long Short-term Memory Recurrent Neural Network. arXiv preprint arXiv:1510.06168.

Wang, P., Qian, Y., Soong, F.K., He, L. and Zhao, H. 2015. A Unified Tagging Solution: Bidirectional LSTM Recurrent Neural Network with Word Embedding. arXiv preprint arXiv:1511.00215.

Zhong, Z. and Ng, H.T. 2010. It Makes Sense: A Wide-coverage Word Sense Disambiguation System for Free Text. In Proceedings of the ACL 2010 System Demonstrations, Association for Computational Linguistics: 78-83.

WebAnno-MM: EXMARaLDA meets WebAnno

Steffen Remus* **Hanna Hedeland** **Anne Ferger** **Kristin Bührig** **Chris Biemann***

***Language Technology Group**
Department of Informatics
Universität Hamburg, Germany
{lastname}@informatik.uni-hamburg.de

Hamburg Centre for Language Corpora (HZSK)
Universität Hamburg, Germany
{firstname.lastname}@uni-hamburg.de

Abstract

In this paper, we present WebAnno-MM, an extension of the popular web-based annotation tool WebAnno, which is designed for the linguistic annotation of transcribed spoken data with time-aligned media files. Several new features have been implemented for our current use case: a novel teaching method based on pair-wise manual annotation of transcribed video data and systematic comparison of agreement between students. To enable the annotation of transcribed spoken language data, apart from technical and data model related challenges, WebAnno-MM offers an additional view to data: a (musical) score view for the inspection of parallel utterances, which is relevant for various methodological research questions regarding the analysis of interactions of spoken content.

1 Introduction

We present WebAnno-MM[1], an extension of the popular web-based annotation tool WebAnno[2] (Yimam et al., 2013; Eckart de Castilho et al., 2014), which allows linguistic annotation of transcribed spoken data with time-aligned media files. Within a project aiming at developing innovative teaching methods, pair-wise manual annotation of transcribed video data and systematic comparison of agreement between annotators was chosen as a way to teach students to analyze and reflect *a*) on a authentic classroom communication, and *b*) on the linguistic transcription as a part of this process.

For the project, a set of video recordings were partly transcribed and compiled into a corpus with metadata on communications and speakers using the EXMARaLDA system (Schmidt and Wörner, 2014), comprising a set of data models, XML transcription and metadata formats and software tools for the creation management and analysis of spoken corpora. The EXMARaLDA system could have been further used to implement the novel teaching method, since it allows for manual annotation of audio and video data and provides methods for visualizing the transcribed data (in HTML format) for further qualitative analysis. However, within the relevant context of university teaching, apart from such requirements, addressing the peculiarities of spoken data, several additional requirements regarding collaborative annotation, user management and data management becomes an increasingly important part of the list of

[1]MM refers to Multi Modal. The application as well as the source can be found at https://github.com/webanno/webanno-mm

[2]https://webanno.github.io

Steffen Remus, Hanna Hedeland, Anne Ferger, Kristin Bührig and Chris Biemann 2019. WebAnno-MM: EXMARaLDA meets WebAnno. *Selected papers from the CLARIN Annual Conference 2018*. Linköping Electronic Conference Proceedings 159: 166–176.

desirable features. The following list illustrates the main necessities for a successful implementation of the project:

- proper handling of spoken data (e.g. speaker and time information must be maintained),
- ease of playback and display of aligned media files must be properly endowed,
- support visualization of transcripts in the required layout, i.e the (musical) score view,
- support complex manual annotation of linguistic data,
- support collaborative (i.e. pair-wise) annotation,
- support the assessment of inter-annotator agreement scores,
- support reliable user management (for student grading).

Furthermore, a web-based user environment is preferred to avoid issues regarding installation, different versions of the software, and data distribution, particularly video content, to the users (students). Another important feature was to use a freely available tool, which allows others to use the teaching method, which was developed within the project, using the same technical set-up for other areas of analysis.

While WebAnno (in its default version) fulfills some of the requirements that are not met by the EXMARaLDA system or any other similar desktop application for transcription and annotation of spoken data, it is primarily designed for annotating sequential data (mostly occurring as written text). Thus, various extensions are required for the tool to interpret and display transcriptions and video data of spoken content, which is aligned by time and thus appears in parallel. Since there are several widely used tools for creating corpora of spoken language, we preferred to rely on an existing interoperable standardized exchange format in order to enable interoperability between the tools with advanced complementary features. For this, we chose the ISO/TEI format, which is the TEI-based ISO standard 'Transcription of spoken language' (ISO/TC 37/SC 4, 2016; Schmidt, 2011).

In Section 2, we will further describe the involved components and related work, in Section 3 we will outline WebAnno-MM in more detail. Section 4 describes the novel teaching method and the use of WebAnno-MM within the university teaching context. In Section 5, we report on another emerging use case for the WebAnno-MM based on the ISO/TEI format, and we present some ideas on how to implement further improvements of the extension in order to open and generalize it for additional, more general use case scenarios related to spoken and multimodal data annotation.

2 Related work

2.1 The EXMARaLDA system

Within the context of our goals, the EXMARaLDA[3] (Schmidt and Wörner, 2014) transcription and annotation tool, including the corresponding data model and XML file format for transcription data, is the most relevant component. The tool was originally developed to support researchers in the field of discourse analysis and research on multilingualism, but has since then been used in various other contexts, e.g. for dialectology, language documentation and even with historical written data. Since spoken and multimodal data inherently displays parallel and overlapping structures, e.g. due to overlapping speech in a conversation or due to gestures accompanying speech, the tool has proven useful in various other contexts where this type of complexity in annotation structure needs to be modeled. EXMARaLDA provides support for common transcription conventions (e.g. GAT, HIAT, CHAT) and can visualize transcription data in various formats and layouts for qualitative analysis.

As shown in Figure 1, the score layout of the interface displays a stretch of speech corresponding to a couple of utterances or intonation phrases, which is well suited for transcription or annotations spanning at most an entire single utterance. Though scrolling through the transcription text is possible, a more comprehensive overview of larger spans of discourse is only available in the static visualizations generated from the transcription data. Another aspect relevant for the current use case is the rather simplistic

[3] http://exmaralda.org

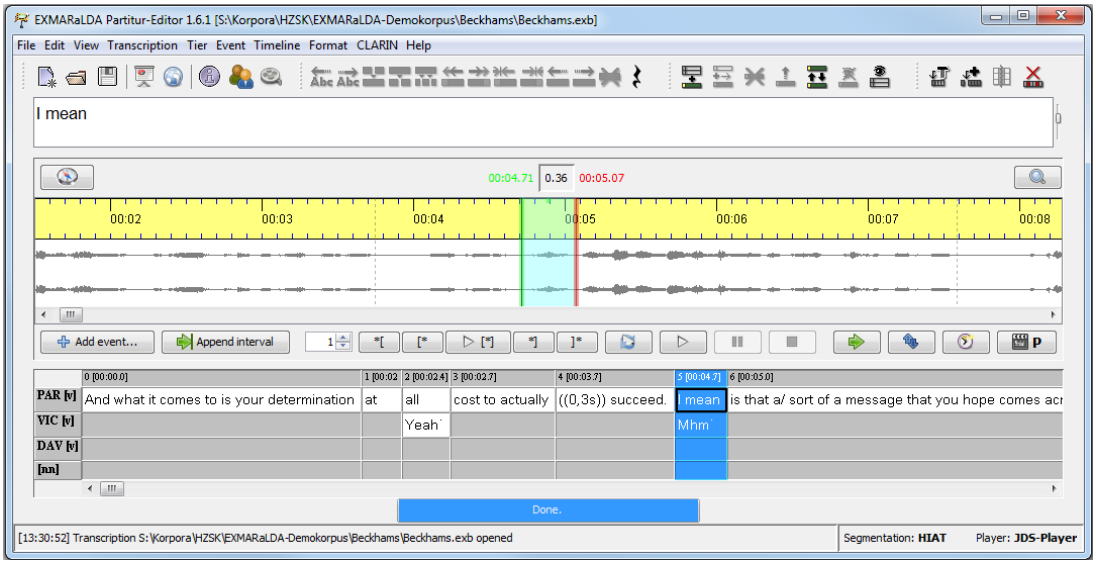


Figure 1: The musical score view of the EXMARaLDA transcription and annotation tool. The transcription is time-aligned with individual tiers for each speaker.

data model of the EXMARaLDA system, which only allows simple span annotations based directly on the transcribed text. On the one hand, this simplicity allows for efficient data processing, but on the other hand, it makes it impossible to model more complex tier dependencies or structured annotations. When annotating phenomena that occur repeatedly and interrelated over a larger span of the discourse, e.g. to analyze how two speakers discuss and arrive at a common understanding of a newly introduced concept, the narrow focus and the simple span annotations make it difficult to perform this task in an efficient and fully expressive manner.

2.2 WebAnno: a flexible, web-based annotation platform for CLARIN

In order to augment texts with linguistic annotations, automatic tools are required that support annotators and principals to collectively create, visualize, analyze, and compare annotations. WebAnno is an annotation platform that provides an interactive web interface accessible through standard web browsers while collecting, storing and processing of data takes place on a centralized server. This paradigm allows to perform shared annotation projects where annotators collectively annotate text documents with information that is immediately available on the server for further processing, e.g. for monitoring or curation purposes. Additionally, WebAnno is developed *by* the community *for* the community, and was first implemented in the CLARIN[4] context. Collaborative effort is made to increase the quality of the project and to address the needs of users who are using the tool. WebAnno offers standard means for linguistic analysis, such as span annotations, which are configurable to be locked to, or to be independent of, token or sentence annotations, relational annotations between spans, and chained relation annotations. Figure 2 shows a screenshot of WebAnno during a standard annotation task.

WebAnno is able to read various pre-defined input formats. The UIMA[5] (Ferrucci and Lally, 2004) framework is the foundation of WebAnno's backend. Hence, the input data, that is specified in a particular format and that might already contain prior annotations, are converted into UIMA's internal representation. UIMA stores text information, specifically the text itself and its annotations, in a stand-off fashion in so-called CASs (Common Analysis Structures). The basic elements needed for utilizing the underlying BRAT[6] visualization (Stenetorp et al., 2011) are then `Sentence` and `Token` annotations, which are ideally specified in the input data, and heuristically created otherwise. While `Sentence` annotations directly influence the visual segments (c.f. Figure 3), `Token` annotations are used for maintaining

[4] `https://www.clarin.eu/`
[5] Unstructured Information Management Architecture: `https://uima.apache.org/`
[6] `http://brat.nlplab.org`

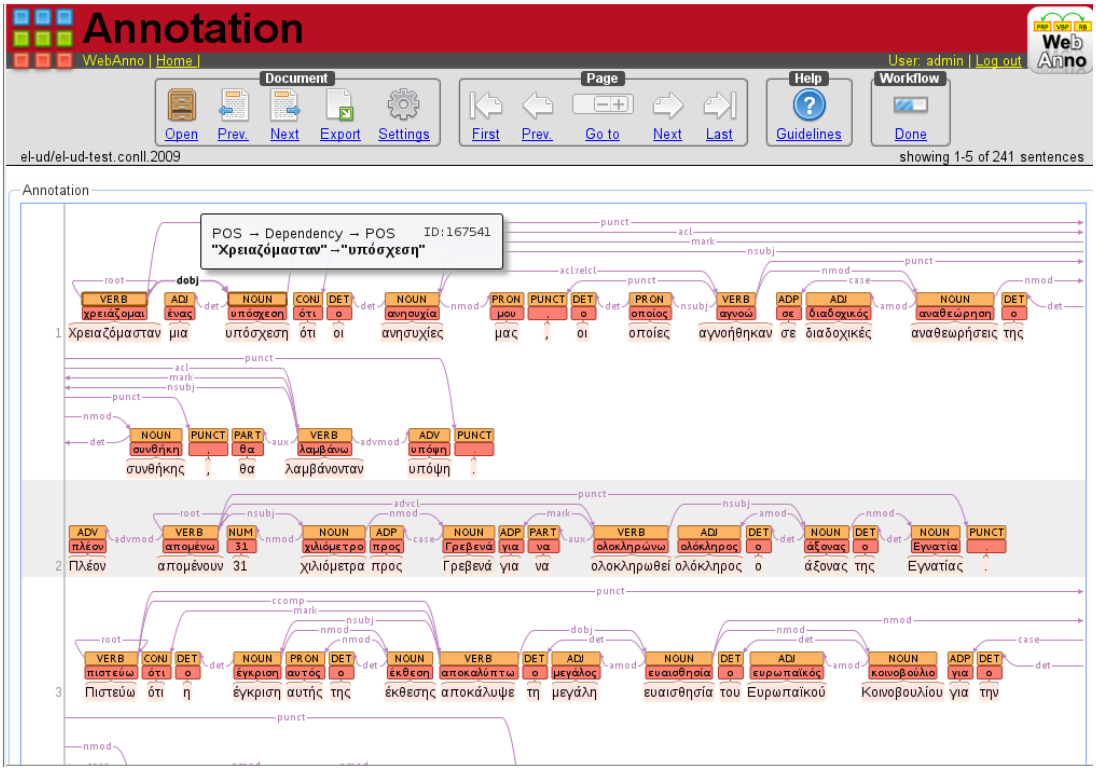


Figure 2: The default WebAnno user view during a standard annotation task. Span annotations are rendered above the annotated span showing their value, and span annotations are connected via relational annotations.

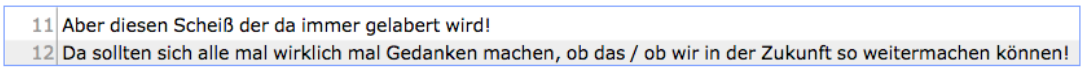


Figure 3: `Sentence` annotations define the visual segments for further annotations.

offsets between the visualization and the backend, and are further used for locking other, higher level annotations to specific token offsets. By using sentences and tokens as basic units, any input text data that is used for annotation is defined to be sequential.

For further analysis of and management of the collected annotations, WebAnno has been equipped with a set of assistive instruments, some examples of which include:

- web-based user- and project management,
- curation of annotations made by multiple users,
- built-in inter-annotator agreement measures such as Krippendorff's α, Cohen's κ and Fleiss' κ, and
- flexible and configurable annotations including extensible tagsets.

All this is available via easy web access for users (annotators), which makes it particularly suitable for research organizations and a perfect fit for the targeted use case in this work. Several extensions have been introduced, one such component is the adaptive learning component introduced by Yimam et al. (2014), where annotations and annotation values are learned by the system during the annotation tasks progress. In this setting, annotators are presented with system generated suggestions that improve by usage.

2.3 The ISO/TEI Standard for Transcription of Spoken Language

The ISO standard ISO 24624:2016 Transcription of spoken language is based on Chapter 8, Transcriptions of Speech, of the TEI Guidelines[7] as an effort to create a standardized solution for transcription data (Schmidt, 2011). Previously, TEI has rarely been used to model spoken language, since its flexibility makes various compliant versions of the format equally possible to model even the most basic elements of transcriptions, i.e. speakers' contributions, including information on their type and relative order and the alignment with media files. A number of well-established formats for transcription data exist and are only to a certain extent interoperable, due to their varying degree of complexity. These are formats of widely used transcription tools, which are usually time-based and group information in different tiers for speakers and various annotation layers. As outlined in Schmidt et al. (2017), most common transcription tool formats, including ELAN (Sloetjes, 2014) and Transcriber (Barras et al., 2000), were taken into account during the standardization process and can be modelled and converted to ISO/TEI. By using the standard, common concepts and structural information, such as speaker or time information, are modeled in a uniform way regardless of the tool used to create the data. It also becomes possible to achieve basic interoperability across transcription convention specific variants, since the standard allows for transcription convention specific units (e.g. utterances vs. intonational phrases) and labels, while still using the same elements for shared concepts.

3 WebAnno-MM: Adapting WebAnno for multimodal content and spoken data

3.1 Transcription, theory and user interfaces

A fundamental difference between linguistic analysis of written and spoken language is that the latter usually requires a preparatory step; the transcription. Most annotations are based not on the conversation or even the recorded signal itself but on its written representation. That the creation of such a representation is not an objective task, but rather highly interpretative and selective, and the analysis thus highly influenced by decisions regarding layout and symbol conventions during the transcription process, was addressed already by Ochs (1979). In particular, different arrangements of the speakers' contributions stress different aspects of conversation. If all contributions are sorted and placed underneath each other, so-called line notation, the conversation might appear more ordered with focus on the transitions, i.e. the turn-taking, than when using a musical score layout. Since the exact onset of a speaker's contributions in relation to other speakers' contributions is visualized accordingly in the score layout, this arrangement of the contributions rather stresses the simultaneity and the overlapping nature of speech.

It is therefore crucial that tools for manual annotation of transcription data respect these theory-laden decisions comprising the various transcription systems in use within various research fields and disciplines. Apart from this requirement on the GUI, the tool also has to handle the increased complexity of "context" inherent to spoken language: While a written text can mostly be considered a single stream of tokens, spoken language features parallel structures through simultaneous speaker contributions or additional non-verbal information or other relevant events in the transcription. In addition to the written representation of spoken language, playback of the aligned original media file is another imperative requirement.

3.2 From EXMARaLDA to ISO/TEI

To allow for transcription system specific characteristics, e.g. units of the transcription such as utterances or intonation phrases, the existing conversion from the EXMARaLDA format to the tool-independent ISO/TEI standard is specific to the conventions used for transcription. For our use case, this was the HIAT transcription system as defined for EXMARaLDA by Rehbein et al. (2004). Apart from the generic transcription tier holding verbal information, and non-verbal or non-linguistic information encoded as incidents in ISO/TEI, the HIAT conventions also define the following optional but standardized annotation layers: *a*) `akz` for accentuation/stress *b*) `k` for free comments *c*) `sup` for suprasegmental information *d*) `pho` for phonetic transcription. Though some common features can be represented in a generic

[7] `http://www.tei-c.org/Guidelines/P5/`

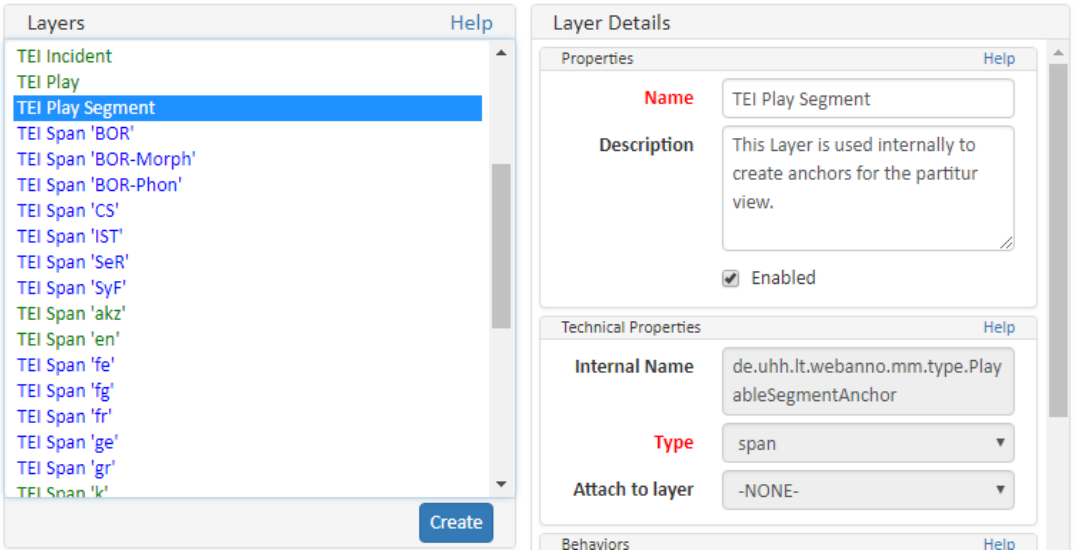


Figure 4: Definition of predefined WebAnno-MM annotation layers (green), and custom annotation layers (blue), which are both resolved and matched during TEI import.

way by the ISO/TEI standard, for reasons described above, several aspects of the representation of conversations must remain transcription convention specific, e.g. the kind of linguistic units defined below the level of speaker contributions.

Furthermore, metadata is stored in different ways for various transcription formats. Within the EXMARaLDA system, metadata on sessions and speakers is stored and managed separated from the transcriptions in the EXMARaLDA Corpus Manager XML format to enhance consistency. The ISO/TEI standard on the other hand, as any TEI variant, can make use of the TEI Header to allow transcription and annotation data and various kinds of metadata to be exported and further processed in one single file, independent of the original format. This approach has also been implemented for the WebAnno-MM extension to be able to retrieve e.g. basic metadata on speakers, such as their age or linguistic knowledge, while annotating.

3.3 Parsing ISO/TEI to UIMA CAS

A major challenge is the presentation of time-aligned parallel transcriptions (and their annotations) of multiple speakers in a sequence without disrupting the perception of a conversation, while still keeping the individual segmented utterances of speakers as a whole, in order to allow continuous annotations. For our use case, we restrict WebAnno-MM to the ISO/TEI standard with HIAT conventions as described in Section 3.2. Annotations in the standardized HIAT annotation format (cf. Rehbein et al., 2004) are recognized as pre-defined annotations, where each TEI (HIAT) type is linked with a matching WebAnno-MM annotation layer on import. By default, unrecognized TEI types (non-HIAT) are merged into a single generic WebAnno-MM annotation layer where the TEI type information is preserved as an annotation feature. Since the TEI standard does generally not restrict annotation types, and in order to still allow the import of HIAT unrelated TEI annotations as new WebAnno-MM layers, which might be required for future projects or research purposes, we adapt the importer, such that any TEI annotation type[8] which occurs in the transcription file and for which a matching WebAnno-MM layer (including certain annotation features) must have been created manually before the actual import, is linked to the particular matching annotation layer. Figure 4 shows the generated ISO/TEI layers in addition to WebAnno's existing annotation layers (e.g. 'Dependency' or 'Lemma'). While 'TEI Incident' or 'TEI Play (Segment)' are necessary layers corresponding to generic components of the ISO/TEI standard, and will thus be generated for all ISO/TEI data, other layers layers have been predefined according to the HIAT transcription

[8]Note that currently only time-based span annotations are supported. Word-based span annotations will be implemented in the future.

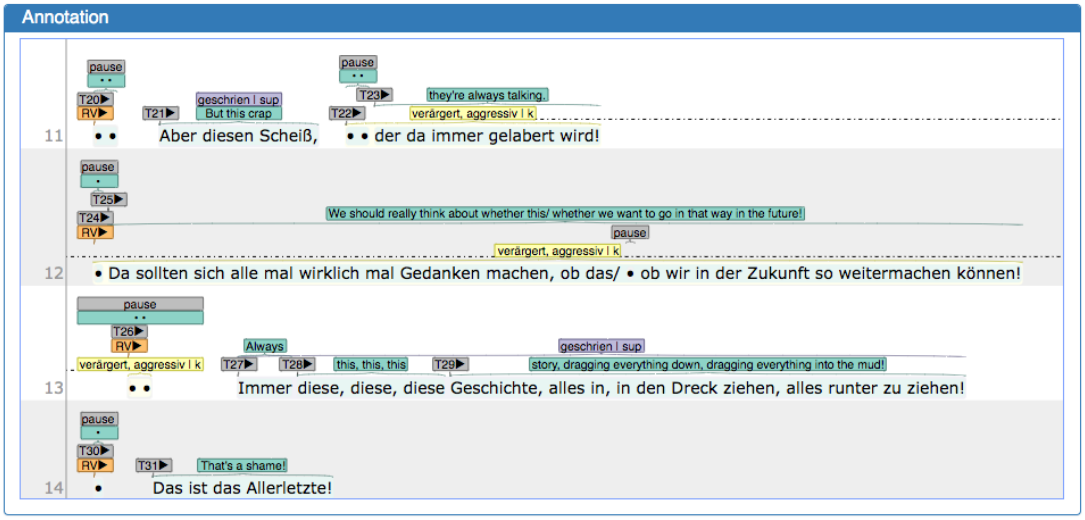


Figure 5: Screenshot of WebAnno-MM's annotation view. Note the additional time marker annotations (with ► as suffix), which are used for synchronizing WebAnno-MM's score view.

conventions used to create the transcripts (`akz`, `en`, `k` and `sup`, cf. Rehbein et al., 2004). Additionally, it is also possible to add custom annotation layers describing project specific annotation, which might be independent of the HIAT standard and can be freely defined in the ISO/TEI. The creation of these new layers follow standard naming policies for automatic matching of TEI span types to the actual annotation layer. The definition must be finished prior the import of data, otherwise the system ignores unknown TEI span types. During the import, the ISO/TEI XML content is parsed and utterances of individual speakers are stored in different 'views' (sub-spaces of CASes with different textual content). Time alignments are kept as metadata within a specific CAS which we call the *speakercas*. `Segments` are considered as the basic unit of the transcription in TEI format and can be considered to be roughly equivalent to sentences. As such, segments are directly converted to sentence annotations, which are sequentially aligned in the order of their occurrence in the so-called *annotation view*.

Note, that a segment within an `annotationBlock` XML element is considered non-disruptive. It can safely assumed that ISO/TEI span annotations are within the time limits of the one utterance that is bound to the annotation block. Hence we can map easily cross-segment annotations, but not cross-utterance annotations. This mainly happens for incident annotations that are speaker independent. We neglect such cases in the annotation view and remark that those occurrences are correctly presented in WebAnno-MM's (musical) score view (explained in detail in the next section).

3.4 New GUI features

In order to show utterances and annotations in an established and widely accepted environment for parallel visualization of transcribed audio content, e.g. similar to EXMARaLDA's score layout of the partitur editor, we adapt the existing html show case demos[9], and call this view the *score view* of WebAnno-MM henceforth. From a user perspective, a new browser window will open on click of a time marker in the annotation view – time markers are implemented as zero width span annotations starting with the respective speaker abbreviation or time marker id and ending with a play button character (►, see Figure 5). All such markers are clickable and trigger the focus change in the score view and start or pause the media replay. The design choice of using two parallel and synchronized browser windows was chosen because of maximum flexibility and the assumption that users have ideally two screens available, where each view sides on a different physical screen. Figure 6 schematically illustrates this operation mode.

Figure 7 shows a screenshot of the adjustable (musical) score view. Metadata (if provided in the TEI source file), such as speaker information, can be accessed via clicking on a specific speaker row, and

[9]EXMARaLDA show case demos are available at `http://hdl.handle.net/11022/0000-0000-4F70-A`

Figure 6: Schematic overview of the synchronization feature and the intended application operation mode: Two monitors, where one holds the browser window of the annotation view and the other holds the browser window of the score view, synchronized by clicking on the respective time markers.

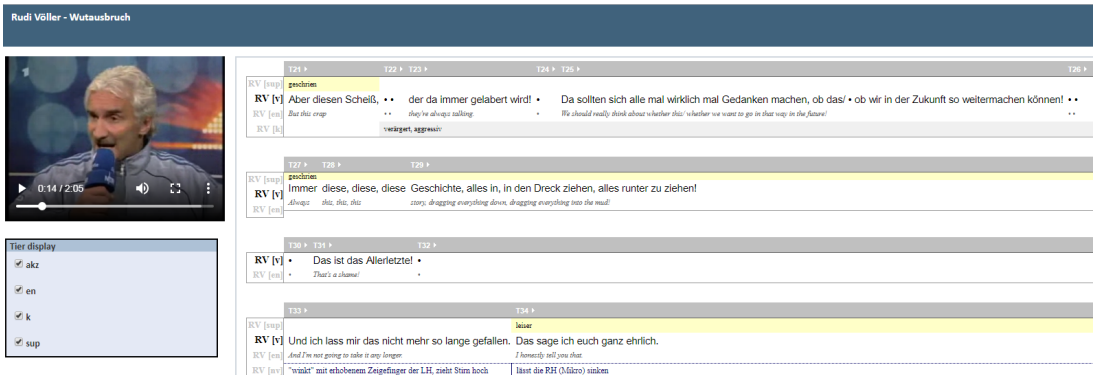


Figure 7: Screenshot of WebAnno-MM's score view, which resembles EXMARaLDA's score view in the partitur editor. WebAnno-MM's annotation view and the score view are synchronized by clicking the respective time markers. The media playback (audio or video) is shown on the top left and reactive to click events on time markers. The parallel tiers are displayed on right which can be activated and deactivated via the checkboxes in the lower left.

further metadata, i.e. relevant to the recording itself can be accessed via clicking the title of the document (cf. Figure 8).

The score view itself is reverse synchronized by clicking on the time marker. The selection of the focus for the annotation view is heuristically determined by using the first marker annotation (independent of the speaker utterance). Upon a synchronizing action, the focus will immediately switch to the other window. Additionally, users are able to select from multiple media formats (if provided during project setup), such as plain audio media, or additionally video media. Also, the score view offers multiple media formats for selection, viewing speaker- or recording related details and a selectable width of the transcribed speaker tracks (rows).

Each segment starts with a marker showing the respective speaker. All markers are clickable and trigger the focus change in the score view and start or pause the media media replay.

For the purpose of managing media per document, a media pane was added to the project settings (cf. Figure 9). Here, media files can be uploaded or web URLs can be specified and linked to uploaded documents. Uploaded media files are hosted within the WebAnno-MM server infrastructure, benefitting from access restrictions through its user management. Additionally, we added support for streaming media files that are accessible in the web by providing a URL instead of a file. Multiple media files can be mapped to multiple documents, which allows user preferred reuse of different media locations of multiple document recordings.

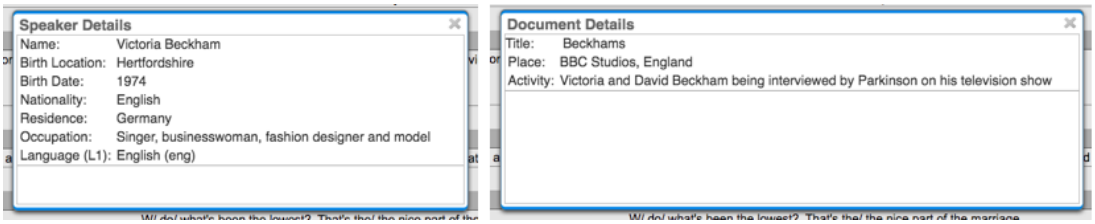


Figure 8: Speaker details (left) and recording details (right), provided by the score view and initially specified in the ISO/TEI source file.

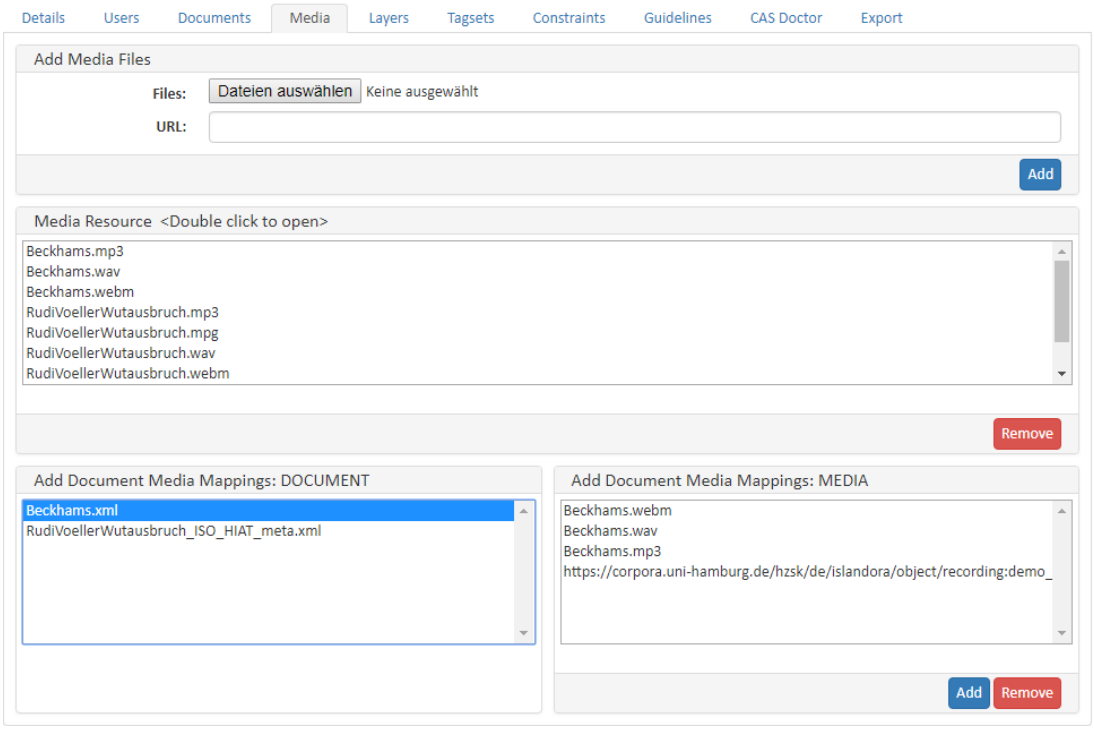


Figure 9: Screenshot of the WebAnno-MM media management pane.

4 WebAnno-MM in practice: An innovative teaching study

As part of a so-called "teaching lab", WebAnno-MM was used by teams of students participating in a university seminar to collaboratively annotate videotaped authentic classroom discourse. Thematically, the seminar covered the linguistic analysis of comprehension processes displayed in classroom discourse. The seminar was addressed to students in pre-service teacher training and students of linguistics. Students of both programs were supposed to cooperate on interdisciplinary teams in order to gain the most from their pedagogic as well as their linguistic expertise. The students had to choose their material according to their own interest from a set of extracts of classroom discourses from various subject matter classes. Benefiting from the innovative ways to decide on units of analysis such as spans, chains, etc., different stages of the process of comprehension were to be identified and then to be described along various dimensions relevant to comprehension. This approach made single steps of analysis transparent for the students, and thus allowed for their precise and explicit discussion in close alignment with existing academic literature. Compared to past seminars with a similar focus, but lacking the technological support, these discussions appeared more thoughtful and more in-depth. The students easily developed independent ideas for their research projects. Students remarked on this very positively in the evaluation of the seminar.

5 Outlook

By implementing an extension of WebAnno, we showed that it is possible to repurpose a linguistic annotation tool for multimodal data. According to the intended use case, the first release focused only on data transcribed according to the HIAT conventions using the EXMARaLDA transcription and annotation tool, thus at first restricting the possible annotation layers. Later releases allow for the creation of custom annotation layers for existing project or research question specific annotation tiers in the transcription data.

Instead of relying on one of the widely used transcription tool formats, we used the ISO/TEI standard, which can model transcription data produced by various tools and according to different transcription conventions, as an exchange format. Obvious next steps would therefore be to extend the interoperability to include full support for the ISO/TEI format. This would require adapted import functionality and transcript visualization functionality for further transcription systems, as well as a generic fallback option. With a more general interoperability, annotation projects could be based on data from several contexts (cf. Ligeois et al. (2015)). Another relevant aspect is supporting the standard-compliant modelling of segment-based annotations, in contrast to annotations that are time-based, i.e. aligned with the base layer and sharing features such as the speaker identity. Segment-based annotations also allow for annotation layers with a more complex structure, which is in some cases required to explicitly model grammatical information (Arkhangelskiy et al., 2019). The word or segment-based annotations are closer to the text-oriented data models used by WebAnno and will be supported in a future release of WebAnno-MM. Other important tasks to take on are extensions of the ISO/TEI standard to model both metadata in the TEI Header and the complex annotations generated in WebAnno in a standardized way.

Acknowledgments

This work was partially funded by the project 'Interaktives Annotieren von Unterrichtskommunikation' in the program Lehrlabor Lehrerprofessionalisierung (L3Prof), Universität Hamburg.[10]

References

Timofey Arkhangelskiy, Anne Ferger, and Hanna Hedeland. 2019. Uralic multimedia corpora: ISO/TEI corpus data in the project INEL. In *Proceedings of the Fifth International Workshop on Computational Linguistics for Uralic Languages*, pages 115–124, Tartu, Estonia.

Claude Barras, Edouard Geoffrois, Zhibiao Wu, and Mark Liberman. 2000. Transcriber: development and use of a tool for assisting speech corpora production. *Speech Communication – Special issue on Speech Annotation and Corpus Tools*, 33(1–2).

David Ferrucci and Adam Lally. 2004. UIMA: An Architectural Approach to Unstructured Information Processing in the Corporate Research Environment. *Natural Language Engineering*, 10(3–4):327–348.

ISO/TC 37/SC 4. 2016. Language resource management – Transcription of spoken language. Standard ISO 2462:2016, International Organization for Standardization, Geneva, Switzerland.

Loc Ligeois, Carole Etienne, Christophe Parisse, Christophe Benzitoun, and Christian Chanard. 2015. Using the TEI as a pivot format for oral and multimodal language corpora. Paper presented at Text Encoding Initiative Conference, Lyon, 28–31, 2015.

Richard Eckart de Castilho, Chris Biemann, Iryna Gurevych, and Seid Muhie Yimam. 2014. WebAnno: a flexible, web-based annotation tool for CLARIN. In *Proceedings of the CLARIN Annual Conference 2014*, pages 1–3, Soesterberg, Netherlands.

Elinor Ochs. 1979. Transcription as theory. In E. Ochs and B.B. Schieffelin, editors, *Developmental pragmatics*, pages 43–72. Academic Press, New York.

Jochen Rehbein, Thomas Schmidt, Bernd Meyer, Franziska Watzke, and Annette Herkenrath. 2004. Handbuch für das computergestützte Transkribieren nach HIAT. *Arbeiten zur Mehrsprachigkeit, Folge B*, 56:1 ff. (in German).

[10] https://www.zlh-hamburg.de/entwicklungsvorhaben/lehrlabore/lehrlabor-lehrerbildung

Thomas Schmidt and Kai Wörner. 2014. EXMARaLDA. In Jacques Durand, Ulrike Gut, and Gjert Kristoffersen, editors, *Handbook on Corpus Phonology*, pages 402–419. Oxford University Press.

Thomas Schmidt, Hanna Hedeland, and Daniel Jettka. 2017. Conversion and annotation web services for spoken language data in clarin. In *Selected papers from the CLARIN Annual Conference*, number 136, pages 113–130, Aix-en-Provence, France. Linköping University Electronic Press, Linköpings Universitet.

Thomas Schmidt. 2011. A TEI-based Approach to Standardising Spoken Language Transcription. *Journal of the Text Encoding Initiative*, 1:1–28.

Han Sloetjes. 2014. ELAN: Multimedia annotation application. In Jacques Durand, Ulrike Gut, and Gjert Kristoffersen, editors, *Handbook on Corpus Phonology*, pages 305–320. Oxford University Press.

Pontus Stenetorp, Goran Topić, Sampo Pyysalo, Tomoko Ohta, Jin-Dong Kim, and Jun'ichi Tsujii. 2011. Bionlp shared task 2011: Supporting resources. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 112–120, Portland, OR, USA.

Seid Muhie Yimam, Iryna Gurevych, Richard Eckart de Castilho, and Chris Biemann. 2013. WebAnno: A flexible, web-based and visually supported system for distributed annotations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 1–6, Sofia, Bulgaria.

Seid Muhie Yimam, Chris Biemann, Richard Eckart de Castilho, and Iryna Gurevych. 2014. Automatic annotation suggestions and custom annotation layers in WebAnno. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 91–96, Baltimore, MD, USA.

SenSALDO: a Swedish Sentiment Lexicon for the SWE-CLARIN Toolbox

Jacobo Rouces Lars Borin
Nina Tahmasebi Stian Rødven Eide
Språkbanken Text
University of Gothenburg, Sweden
jacobo.rouces|lars.borin@gu.se
nina.tahmasebi|stian.rodven.eide@gu.se

Abstract

The field of *sentiment analysis* or *opinion mining* consists in automatically classifying text according to the positive or negative sentiment expressed in it, and has become very popular in the last decade. However, most data and software resources are built for English and a few other languages. In this paper we compare and test different corpus-based and lexicon-based methods for creating a sentiment lexicon. We then manually curate the results of the best performing method. The result, SenSALDO, is a comprehensive sentiment lexicon for Swedish containing 7,618 word senses as well as a full-form version of this lexicon containing 65,953 items (text word forms). SenSALDO is freely available as a research tool in the SWE-CLARIN toolbox under an open-source CC-BY license.

1 Introduction

The field of *sentiment analysis* or *opinion mining* consists in automatically classifying text according to the positive or negative sentiment expressed in it, and has become very popular in the last decade (Pang and Lee, 2008). However, most data and software resources are built for English or a few other languages, and there is still a lack of resources for most languages. While often discussed in the NLP literature as a business-intelligence tool – helping online businesses keep track of customer opinion about their goods and services – there have also been a number of studies where sentiment analysis has been applied to research data in the humanities and social sciences (HSS) (Bentley et al., 2014; Eichstaedt et al., 2015; Sprugnoli et al., 2016; Thelwall, 2017). This has prompted inquiries by Swedish HSS researchers as to whether the Swedish CLARIN infrastructure could provide this kind of tool also for Swedish textual data. For this reason, at the CLARIN B center Språkbanken Text (University of Gothenburg) we initiated a concerted effort aiming at the development of a Swedish sentiment lexicon for the SWE-CLARIN toolbox.

The development of this resource – *SenSALDO* – has been done in three steps, described below: (1) creation of a gold standard word-sense list (section 3); (2) implementation and evaluation of different automatic methods for creating the sentiment lexicon (section 4); and (3) manual curation of the results of the best performing method (section 5).

In SenSALDO, each word sense has two annotations: a coarse-grained label with three possible values ('positive', 'neutral', and 'negative') and a more fine-grained score in the range $[-1, 1]$.

2 Lexical resources for sentiment analysis: state of the art

In recent years, a wide variety of methods has been used for building sentiment lexicons. Unsurprisingly, most of this work has focused on English, although some efforts targeting other languages have also been reported in the literature.

Some methods rely on corpus analysis (making use of word co-occurrence, syntactic patterns, or distant-supervision signals) and others on existing lexicons (usually utilizing some sort of sentiment label propagation exploiting the structure of the lexicon), although both approaches can be combined (Devitt

Jacobo Rouces, Lars Borin, Nina Tahmasebi and Stian Rødven Eide 2019. SenSALDO: a Swedish Sentiment Lexicon for the SWE-CLARIN Toolbox. *Selected papers from the CLARIN Annual Conference 2018*. Linköping Electronic Conference Proceedings 159: 177–187.

and Ahmad, 2013). In some cases manual annotation is used as well, either to create seeds or to curate results.

Among the English sentiment lexicons built with mostly-automatic lexicon-driven methods, SentiWordNet (Baccianella et al., 2010) has become a popular resource. It was created by combining a semi-supervised learning step that uses existing relations between WordNet 3.0 entries (Fellbaum, 1998), (such as *synonymy*, *antonymy*, and *related with*), and a random-walk step over a graph built using the implicit *definiens–definiendum* relation between words in the entries and words in the glosses of the entries (Esuli and Sebastiani, 2007). However, the use of these relations requires a structure like that of WordNet, which in turn requires a considerable amount of manual effort by trained lexicographers.

Among the English lexicons built using corpus-driven approaches, SENTPROP (Hamilton et al., 2016) constitutes a recent state-of-the-art approach where a directed weighted graph of terms is constructed using the nearest neighbors in the space of word embeddings obtained from applying singular value decomposition to the positive pointwise mutual information matrix obtained from the corpus. Then, it uses random walks in a similar fashion to SentiWordNet.

A common problem when using label propagation is that words far away from seeds get low values just by virtue of their distance, which should not be the case. Yazidi et al. (2015) propose a solution: at each iteration, a fixed number of "informative words" are selected as new seeds for labeling according to different criteria.

For Swedish, the language of interest in our case, much less work has been reported in the literature. However, two openly available sentiment lexicons existed prior to the work reported here, presented by Rosell and Kann (2010) and Nusko et al. (2016). In addition, some Swedish sentiment lexicons or word lists have been produced by automatic translation of corresponding English resources, e.g., by Mohammad and Turney (2010)[1] and Chen and Skiena (2014). Looking at these existing resources, there are obvious ways in which they can be improved. With automatically translated resources, there are more than a few strange translations or translation errors. In many of these resources the lexical items are undisambiguated text words or lemmas, i.e., all senses of polysemous words are conflated in the lexicon, even though the different senses may of course have different sentiment values. In the case of the resources organized by lemmas, there is generally no attempt at full-form expansion, i.e., having the lexicon cover all inflected forms of a lemma, which obviously makes the lexicon less useful for automated text processing. Both these issues are addressed in the work reported here, at the same time that we have been able to draw inspiration from the work of both Rosell and Kann (2010) and Nusko et al. (2016). In this sense, our lexicon builds on their earlier efforts.

3 Step 1: a gold-standard sentiment word-sense list for Swedish

The first step of our work consisted in producing a gold-standard list of sentiment-annotated Swedish word senses. This work has been reported elsewhere, and here we just provide some necessary background information. For the details, see Rouces et al. (2018a) and Rouces et al. (2018b).

SenSALDO is based on SALDO, a computational lexicon for Swedish composed, among other components, of word senses as entries and semantic relations – called *descriptors* – connecting word senses. There are two kinds of descriptors: The *primary descriptor* is obligatory. It connects an entry to exactly one other word sense (also a SALDO entry[2]). This parent word sense is a close semantic neighbor which is also more central, which means that is typically structurally simpler, stylistically more neutral, acquired earlier by a first-language learner and more frequent in usage. Any number of secondary descriptor relations provide additional semantic properties of the entry that are not conveyed by the primary descriptor, such as an inversion or negation or some important semantic argument in a hypothetical definition. The primary descriptor structure forms a tree and the secondary descriptors define a directed acyclic graph. For a detailed description of the organization of SALDO and a discussion of the underlying linguistic

[1] `http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm`

[2] Except in the case of 41 top entries, which are given an artficial primary descriptor in order to make all of SALDO into a single rooted tree.

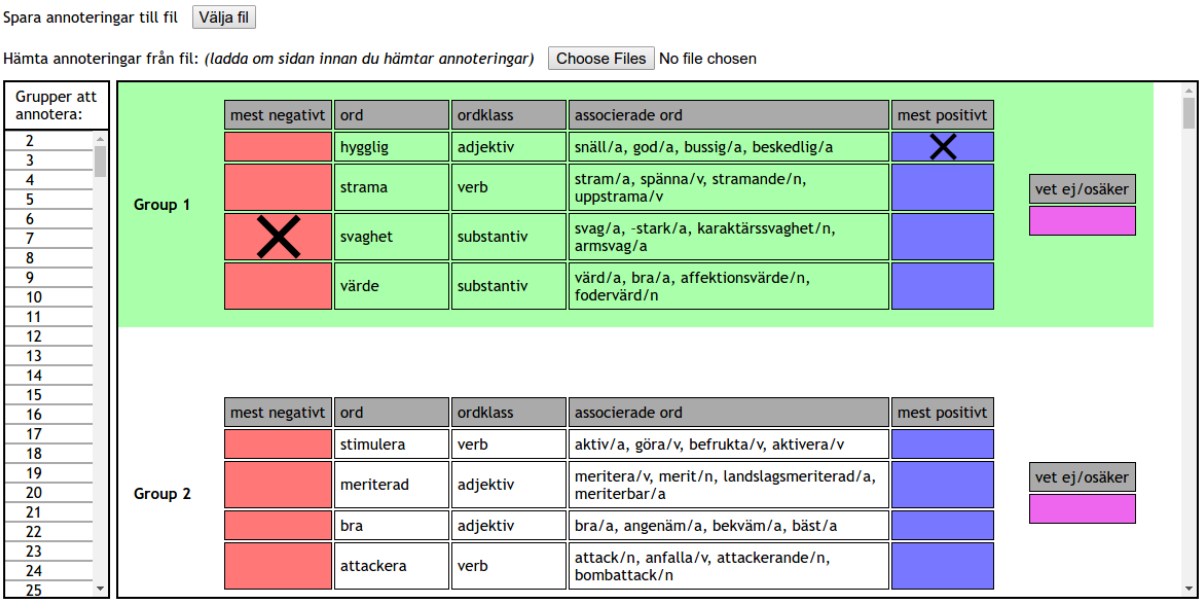


Figure 1: Screenshot for the best–worst scaling annotation interface. The labels displayed for each group are (from left to right) 'most negative', 'word', 'part of speech', 'associated words', 'most positive', 'don't know/uncertain'.

theoretical and methodological principles informing it, see Borin et al. (2013). For the work described here, we used the current stable version SALDO v. 2.3, which contains 131,020 word senses.[3]

First, an initial sampling from SALDO was done following the distribution given by the estimated frequency of each word sense in the *Swedish Culturomics Gigaword Corpus* (Eide et al., 2016), which is a one-billion-word mixed-genre corpus of written Swedish.[4] We sampled ∼2,200 open-class words (nouns, verbs, adjectives and interjections), which were annotated by three annotators (the three last authors of this paper), with 200 overlapping items in order to estimate interannotator agreement. This annotation was done using discrete labels with three possible values (-1, 0, and $+1$, for negative, neutral, and positive sentiment, respectively).

After this, four external annotators were employed to annotate a subset of the 2,200 items, such that at least two of the initial annotators had asigned a non-neutral value to each item. The resulting 278 word senses were annotated using *best–worst scaling* (Kiritchenko and Mohammad, 2016) through a web interface developed for this purpose, shown in figure 1.

The histograms in figure 2 show the distribution of the sentiment values obtained with direct and best–worst scaling annotation, illustrating the effectiveness of the preliminary filtering steps in ensuring that the best–worst scaling annotators were presented mainly non-neutral items.

4 Step 2: method evaluation

The methods that we compare can be divided in two categories: graph-based algorithms using the SALDO descriptors and other lexicon-based relations, and corpus-based methods using dimensions from word embeddings as features for different classifiers.

We model the sentiment associated to a word sense using a value in the interval $[-1, 1]$, where $+1$ represents a totally positive sentiment and -1 represents a totally negative sentiment. After having considered using a three-dimensional model like that of SentiWordNet (Baccianella et al., 2010), we found that experimental evidence indicated that the average overlap between positivity and negativity in the same word was very low (Rouces et al., 2018b).

[3]SALDO is freely available (under a CC-BY license) at `https://spraakbanken.gu.se/eng/resource/saldo`.

[4]The corpus is freely available (under a CC-BY license) at `https://spraakbanken.gu.se/eng/resource/gigaword`.

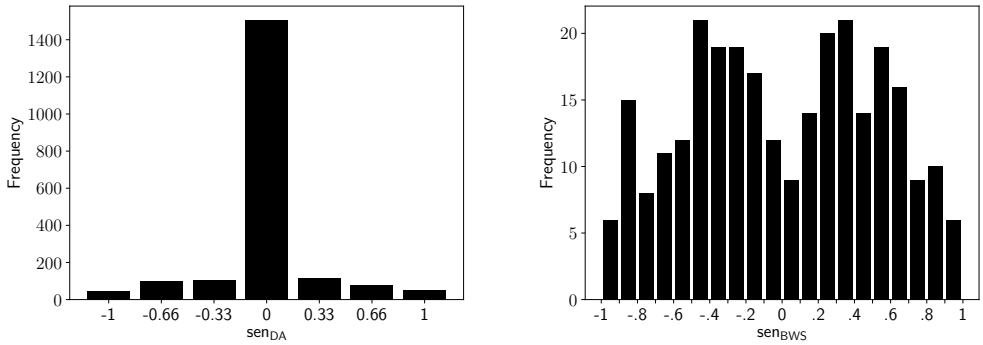


Figure 2: Histograms of the sentiment values resulting from direct annotation (left) and best–worst scaling annotation (right)

We experiment with different approaches, which we describe below. We extend the methods of Rosell and Kann (2010) and Nusko et al. (2016) and we also try a corpus-oriented approach similar to the one described by Hamilton et al. (2016). For all methods, we produce continuous scores as well as discrete labels: $+1$ (positive), 0 (neutral), -1 (negative). What is relevant about the continuous scores is not their magnitudes but the relative order that they produce. The values and their distributions depend on idiosyncrasies of the methods employed and do not necessarily resemble what would be produced by direct human annotations, but they can be fit to any desired distribution. The discrete labels are less fine-grained, but may be more appropriate for certain applications.

4.1 Inheritance over graph

Nusko et al. (2016) propose a tree traversal method on the tree defined by the primary descriptor relation between SALDO entries. This method starts with 6 seeds with a manually assigned polarity and recursively calculates the sentiment of children based on the sentiment of the parent.[5] The algorithm calculates a confidence score for each sentiment, which decreases at a constant rate with the distance to the original seed (steps of -0.25 from a confidence of 1 for the descendants of the core words), and sets a threshold of 0.5 as the lowest acceptable confidence. It also uses secondary descriptors, but only when the secondary descriptor is *inte* ‘not’, which indicates that the child and parent have opposite semantic values and therefore the sign on the sentiment value should also be inverted, or a strength modifier like *lite* ‘a little’, or *enastående* ‘outstanding’. It obtains a sentiment for 2,133 entries. Three annotators independently labeled 150 of the entries as positive, negative or neutral, and for 117 of these the annotators were in full agreement. From these a 71% precision was obtained. The 150 entries were sampled using equally sized stratification over the three confidence levels.

Our first method is a modified and extended version of this method. The extensions allows it to inherit semantic values from the secondary descriptors too. In particular, the traversal occurs over the directed acyclic graph defined by using both primary and secondary descriptors. In this way, the secondary descriptors of an entry are used not only for polarity inversion or intensification, but also their sentiment value is used, although with a lower weight.

The algorithm cannot use a simple breadth-first exploration over this graph, because for a given node, in general, some incoming neighbors will be at a different distance from the seed set than others, and the node will be reached before all the incoming neighbors have been calculated. This prevents all elements in the frontier to be expanded in a single iteration. Depth-first is not suitable for similar reasons.

Therefore, different partially-successful passes over the frontier have to be tried. However, even when doing this, the algorithm will stagnate as soon as some secondary descriptors are not reachable from the given seed words. For this reason, the algorithm applies a best-effort mechanism when the previous approach stagnates, whereby the node with the lowest possible number of unreached secondary-descriptor

[5] In Nusko et al. (2016) the seeds and their children are referred to as “core words” and “seeds” respectively.

incoming nodes is chosen, and the sentiment is calculated for this node ignoring its unreached incoming nodes, and a new pass is performed over the frontier. Primary descriptors are never ignored. A priority queue is used for the nodes with unreached secondary-descriptor incoming nodes. If it is still not possible to calculate new nodes from all their parents, the process is repeated until it is possible, or the queue and the frontier are empty.

This method outputs scores, so in order to obtain discrete labels we apply thresholds. The thresholds are obtained from the percentiles of each class in a training set obtained from sampling two thirds of the gold standard, which constitutes a very basic kind of supervised learning. The other third is used for testing.

4.2 Random paths over graphs

For our second experimental setup, we develop an adaptation of the method by Rosell and Kann (2010), who developed a Swedish sentiment lexicon using random walks over a graph of synonyms and 4 positive and 4 negative seed words. The graph was built using the *Synlex/People's Dictionary of Synonyms* (Kann and Rosell, 2005), which used Swedish–English lemma pairs concatenated with their inverse relation to generate candidate synonym pairs. The pairs were filtered by grading and then averaging the grades. The result of Synlex was 16,006 words with 18,920 weighted pairs, which were used as edges of the graph in the random walks.

Our modification consists of adapting Synlex to use SALDO word senses instead of Swedish sense-ambiguous lemmas (the adaptation was done by a trained linguist, adapting the original weights to the $(0, 1]$ interval), and the union of the following sets of edges with an heuristic weight of 0.5.

- The edges defined by primary descriptors in SALDO. This component ensures that there are no isolated nodes, since every node has one primary descriptor.
- The edges defined by secondary descriptors in SALDO.
- The edges that connect SALDO entries that have the same primary descriptor (siblings). This creates a relation which is often often tantamount to co-hyponymy.

The discrete labels are obtained using the same thresholding method as in the inheritance-based method.

4.3 Classification over word2vec

As opposed to the previous methods, which are purely lexicon-driven, the third approach is partly corpus-based. We use already existing vector representations of SALDO word senses derived from *word2vec* lemma embeddings (Johansson and Nieto Piña, 2015) by means of solving a constrained optimization problem. The vector space dimensionality is 512 and the source for the vector representations was the Gigaword corpus (see section 3). Because the elements of the vector space are SALDO word senses, and the problem solved in Johansson and Nieto Piña (2015) uses the SALDO descriptor relations, this is not a purely corpus-based approach but a mixed one. We train a logistic regression classifier (word2vec-logit) and a support vector classifier with a radial basis function kernel (word2vec-svc-rbf). All the classifiers used a one-vs-rest approach of the three-class classification. For the classifiers we used 5-fold cross-validation stratified by the (positive, neutral, negative) classes. For each fold, the SVM/RBF meta-parameters (C, γ) were estimated using 5-fold cross-validation over the training set.

The classifiers' final output are discrete labels (positive/pos, neutral/neu, negative/neg), but scores are obtained computing $p((pos) - p(neg))$, where p is the probability for a given entry to belong to the positive or negative classes. For the logit classifier, p is straightforward. For the support vector classifier, we use an extension of Platt scaling for multiple classes (Wu et al., 2004).

4.4 Results

For training and testing the different methods, we used the direct annotation gold standard developed by Rouces et al. (2018b) (see section 3), which contains of 1,998 entries from SALDO entries labeled as negative (-1), neutral (0), or positive $(+1)$. The values were averaged over three annotators (so if an entry is labeled as positive by two annotators and as neutral by one, the final value would be $2/3$).

Table 1 shows the results for each method. We employ two different sets of measures for measuring the quality of the gold standard: ones based on ranks and (Spearman rank-order correlation coefficient ($\rho \in [-1, 1]$, the p-normalized Kendall tau distance ($\tau_\in [0, 1]$, and Kendall's tau-b (τ_b)) others based on discrete labels (precision, recall and confusion matrix).

- The rank-based measures are the Spearman rank-order correlation coefficient (ρ) (Kokoska and Zwillinger, 2000), in the interval $[-1, 1]$ (Myers and Well, 2003), the p-normalized Kendall tau distance (τ_p) (Fagin et al., 2004) in the interval $[0, 1]$ (the one used in (Baccianella et al., 2010)), and Kendall's tau-b (τ_b) (Kendall, 1945) (the one used in (Rothe et al., 2016)). Both τ_p and τ_b are suited to handle ties —which in our case means word senses with equal sentiment values— but they do so in different ways. For additional testing, in addition to the direct annotation values in the test set, we also use more fine-grained sentiment values of 278 entries that are available as part of the same gold standard (Rouces et al., 2018b), which were obtained using Best-Worst Scaling (BWS) and also comprised in the $[-1, 1]$ range. The reason for this is that these values are more fine-grained than the Direct Annotation (DA) values (which due to the use of 3 annotators, they range over only 7 possible values), and therefore ties are less common in the gold standard, making some ranking comparison algorithms more suitable. Since the BWS values were created only for the entries annotated as non-neutral by the DA scoring ($|\text{value}| \geq 0.5$), they cannot all be used for testing (or else the training set would be too biased towards neutral elements). Therefore, the intersection of the DA test set and the entries with BWS value is used for applying the rank-based measures.

- The measures based on discrete labels are the precision and recall values for each label, derived from the confusion matrix.

	DA										BWS
	ρ	τ_p	τ_b	precision	recall	acc.	confusion matrix				τ_b
							GS	SL			
								pos	neu	neg	
graph inheritance	0.39	0.39	0.38	pos: 0.28 neu: 0.91 neg: 0.33	pos: 0.26 neu: 0.90 neg: 0.42	0.82	pos neu neg	10 23 3	28 391 12	1 21 11	0.49
graph inheritance ext	0.33	0.42	0.32	pos: 0.22 neu: 0.90 neg: 0.27	pos: 0.21 neu: 0.89 neg: 0.35	0.81	pos neu neg	8 26 2	30 386 15	1 23 9	0.46
graph random paths	0.30	0.31	0.24	pos: 0.25 neu: 0.90 neg: 0.39	pos: 0.23 neu: 0.90 neg: 0.50	0.82	pos neu neg	9 26 1	29 390 12	1 19 13	0.46
word2vec +logit	0.47	0.21	0.38	pos: 0.37 neu: 0.93 neg: 0.46	pos: 0.54 neu: 0.88 neg: 0.52	0.84	pos neu neg	15 25 1	13 301 11	0 15 13	0.61
word2vec +svc /rbf	**0.55**	**0.15**	**0.45**	pos: 0.65 neu: 0.92 neg: 0.65	pos: 0.46 neu: 0.96 neg: 0.44	**0.89**	pos neu neg	13 7 0	15 328 14	0 6 11	**0.62**

Table 1: Results for evaluating the different methods for constructing the sentiment lexicon in Swedish. Note that the Kendall tau τ_p is a distance, and therefore it is inversely related to the Spearman correlation ρ. GS and SL stand for gold standard and sentiment lexicon respectively.

SentiWordNet is reported to have τ_p values of 0.281 and 0.231 for positive and negative dimensions (their sentiment model has 2 degrees of freedom). All our embeddings-based methods outperform both measures (τ_p is a distance, and therefore lower values are desired). (Rothe et al., 2016) reports $\tau_b = 0.654$. We obtain $\tau_b = 0.45$ when testing against the DA values, which is significantly lower. However, this probably owes to τ_b penalizing the big amount of ties in the DA values (61.95% of the possible pairs), as the method obtains $\tau_b = 0.63$ (a very close value) when testing against the BWS values, where ties are much less common (0.63%).

word sense ID	gloss	value	label
ond..4	'bad'	-0.9959	neg
farlig..1	'dangerous'	-0.9919	neg
villa..2	'illusion'	-0.9878	neg
kriminalitet..1	'criminality'	-0.9838	neg
skrämma..1	'frighten'	-0.9797	neg
fel..2	'wrong (a)'	-0.9757	neg
problem..1	'problem'	-0.9716	neg
misskreditera..1	'discredit'	-0.9675	neg
reaktionär..1	'reactionary'	-0.9635	neg
angrepp..1	'attack (n)'	-0.9594	neg
förfördela..1	'wrong (v)'	-0.9554	neg
brottslig..1	'criminal (a)'	-0.9513	neg
risk..1	'risk (n)'	-0.9473	neg
steka..2	'dismiss'	-0.9432	neg
absurd..1	'absurd'	-0.9391	neg
server..1	'server'	-0.0426	neu
ställe..1	'place (n)'	-0.0385	neu
förhållande..1	'relationship'	-0.0345	neu
markägare..1	'land owner'	-0.0304	neu
radio..1	'radio'	-0.0264	neu
sälja..1	'sell'	-0.0223	neu
offentlighet..1	'public (n)'	-0.0183	neu
manus..1	'manuscript'	-0.0142	neu
positiv..2	'positive (charge)'	-0.0101	neu
älvstrand..1	'riverside'	-0.0061	neu
molnet..1	'the cloud'	-0.0020	neu
flagga..2	'flag (v)'	0.0020	neu
reglera..2	'regulate'	0.0061	neu
resenär..1	'traveller'	0.0101	neu
läge..1	'position (n)'	0.0142	neu
inkomstskatt..1	'income tax'	0.0183	neu
kurator..1	'therapist'	0.0223	neu
land..2	'field, plot (n)'	0.0264	neu
distrikt..1	'district'	0.0304	neu
likartad..1	'similar'	0.0345	neu
fusion..3	'fusion (music)'	0.0385	neu
surdeg..1	'sourdough'	0.0426	neu
uppryckning..1	'improvement, recovery'	0.9635	pos
god..2	'tasty'	0.9675	pos
riktig..2	'genuine'	0.9716	pos
överraska..1	'surprise (v)'	0.9757	pos
hjälpa..1	'help (v)'	0.9797	pos
välsignelse..1	'blessing'	0.9838	pos
stöd..2	'support, aid (n)'	0.9878	pos
bra..3	'good'	0.9919	pos
äga..3	'rock, excel (v)'	0.9959	pos
fantastisk..1	'fantastic'	1.0000	pos

Table 2: Examples of sentiment values and labels.

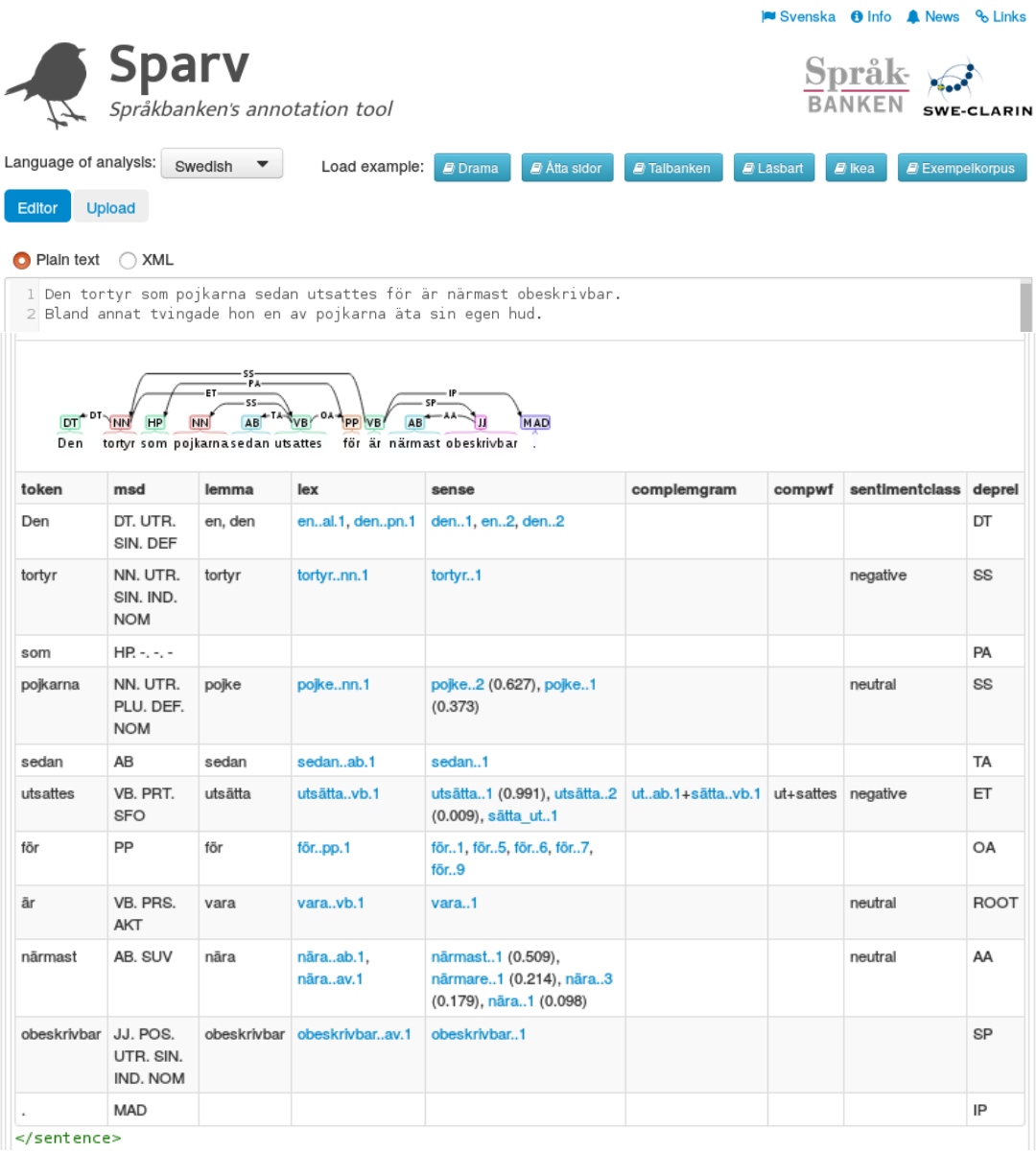
The method word2vec-svc-rbf performed consistently better than the rest, and therefore we have used it for the input to the manual curation step. Table 2 shows some examples of sentiment scores obtained using this method.

5 Step 3: Manual curation

In order both to get a better sense for the accuracy of the word2vec-svc-rbf method and in order to enhance the quality of the resulting dataset, this has been manually curated, as described in the following.

The outcome of the automatic sense-label assignment was a list of SALDO word senses labelled with a score in the interval $[-1, 1]$ assigned by the word2vec-svc-rbf method, and a sentiment label – one of -1, 0 or $(+)1$ – computed on the basis of the score. The resulting list contained 69,785 word senses, out of which 5,118 were labeled as non-neutral (3,508 negative and 1,610 positive items).

For the manual curation, we took all non-neutral items, plus the top 2,500 neutral items as determined



token	msd	lemma	lex	sense	complemgram	compwf	sentimentclass	deprel
Den	DT. UTR. SIN. DEF	en, den	en..al.1, den..pn.1	den..1, en..2, den..2				DT
tortyr	NN. UTR. SIN. IND. NOM	tortyr	tortyr..nn.1	tortyr..1			negative	SS
som	HP. -. -. -							PA
pojkarna	NN. UTR. PLU. DEF. NOM	pojke	pojke..nn.1	pojke..2 (0.627), pojke..1 (0.373)			neutral	SS
sedan	AB	sedan	sedan..ab.1	sedan..1				TA
utsattes	VB. PRT. SFO	utsätta	utsätta..vb.1	utsätta..1 (0.991), utsätta..2 (0.009), sätta_ut..1	ut..ab.1+sätta..vb.1	ut+sattes	negative	ET
för	PP	för	för..pp.1	för..1, för..5, för..6, för..7, för..9				OA
är	VB. PRS. AKT	vara	vara..vb.1	vara..1			neutral	ROOT
närmast	AB. SUV	nära	nära..ab.1, nära..av.1	närmast..1 (0.509), närmare..1 (0.214), nära..3 (0.179), nära..1 (0.098)			neutral	AA
obeskrivbar	JJ. POS. UTR. SIN. IND. NOM	obeskrivbar	obeskrivbar..av.1	obeskrivbar..1				SP
.	MAD							IP

Figure 3: Sentiment annotation in Sparv

by corpus frequency in the Gigaword Corpus (described in section 4.3 above). The curation consisted simply in checking the sentiment labels for all the 7,618 word senses in the resulting list, and correcting them if needed.

The resulting list has more neutral, and consequently less positive and negative items than the original: 2,640 neutral, 1,584 positive, and 3,394 negative items. A detailed analysis of the differences is still pending.

6 Summing up and looking ahead

We have described the development of SenSALDO, a Swedish sentiment lexicon containing 7,618 word senses as well as a full-form version of this lexicon containing 65,953 items (text word forms), for the SWE-CLARIN toolbox.[6]

Merely providing the downloadable lexicon is generally not sufficient for the user community targeted by CLARIN. For this reason, as a first step in the direction of more user-friendliness we have included

[6]The first version of this resource – SenSALDO v. 0.1 – is freely available for downloading under a CC-BY license from Språkbanken Text: `https://spraakbanken.gu.se/eng/resource/sensaldo`

sentiment annotation based on SenSALDO in Språkbanken Text's online annotation tool *Sparv*[7] (see figure 3) and the new document-oriented infrastructure component *Strix*, with the aim to provide document filtering based on sentiment (see figure 4).[8]

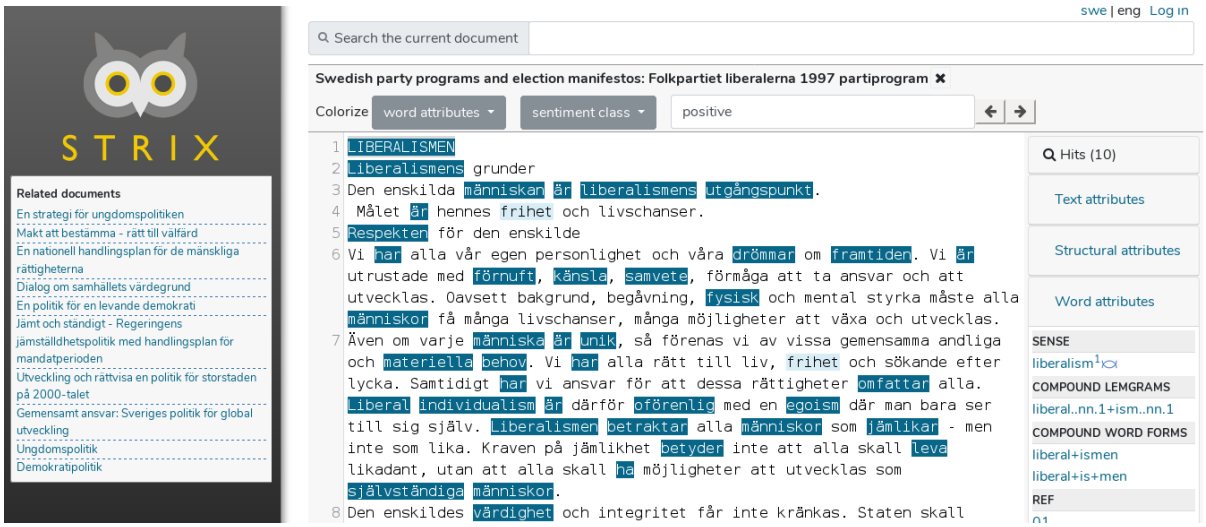


Figure 4: Sentiment annotation in Strix

In order to provide sentiment analysis as a standard tool in the SWE-CLARIN toolbox, we are currently pursuing two lines of development.

Firstly, there is a natural extension to the present version of SenSALDO, namely one where we ensure that all sentiment values for all lemmas present in it are accounted for. Because of the way SALDO is organized and because the lexical units considered in the work described here are *word senses*, there is no guarantee that this will be the case. For example, SenSALDO contains the information that the word sense `suga..2` carries a negative sentiment, which is correct (it means 'suck', as in *this situation sucks*). When generating the full-form version of SenSALDO, all forms of the lexeme *suga (v)* are given the sentiment label 'negative' (-1), but in fact the SALDO word sense `suga..1` 'suck (with mouth or instrument)' links to the same lexeme, and this word sense is arguably neutral wrt its sentiment value, so that SenSALDO ought to tell us that all forms of *suga (v)* occur with both negative and neutral sentiment. Complementing SenSALDO to reflect this is a straightforward enhancement which we are planning to implement in the next release of the resource.

Secondly, any sentiment-analysis software will require some means of evaluation, regardless of whether it is based on a sentiment lexicon or not. Pure machine-learning approaches to sentiment analysis rely on annotated training data. To meet both these needs, we are now in the final stages of preparing a Swedish gold-standard corpus for aspect-based sentiment analysis, consisting of approximately 1.5 million words from three different sources, two newspapers belonging to opposite ends of the political left–right spectrum, and an online discussion forum. The newspaper material consists of editorials and opinion pieces, and the topic for the whole corpus is immigration. This work is described in Rouces et al. (forthcoming).

An additional obviously very useful extension would be to add sentiment values to our diachronic lexicons (Borin and Forsberg, 2017), in order to support for instance historical research and studies in conceptual history such as that by Viklund and Borin (2016). This must remain a plan for the future, however.

Finally, we are in the process of using SenSALDO for developing both a sentence-level and an aspect-based sentiment analysis system for Swedish text, combining the polarity of terms according to syntax-based rules of compositionality. This will be complemented with information derived from annotated

[7] `https://spraakbanken.gu.se/sparv`
[8] `https://spraakbanken.gu.se/strix/`

corpora, which can cover cases that the lexicon-based approach cannot cover either due to limited coverage or non-compositional expressions.

With the already accomplished work presented above and the ongoing activities described in this section, we will soon be able to offer sentiment-analysis tools to Swedish researchers which are on a par with what is available for English.

Acknowledgements

This work has been supported by a framework grant (*Towards a knowledge-based culturomics*;[9] contract 2012-5738) as well as funding to Swedish CLARIN (*Swe-Clarin*;[10] contract 2013-2003), both awarded by the Swedish Research Council, and by infrastructure funding granted to Språkbanken by the University of Gothenburg.

References

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of LREC 2010*, pages 2200–2204.

R. Alexander Bentley, Alberto Acerbi, Paul Ormerod, and Vasileios Lampos. 2014. Books average previous decade of economic misery. *PLoS ONE*, 9(1):e83147.

Lars Borin and Markus Forsberg. 2017. A diachronic computational lexical resource for 800 years of Swedish. In Caroline Sporleder, Antal van den Bosch, and Kalliopi Zervanou, editors, *Language technology for cultural heritage*, pages 41–61. Springer, Berlin.

Lars Borin, Markus Forsberg, and Lennart Lönngren. 2013. SALDO: A touch of yin to WordNet's yang. *Language Resources and Evaluation*, 47(4):1191–1211.

Yanqing Chen and Steven Skiena. 2014. Building sentiment lexicons for all major languages. In *Proceedings of ACL 2014*, pages 383–389. ACL.

Ann Devitt and Khursid Ahmad. 2013. Is there a language of sentiment? An analysis of lexical resources for sentiment analysis. *Language Resources and Evaluation*, 47(4):475–511.

Johannes C. Eichstaedt, Hansen Andrew Schwartz, Margaret L. Kern, Gregory Park, Darwin R. Labarthe, Raina M. Merchant, Sneha Jha, Megha Agrawal, Lukasz A. Dziurzynski, Maarten Sap, Emily E. Weeg, Christopherand Larson, Lyle H. Ungar, and Martin E. P. Seligman. 2015. Psychological language on Twitter predicts county-level heart disease mortality. *Psychological Science*, 26(2):159–169.

Stian Rødven Eide, Nina Tahmasebi, and Lars Borin. 2016. The Swedish culturomics gigaword corpus: A one billion word Swedish reference dataset for NLP. In *Proceedings of the* From Digitization to Knowledge *workshop at DH 2016, Kraków*, pages 8–12, Linköping. LiUEP.

Andrea Esuli and Fabrizio Sebastiani. 2007. Random-walk models of term semantics: An application to opinion-related properties. *Proceedings of LTC 2007*, pages 221–225.

Ronald Fagin, Ravi Kumar, Mohammad Mahdian, D. Sivakumar, and Erik Vee. 2004. Comparing and aggregating rankings with ties. In *Proceedings of the Twenty-third ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '04, pages 47–58, New York, NY, USA. ACM.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, Mass.

William L. Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. 2016. Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of EMNLP 2016*, pages 595–605, Austin. ACL.

Richard Johansson and Luis Nieto Piña. 2015. Embedding a semantic network in a word space. In *Proceedings of NAACL-HLT 2015*, pages 1428–1433, Denver. ACL.

Viggo Kann and Magnus Rosell. 2005. Free construction of a free Swedish dictionary of synonyms. In *Proceedings of NODALIDA 2010*, Joensuu. University of Eastern Finland.

[9] https://spraakbanken.gu.se/eng/culturomics

[10] https://sweclarin.se/eng

Maurice G Kendall. 1945. The treatment of ties in ranking problems. *Biometrika*, pages 239–251.

Svetlana Kiritchenko and Saif M. Mohammad. 2016. Capturing reliable fine-grained sentiment associations by crowdsourcing and best–worst scaling. In *Proceedings of NAACL 2016*, pages 811–817, San Diego. ACL.

Stephen Kokoska and Daniel Zwillinger. 2000. *Standard Probability and Statistics Tables and Formulae*. Chapman & Hall / CRC.

Saif Mohammad and Peter Turney. 2010. Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, Los Angeles. ACL.

Jerome L Myers and A. (Arnold) Well. 2003. *Research Design and Statistical Analysis* . Mahwah, N.J. : Lawrence Erlbaum Associates, 2nd ed edition.

Bianka Nusko, Nina Tahmasebi, and Olof Mogren. 2016. Building a sentiment lexicon for Swedish. In *Proceedings of the* From Digitization to Knowledge *workshop at DH 2016, Kraków*, pages 32–37, Linköping. LiUEP.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2):1–135.

Magnus Rosell and Viggo Kann. 2010. Constructing a Swedish general purpose polarity lexicon: Random walks in the People's dictionary of synonyms. In *Proceedings of SLTC 2010*, pages 19–20, Stockholm. KTH.

Sascha Rothe, Sebastian Ebert, and Hinrich Schütze. 2016. Ultradense word embeddings by orthogonal transformation. *arXiv preprint arXiv:1602.07572*.

Jacobo Rouces, Lars Borin, Nina Tahmasebi, and Stian Rødven Eide. 2018a. Defining a gold standard for a Swedish sentiment lexicon: Towards higher-yield text mining in the digital humanities. In *Proceedings of DHN 2018*, pages 219–227, Aachen. CEUR-WS.org.

Jacobo Rouces, Nina Tahmasebi, Lars Borin, and Stian Rødven Eide. 2018b. Generating a gold standard for a Swedish sentiment lexicon. In *LREC 2018*, pages 2689–2694, Miyazaki. ELRA.

Jacobo Rouces, Lars Borin, and Nina Tahmasebi. forthcoming. Tracking attitudes towards immigration in Swedish media. In *Proceedings of DHN 2019*, Aachen. CEUR-WS.org.

Rachele Sprugnoli, Sara Tonelli, Alessandro Marchetti, and Giovanni Moretti. 2016. Towards sentiment analysis for historical texts. *Digital Scholarship in the Humanities*, 31(4):762–772.

Mike Thelwall. 2017. Sentiment analysis. In Luke Sloan and Anabel Quan-Haase, editors, *The SAGE Handbook of Social Media Research Methods*, pages 545–556. SAGE, London.

Jon Viklund and Lars Borin. 2016. How can big data help us study rhetorical history? In *Selected papers from the CLARIN annual conference 2015*, pages 79–93, Linköping. LiUEP.

Ting-Fan Wu, Chih-Jen Lin, and Ruby C Weng. 2004. Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, 5(Aug):975–1005.

Anis Yazidi, Hugo Lewi Hammer, Aleksander Bai, and Paal Engelstad. 2015. On enhancing the label propagation algorithm for sentiment analysis using active learning with an artificial oracle. In Leszek Rutkowski, Marcin Korytkowski, Rafal Scherer, Ryszard Tadeusiewicz, Lotfi A. Zadeh, and Jacek M. Zurada, editors, *Artificial Intelligence and Soft Computing*, pages 799–810, Cham. Springer International Publishing.

Using Apache Spark on Hadoop Clusters as Backend for WebLicht Processing Pipelines

Soheila Sahami
NLP Group
Leipzig University, Germany
sahami@informatik.uni-leipzig.de

Thomas Eckart
NLP Group
Leipzig University, Germany
teckart@informatik.uni-leipzig.de

Gerhard Heyer
NLP Group
Leipzig University, Germany
heyer@informatik.uni-leipzig.de

Abstract

Modern annotation tools and pipelines that support automatic text annotation and processing have become indispensable for many linguistic and NLP-driven applications. To simplify their active use and to relieve users from complex configuration tasks, Service-oriented architecture (SOA) based platforms – like CLARIN's WebLicht – have emerged. However, in many cases the current state of participating endpoints does not allow processing of "big data"-sized text material or the execution of many user tasks in parallel. A potential solution is the use of distributed computing frameworks as a backend for SOAs. These systems and their corresponding software architecture already support many of the features relevant for processing big data for large user groups. This submission describes such an implementation based on Apache Spark and outlines potential consequences for improved processing pipelines in federated research infrastructures.

1 Introduction

There are several approaches to make the variety of available linguistic applications – i.e. tools for preprocessing, annotation, and evaluation of text material – accessible and to allow their efficient use by researchers in a service-oriented environment. One of those, the WebLicht execution platform (Hinrichs et al., 2010), has gained significance – especially in the context of the CLARIN project – because of its easy-to-use interface and the advantages of not being confronted with complex tool installation and configuration procedures, or the need for powerful local hardware where processing and annotation tasks are executed.

The relevance of this general architecture can be seen when considering the increasing relevance of "cloud services" in the current research landscape (in projects like the European Open Science Cloud EOSC) and the rising number of alternative platforms. Comparable services like Google's *Cloud Natural Language*, *Amazon Comprehend*, *GATE Cloud* (Gate Cloud, 2018), or the completed *AnnoMarket* project are typically tight to some form of business model and show the significance – including a commercial one – of those applications. It has to be seen how a platform like WebLicht that is mostly driven by its participating research communities can compete with those offerings. However, some of the shortcomings that could be reasons to use alternative services may be reduced in the context of the CLARIN infrastructure as well. Potential problems may include the following areas:

- Support of processing large amount of text material (so called "big data") without loosing the mentioned benefits of a service-oriented architecture.
- Efficient use of parallelization, including the parallel processing of large document collections and the support of large user groups.

Soheila Sahami, Thomas Eckart and Gerhard Heyer 2019. Using Apache Spark on Hadoop Clusters as Back-end for WebLicht Processing Pipelines. *Selected papers from the CLARIN Annual Conference 2018*. Linköping Electronic Conference Proceedings 159: 188–195.

- Open accounting of used resources (ranging from used hardware resources to financial costs) for enhancing user acceptance of services and workflows by making hidden costs more transparent.

Using parallel computing approaches to improve the performance and workload on available hardware is a common topic in computer science. Several approaches have been established over time, including a variety of libraries, distributed computing frameworks, and dedicated computing hardware for different forms of parallelization. This submission proposes using the Apache Spark[1] framework on Hadoop clusters as backend for a WebLicht processing endpoint to address the aforementioned issues. A first prototypical implementation suggests the benefits of this approach.

The following two sections will describe the technical details of this demonstrator. In section 4 some of the general outcomes – like potential consequences for improving the performance, user satisfaction and clarity of the tasks – will be discussed and are followed by a brief summary of this contribution in section 5.

2 Synchronous Communication and Interaction in a SOA

Service-oriented architectures (SOAs) are an architectural approach that supports a group of services – as discrete units of functionality – to communicate with each other. The basic principles of SOAs are their independence of users, products and technologies. Some of SOA-based applications have several limitations such as the inability to develop highly interactive or completely customizable applications or lack of supporting synchronous interactions (Papazoglou, 2003), (Erl, 2005).

CLARIN's WebLicht[2] is a SOA-based processing environment that provides chains of language resources and tools (LRT) as distributed and independent services and covers a wide range of linguistic applications and several languages (Hinrichs et al., 2010). It facilitates the users' processes and relieves them from complicated configuration tasks. In our working prototype, the atomic feature of the implemented tools make them appropriate to integrate into WebLicht and other SOAs as well.

WebLicht and comparable environments are mostly based on synchronous communication and interaction between users and the framework. The typical WebLicht workflow relies on users issuing – simple or combined – tasks and waiting for the final results for their export or further analysis. In general, this assumes synchronous interactions which are hardly feasible for big data analysis that may require execution times of several days or more.

However, WebLicht's popular Web-interface does not provide any adequate handling of long-term processes, processing of document collections or (very) large input data in general. Although it should be noted that with *WebLicht as a Service*, a first alternative approach exists for WebLicht, that mitigates some of these problems[3].

3 Technical Approach

In this section we discuss the techniques utilized in our implementation including Apache Hadoop and Apache Spark and their strengths in comparison with other techniques. We also describe the implemented linguistic applications and compare their efficiency with similar tools that use different frameworks.

3.1 Apache Hadoop

Apache Hadoop is a framework to process large-scale data in a distributed computing environment and is a popular framework for applications that deal with massive data and where the response time is significant. Its large ecosystem consists of the Hadoop Kernel, MapReduce, HDFS and some other components such as YARN, Apache Hive and Zookeeper (Apache Hadoop, 2019).

Apache Hadoop has a fault-tolerant distributed storage system called Hadoop Distributed File System (HDFS) that supports storage of massive amounts of data. HDFS is highly fault-tolerant, suitable for applications that support large data sets, easily deployable on low-cost hardware and provides high throughput access to data. It scales up incrementally and is able to handle the failure of

[1] https://spark.apache.org/
[2] https://weblicht.sfs.uni-tuebingen.de
[3] https://weblicht.sfs.uni-tuebingen.de/WaaS/

storage infrastructure without losing data by storing three complete copies of each block of data redundantly on three different servers (Bhosale and Gadekar, 2014).

MapReduce (Dean and Ghemawat, 2004) – which is the processing pillar in the Hadoop ecosystem – is a programming model and associated implementation that allows processing of data up to the size of multiple terabyte in parallel on large clusters (with thousands of nodes) in a reliable, fault-tolerant way. A MapReduce job divides input data into independent blocks. Users define the *Map* functions that process these data blocks in parallel and generate intermediate results as set of key-value pairs. The key-value based approach is utilized for the challenging task of combining processed data in a distributed environment. *Reduce* functions merge the intermediate values based on the same intermediate keys to produce the output. Input and output of the functions are stored in HDFS; the framework supervises and monitors scheduled tasks, and re-executes failed tasks (Bhosale and Gadekar, 2014), (Apache Hadoop, 2019).

Hadoop has some particular properties that make it more eligible (White, 2012). For instance, it can increase access speeds – the rate at which data can be read or written from or to drives – by reading and writing data from and to multiple disks in parallel instead of serial accesses and allows shorter analysis times by parallel execution of tasks. Furthermore, replication and redundant copies of data is a central component of Hadoop to avoid data loss in the event of hardware failure.

In comparison with relational database management systems (RDBMS) Hadoop is more efficient in many respects. RDBMS are able to do large-scale batch analysis on several disks but there are several problematic issues resulting from the used architecture. For example, updating a large portion of data – which often comes with sort or merge operations – is less efficient than using MapReduce. Furthermore, seek time latency – the time required to move the disk's head to a particular place on the disk to read or write data – when processing big data is considerable, especially in unstructured resources like texts. The approach of data access patterns in MapReduce – read, process and write the entire data set at once – decreases this latency.

In High Performance Computing (HPC) and Grid Computing platforms, large-scale data processing is executed using Message Passing Interface (MPI) which is distributing tasks across a cluster of machines and utilizing a shared file system. However, accessing very large data volumes (up to several terabytes) makes network bandwidth a potential bottleneck that can lead to idle computing nodes. MapReduce uses the data locality feature: data is collected by corresponding computing nodes; data is stored locally which improves access times (White, 2012).

3.2 Apache Spark

Apache Spark is also a general-purpose cluster computing framework for big data analysis with an advanced in-memory programming model and upper-level libraries and APIs. It uses a multi-threaded model where splitting tasks on several executors improves processing times and fault tolerance.

Compared to MapReduce, Apache Spark uses a data-sharing abstraction called Resilient Distributed Dataset (RDD); individual operations (i.e. Map and Reduce) are similar. RDDs are In-Memory Databases (IMDB) which are designed to run completely in RAM. Using this extension, Apache Spark is able to perform processing workloads easier and more efficient and provides large speedups. RDDs are transient: every time they are used they are recomputed. If data is used more often, users can persist individual RDDs in memory for a faster reuse.

One of the key properties of RDDs is their design as fault-tolerant collections, which are capable to recover lost data after a failure occurs and can be processed and manipulated in parallel. In other distributed computing frameworks, fault tolerance is achieved by data replication or check pointing while Spark uses a different approach called lineage. When building an RDD, the graph of used transformations is kept and if any failure occurs, the operations are re-run to rebuild lost results. As the RDDs are stored in memory, rewriting recovered data is faster than writing operations over the network. Thus, lineage-based recovery can save both execution time and storage space ((Zaharia et al., 2016), (Hamstra and Zaharia, 2013), (Salloum et al., 2016)).

3.3 Scalability

In parallel computing, beside the performance that increases by distributing executions over several processing cores, the efficiency and scalability of applications are also desirable attributes.

Scalability relates to the capability of a process to deal with a growing amount of data. As one of the performance metrics for parallel implementation is the runtime, the scalability is measured by the speedup that is the ratio of runtime using one executor to n executors. Efficiency is calculated by the speedup divided by the number of executors. In the ideal case, the highest value for efficiency shows a linear growing speedup. In reality and for typical use cases, a sub-linear speedup also shows an improvement in efficiency ((Hill, 1990), (Bondi, 2000)).

3.4 Tools

In the context of our implementation, a variety of typical NLP tools – including sentence segmentation, pattern-based text cleaning, tokenizing, language identification, and named entity recognition – were implemented[4]. These tools use Hadoop as their framework, Apache Spark as execution engine and store the input data and outputs on HDFS. The tools are atomic services that have the potential to be integrated in any SOA-based annotation environment.

In order to execute these tools, we have used a cluster provided by Leipzig University Computing Center (Meyer et al., 2018). Table 1 illustrates hardware and configuration of this cluster (Lars-Peter Meyer, 2018).

Number of nodes	CPUs	Hard drives	RAM	Network
90	6 cores per node	>2 PB in total	128 GB per node	10 Gbit/s Ethernet

Table 1: Cluster Characteristics

During the execution, as the first step, input text files are read from HDFS and loaded in RDDs. Those RDDs are distributed over the allocated cluster hardware and are processed by several executors in parallel. The results are provided by merging processed RDDs and the finalized outputs again are stored on HDFS.

Optimum hardware configuration for each job can be set dynamically considering volume and type of input data as well as the selected processing pipeline which may consist of a single or even multiple tools. The specific configuration is determined automatically based on empirical values taken from previous runs and takes the current workload of the underlying cluster into account.

For the subset of these tasks that is supported by WebLicht's text corpus format (TCF) (Heid et al., 2010) (i.e. tokenization and sentence segmentation) converters between TCF and the RDDs were written. As a result, the endpoint is structured as depicted in Figure 1.

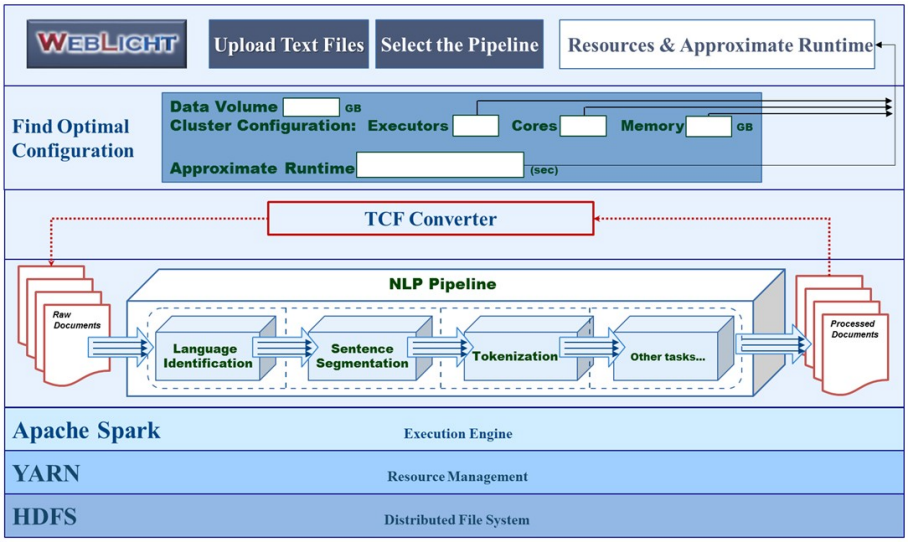


Figure 1: WebLicht with a Spark-based Backend.

[4] http://hdl.handle.net/11022/0000-0007-CA50-B

For the evaluation of the implemented solution benchmarks were carried out. The benchmarks were executed to show the impact of parallelization for every task. The diagrams in Figures 2 and 3 show runtimes for various data volumes with comparable characteristics using different cluster configurations. They illustrate the effect of configuration variables on concrete process runtimes and especially the impact of parallelization (i.e. the number of executors). Using these results, for every batch of input data a cluster configuration can be estimated that constitutes an acceptable trade-off between allocated resources and the expected runtime.

4 Results

The implemented tools already provide added value to the WebLicht platform. In this section, we will describe some of the more general outcomes.

4.1 Performance Improvements

The central advantage of the described implementation is its ability to use state-of-the-art cluster computation technology to process input material in massive scale. Parallelization in general has the potential to achieve major performance improvements; using Apache Hadoop and Spark – on the basis of large scale clusters – is a promising approach in this area. As a result, we were able to process huge amounts of data with a significant decrease in resources compared with sequential approaches.

Table 2 compares the runtimes for processing 6.5 GB input data for the subset of tools that can be currently integrated into WebLicht, using our parallel implementation and an existing sequential application. Both tests used the same hardware configuration with 8 GB memory and one executor. The measured runtimes illustrate that parallelization using Apache Spark can generate results in significantly less amount of time when the same input texts were processed using equal resources (memory and CPU).

	Segmenting	Tokenizing
Existing Sequential Implementation	8,200 sec	19,860 sec
Parallel Implementation	357 sec	781 sec

Table 2: Runtime for processing 6.5 GB input text using sequential and parallel implementations

The diagrams in Figure 2 and 3 show the scalability of the tools. As expected, by increasing the number of executors, runtime declines and efficiency improves. It is worth mentioning, that the number of executors cannot grow infinitely and exceeding available resources will lead to negative effects on the efficiency.

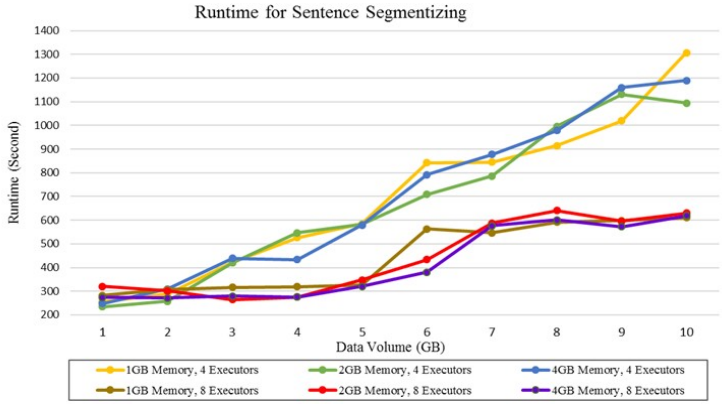


Figure 2: Segmentation 1 to 10 GB Text Data using 4 or 8 Executors and 1, 2 or 4 GB RAM.

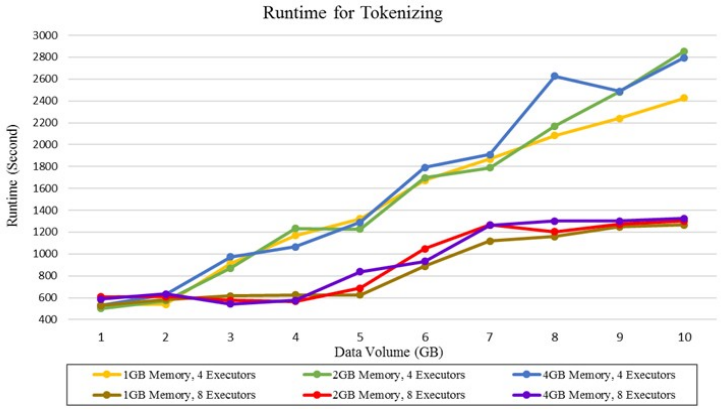


Figure 3: Tokenizing 1 to 10 GB Text Data using 4 or 8 Executors and 1, 2 or 4 GB RAM.

It is also noteworthy that the total runtime for the pipeline including sentence segmentation and tokenizing as a combined task depends on the most time-consuming sub-task – tokenizing in this experiment – regardless of the volume of documents and the allocated resources. That is because combined tasks save I/O time for writing the segmented documents which is the main part of the overall sentence segmentation runtime.

4.2 Benefits of Asynchronous Workflows

The aforementioned focus on synchronous communication patterns is problematic when dealing with very large input data and resulting growing execution times of scientific workflows. In those cases, a direct interaction with the user can not be expected and alternatives have to be evaluated. One of the possible approaches is the automated execution of pipelines as it is supported by *WebLicht as a Service*. However, this already requires some deeper technical knowledge from the end user and contradicts partially the seamless integration of pipeline results in a federated Web-based infrastructure.

For a systematic support of big data processing in the context of WebLicht pipelines, changes in the default workflows and user interfaces might be helpful. This may comprise an improved support for the processing of document collections – in contrast to a more document-centric approach – and a stronger focus on data storage platforms that support workspaces for individual users like B2DROP. This infrastructure component has the ability to function both as a primary means of storage for individual users and user groups, but also as a central entry point to start new workflows in a data-oriented environment. Its function would include both being a potential provider of input data for WebLicht but also as the default storage space of intermediate and final results. User information about status and outcome of scheduled processing jobs can be transfered via Email or job-specific status pages[5]. Those status reports should be seen as an important means to inform users about used hardware resources, required runtimes, and relevant process variables. For increasing user acceptance of the overall system, they may also contain information about required financial resources that would have been necessary to perform the same task using a commercial platform.

4.3 Technical Costs for Accountability and User Motivation

Resulting execution times for a specific configuration (i.e. number of used executors, allocated memory, etc.) are valuable information for estimating requirements and runtime behavior of every task. Based on empirical data, runtimes for new tasks can be estimated considering their general characteristics (i.e. size of input data and used tool configuration). This estimate provides several added values to our system that are briefly described in the following.

[5] A functionality that is already supported by other comparable frameworks.

- **Balance between resources and users satisfaction**: This valuable information helps to find an optimal balance between number of parallel user tasks, available hardware configuration, and waiting times that are still acceptable for the users.
- **Processing time**: It provides the users the approximate process time for the requested services, which may help to increase their willingness to wait for results.
- **Parallel users**: Available resources and estimated runtimes for requested tool chains can be used to define the number of parallel users on the system dynamically.
- **Financial costs**: Using services (resources and tools) has financial costs in commercial NLP platforms like Amazon Comprehend[6] and Google Cloud NLP[7]. Prices are defined based on the chosen service and size of the data (like number of characters). Making these potential costs more visible can encourage users to carefully select a proper configuration. This information – next to required resources and run times – can also be taken as a basis to calculate the technical cost for each tool chain individually.

Having this information, the tool chain is not a black box anymore and this transparency can increase the user's motivation and satisfaction.

As an instance, results of sentence segmentation a 5 GB document using 8 executors with 8 GB memory will be ready after 10 minutes and using 2 executors with 4 GB memory after 25 minutes. Administrator of the system can suggest different configuration considering the available resources and the user has the options to decide based on the waiting time and probable technical cost.

5 Summary

In this working prototype, we have presented an alternative approach to use NLP tools for processing large text data and handling large user groups using parallelization in a cluster environment. Using this approach, besides simple runtime improvements – processing more data using affordable resources and less runtime –, several results were achieved.

Dynamic resources configuration was introduced for each individual NLP task that constitutes an acceptable trade-off between allocated resources and expected runtime. The capability of open accounting of required resources – ranging from used hardware resources to financial costs – to make hidden costs more transparent is seen by the authors as a central means to enhance user acceptance of services in an environment like WebLicht, or CLARIN in general. Furthermore, it was outlined how a stronger focus on asynchronous communication in a federated research environment has the potential for seamless integration even in the case of long-term annotation processes or the processing of big data resources.

Acknowledgement

Computations for this work were done with resources of Leipzig University Computing Center.

References

[Apache Hadoop2019] Apache Hadoop. 2019. Apache Hadoop Documentation. Online. Date Accessed: 11 Jan 2019. URL http://hadoop.apache.org/.

[Bhosale and Gadekar2014] Harshawardhan S Bhosale and Devendra P Gadekar. 2014. A review paper on big data and hadoop. International Journal of Scientific and Research Publications, 4(10):1–7.

[Bondi2000] André B Bondi. 2000. Characteristics of scalability and their impact on performance. In Proceedings of the 2nd international workshop on Software and performance, pages 195–203. ACM.

[Dean and Ghemawat2004] Jeffrey Dean and Sanjay Ghemawat. 2004. Mapreduce: Simplified data processing on large clusters. In OSDI'04: Sixth Symposium on Operating System Design and Implementation, pages 137–150, San Francisco, CA.

[Erl2005] Thomas Erl. 2005. Service-oriented architecture: Concepts, Technology, and Design. Prentice Hall PTR.

[6] https://aws.amazon.com/comprehend/pricing
[7] https://cloud.google.com/natural-language/pricing

[Gate Cloud2018] Gate Cloud. 2018. GATE Cloud: Text Analytics in the Cloud. Online. Date Accessed: 11 Apr 2018. URL https://cloud.gate.ac.uk/.

[Hamstra and Zaharia2013] Mark Hamstra and Matei Zaharia. 2013. Learning Spark: lightning-fast big data analytics. O'Reilly & Associates.

[Heid et al.2010] Ulrich Heid, Helmut Schmid, Kerstin Eckart, and Erhard W Hinrichs. 2010. A Corpus Representation Format for Linguistic Web Services: The D-SPIN Text Corpus Format and its Relationship with ISO Standards. In Proceedings of LREC 2010.

[Hill1990] Mark D Hill. 1990. What is scalability? ACM SIGARCH Computer Architecture News, 18(4):18–21.

[Hinrichs et al.2010] Erhard Hinrichs, Marie Hinrichs, and Thomas Zastrow. 2010. WebLicht: Web-based LRT services for German. In Proceedings of the ACL 2010 System Demonstrations, pages 25–29. Association for Computational Linguistics.

[Lars-Peter Meyer2018] Lars-Peter Meyer. 2018. The Galaxy Cluster. Online. Date Accessed: 12 Apr 2018. URL https://www.scads.de/de/aktuelles/blog/264-big-data-cluster-in-shared-nothingarchitecture-in-leipzig.

[Meyer et al.2018] Lars-Peter Meyer, Jan Frenzel, Eric Peukert, René Jäkel, and Stefan Kühne. 2018. Big data services. In Service Engineering, pages 63–77. Springer.

[Papazoglou2003] Mike P Papazoglou. 2003. Service-oriented computing: Concepts, characteristics and directions. In Web Information Systems Engineering, 2003. WISE 2003. Proceedings of the Fourth International Conference on, pages 3–12. IEEE.

[Salloum et al.2016] Salman Salloum, Ruslan Dautov, Xiaojun Chen, Patrick Xiaogang Peng, and Joshua Zhexue Huang. 2016. Big data analytics on Apache Spark. International Journal of Data Science and Analytics, 1(3-4):145–164.

[White2012] Tom White. 2012. Hadoop: The definitive guide. O'Reilly Media, Inc.

[Zaharia et al.2016] Matei Zaharia, Reynold S Xin, Patrick Wendell, Tathagata Das, Michael Armbrust, Ankur Dave, Xiangrui Meng, Josh Rosen, Shivaram Venkataraman, Michael J Franklin, et al. 2016. Apache spark: a unified engine for big data processing. Communications of the ACM, 59(11):56–65.

Bulgarian Language Technology for Digital Humanities: a focus on the Culture of Giving for Education[1]

Kiril Simov
IICT-BAS
Sofia, Bulgaria
kivs@bultreebank.org

Petya Osenova
IICT-BAS
Sofia, Bulgaria
petya@bultreebank.org

Abstract

The paper presents the main language technology components that are necessary for supporting the investigations within the digital humanities with a focus on the culture of giving for education. This domain is socially significant and covers various historical periods. It also takes into consideration the social position of the givers, their gender and the type of the giving act (last posthumous will or financial support in one's lifetime). The survey describes the adaptation of the NLP tools to the task as well as the various ways for improving the targeted extraction from the specially designed corpus of texts related to giving. The main challenge was the language variety caused by the big time span of the texts (80-100 years). We provided two initial instruments for targeted information extraction: statistics with ranked word occurrences and content analysis. Even in this preliminary stage the provided technology proved out to be very useful for our colleagues in sociology, cultural and educational studies.

1 Introduction

Language technology can help in the extraction of useful and focused content from domain texts. We have already worked on a number of such tasks related to Digital Humanities. For example, in the eLearning area (enriching learning objects content or positioning the learner against a predefined level of expected knowledge) – see in Monachesi et al., 2006; in iconography (describing the icons with the help of an ontology for a better comparison and typology) – see in Staykova et al., 2011, etc. Such projects are reflected in the creation of our language resources and technologies during the years – see Zhikov et al., 2013, Savkov et al., 2012, and Simov et al., 2004.

In this paper we focus on the culture of giving for education. Our work was part of the national project (2015-2017) entitled *Culture of giving in the sphere of education: social, institutional and personality dimensions*, coordinated by the Institute for the Study of Societies and Knowledge at Bulgarian Academy of Sciences. Our sociology colleagues adopted two main approaches in their survey: a) application of software developed especially for the content analysis of historical documents; b) application of the theory of planned behavior to the study of philanthropy. Our work was part of the former approach but with influence on the latter.

The collected corpus comprises texts with a time span of 80-100 years. The task was to extract relevant information with the help of statistics and content analysis for displaying the tendencies in the area of giving from the perspective of the language/phrasing/terminology, the social and economic

Kiril Simov and Petya Osenova 2019. Bulgarian Language Technology for Digital Humanities: a focus on the Culture of Giving for Education. *Selected papers from the CLARIN Annual Conference 2018*. Linköping Electronic Conference Proceedings 159: 196–204.

context. Thus, the initial steps include: the creation of a specialized corpus, the creation of a web-based concordance tool and presenting useful statistics and content analysis over the corpus, and adaptation of the existing basic NLP tools.

Our work aims towards the ideas presented in Fokkens et al., 2018. The similarities are as follows: we similarly aim at receiving structured data as output from the NLP processing; we encode metadata characteristics that are common for the corpus (birth date, death date, place of birth, gender, names, etc.); we provide help for getting information on various thematic questions, such as the terminology change through the periods, the target groups preferences, the social behaviour of the givers, etc. The differences are as follows: we do not work with digitized biography dictionaries, but with a specialized corpus of givers' wills that include biographical information; we have not progressed yet to cover also prosopographical information, i.e. to measure characteristics of well-defined groups. However, we envisage this task as our future direction. Last, but not least, the NLP pipeline that was used here does not incorporate any Wordnet concepts or semantic roles despite the fact that we have a word sense disambiguation module for processing newspaper data. These modules will be added later for such specific tasks as the one reported in the paper.

The structure of the paper is as follows: section 2 describes the corpus and its processing; section 3 focuses on the statistics and content analysis; section 4 outlines our efforts in linking the named entities in the corpus – people, locations, organizations; section 5 concludes the paper.

2 Corpus and processing

The specialized corpus of giving for education (abbreviated as CoDar) consists of separate documents from the period after the liberation of Bulgaria (from 1878 onward) until the middle of XX century. Since the aim of the sociologists is to investigate the incentives behind the decision to support education as well as the attitude of the donors together with the most significant causes, the resource includes last will documents, various acts of giving – letters, notarized acts of giving, constitutive documents of charity funds and foundations.

The texts have been gathered from various libraries and then – scanned and digitized. They were represented in an XML format. The following types of information were added: metadata, structural and linguistic ones. The *metadata* provides information about: the title of the document and its type (last will, document of giving, etc.), the place and the time of the document emergence; the gender and the social status of the donor/donors. The *structural information* provides the text, divided into paragraphs and sentences. The *linguistic information* provides parts-of-speech, morphosyntactic characteristics and dependency syntactic analysis.

The NLP modules for Bulgarian that have been adapted to the specificities of the corpus are as follows: a tokenizer, a morphological analyzer, a Named Entities recognition and linking module, a lemmatizer and a parser (Savkov et al. 2012). Our state-of-the-art morphosyntactic tagging reaches 97.98 % accuracy (Georgiev et al. 2012). The lemmatization module that depends on the tagging has 95 % accuracy (Savkov et al. 2012), and the dependency parser - 91 % (Simova et al. 2014). Our tagset is rich - it comprises 680 tags[2].

When applied to the specific data, the main problems in the tokenization were related to the proper handling of the abbreviations, especially of titles, named entities, temporal expressions, etc. The morphological analyzer, which is a combination of a morphological dictionary and statistical components, had as its main challenges: rare or archaic words and different orthographical codifications. The lemmatizer depends on the results from the morphological analyzer. We rely on the Bulgarian inflectional lexicon containing near 110 000 lemmas (Popov et al. 1998 and Popov et al. 2003) which is used to map each word form to the lemma given that the grammatical features are predicted correctly. This mapping is almost 100 % accurate. Thus, the main difficulty was the assignment of the word form of a rare word to its lemma. The parser also depends on the previous

[2] http://bultreebank.org/wp-content/uploads/2017/04/BTB-TR03.pdf

steps. Apart from that, the parser had problems with the syntactically different codifications in the contemporary Bulgarian and the texts related to the past times. At this first stage, the parser was used mainly as a source of extracting key phrases. Thus the errors were not viewed as crucial.

The morphological tagging step showed the following situation for the three historical periods (see next section for details on the periodization). The qualitative analysis displayed similar problems for all periods: recognition of dates, names, abbreviations; usage of old case forms, old orthographic forms (including malapropisms, such as '*мие*' (wash) instead of '*ми е*' (to me); '*са*' (are) instead of '*се*' (refl)). In Period 2 in addition come the problematic cases with a wrong gender and a wrong part-of-speech. In Period 3 there appear the wrongly analysed name parts. The quantitative analysis in Table 1 showed the following error rates:

Period	**Error Rate**
Period 1	3.3 %
Period 2	7 %
Period 3	12 %

Table 1: Error rates of morphosyntactic tagging.

At first sight it might seem strange that the error rate of the oldest period is lower than the error rate of the most recent period. It should be vice versa, since the language of the third period is closer to nowadays language. Here, however, the role of the corpus sources has to be taken into account. Many of the documents representing the first and the second period came in a normalized way because they were published in books before the start of the project, while many of the texts from the third period were in their authentic form – they were scanned from archive documents in several large cities in Bulgaria.

3 Statistics and Content Analysis

Two of the most effective ways for observing the behaviour of various words, collocations and phrases are: (a) the statistics over the keywords in some domain and (b) their context-expanding concordances.

On the basis of a frequency analysis over the specialized corpus of giving (CoDar) in comparison to the Bulgarian National Reference Corpus[3] (being a general corpus), frequency lists have been produced for three historical periods: *before 1919* (the Bulgarian Renaissance and the end of the First World War) – 49698 word forms; *between 1919 and 1930* (the period of crisis after the First World War) – 46031 word forms, and *after 1930 to early 1940s* (the years of stability, the Second World War and the first years after 09.09.1944) – 66373 word forms. The content analysis of the corpus showed that a) during all the three mentioned periods the acts of giving aimed at raising the education among the Bulgarian population and it targeted mainly students with modest financial possibilities; b) the texts content reflects the influence of the historical development in Bulgaria during the three periods on the campaigns of giving; c) the orthographic and grammatical style follows the norms that held in the respective period.

For extracting the key words from the corpus we used the program *AntConc* (see Anthony, 2014). We compared the lemmatized version of CoDar corpus with a frequency list constructed on the basis of the lemmatized version of the Bulgarian National Reference Corpus. In this way a list of keywords for each subcorpus was created. The visualization (in form of a word cloud) has been done through the

[3] http://webclark.org/

following web service: https://www.jasondavies.com/wordcloud/. The 'cloud' representations of words for the three periods are outlined on Figures 1, 2 and 3 below.

From the visualizations it can be seen that in the period before 1919 the important keywords are: '*фонд*' (fund), '*училище*' (school), '*ефория*' (board), '*завещание*' (will), '*завет*' (legacy), '*книжа*' (securities), '*лев*' (lev – Bulgarian currency), '*пари*' (money), '*капитал*' (capital); '*София*' (Sofia), '*Букурещ*' (Bucharest). In the period between 1919 and 1930 the following keywords appear: '*банка*' (bank), '*имот*' (estate), '*сума*' (sum), '*съвет*' (council), '*учител*' (teacher); '*Габрово*' (Gabrovo). After 1930 the following words show high frequency: '*гимназия*' (high school), '*година*' (year), '*народен*' (national), '*беден*' (poor). In general, however, the keywords in the three periods are semantically coherent. They belong to the domains of education and finances.

Fig. 1: Keywords for the period *before 1919*.

Fig. 2: Keywords for the period *between 1919 and 1930*.

Fig. 3: Keywords for the period *after 1930*.

In Table 2 below we give the first ten most frequent words from the lists with ranked keywords for the three periods.

While in the first period the concept of *will* dominates, in the second and third one this is the concept of *fund*. As third in the ranking list during the first period comes the Bulgarian word '*училище*' (school), while in the next two periods it is the Bulgarian word '*сума*' (sum). In the first and second periods the fourth place has been taken by the same concept *board*, but with two lexicalizations. In the third period the word is *the secondary school*. In case of the board concept the terminology change can be traced.

Ranking of keywords for the three periods					
Before 1919		*Between 1919 and 1930*		*After 1930*	
Word	**Rank**	**Word**	**Rank**	**Word**	**Rank**
завещание (will)	7.87	*фонд* (fund)	6.42	*фонд* (fund)	7.12
фонд (fund)	4.22	*завещание* (will)	5.73	*завещание* (will)	5.85
училище (school)	3.42	*сума* (sum)	3.67	*сума* (sum)	3.69
ефория (board of trustees)	2.71	*настоятелство* (board of trustees)	3.40	*гимназия* (secondary school)	2.78
имот (property)	2.23	*беден* (poor)	3.11	*беден* (poor)	2.60
сума (sum)	2.19	*училище* (school)	2.43	*просвещение* (education)	2.54
лихва (interest)	2.14	*завещавам* (leave one's will)	2.40	*лихва* (interest)	2.51
МНП (Ministry of national education)	2.14	*лихва* (interest)	2.35	*гимназията* (the secondary school)	2.07
душеприказчици (confessors)	1.93	*гимназия* (secondary school)	1.69	*дарение* (donation)	2.06
завещавам (leave one's will)	1.76	*дарение* (donation)	1.66	*завещавам* (leave one's will)	1.80

Table 2: The first 10 words with the highest rank, presented per period.

The lexicalization of the concept *board* '*ефория*' (ephoria) changes as follows: in the first period it takes fourth position, in the second period – 280th position and in the third period – 602nd position. On the contrary, the other lexicalization '*настоятелство*' comes as 426th in the first period, as 4th in the second period and as 14th – in the third period. In all the periods within the first 10 words there appears the verb '*завещавам*' (leave by will). Thus, as a whole, mainly the terminology has changed, not the content.

Of special interest are the keywords that appear in one or two periods, but not in all three. Such an example is the adjective *беден* (poor). It the second and third periods it takes fifth position of frequency, but it does not appear within the first 10 most frequent words in the first period. Thus, an assumption might be made that the giving after 1919 was oriented exclusively to the poor students. In the period before 1919 also the word '*имот*' (property) has been used frequently. This means that giving through estates was a charity form that was not so popular in the periods after 1919. The data might give hints on the distribution of roles within society. For example, before 1919 the roles of the members of the boards of trustees as well as the executors were more popular, while in the next two periods the role of the legator became the most frequent one.

The concordancing service has been customized on the base of the *webclark.org* concordancer. Several use cases have been tested, such as: finding information about female donors or executors of wills (we got 20 results); finding information about the beneficiaries of the donors (we got around 56 results); finding cases on what the support has been given for (we got 70 results where the preferences concern the schools and then – some specific persons).

Concerning the first use case, the names of the donors can be derived through the keyword *donor* or through the metadata search, or both. It is interesting to get the information about the female donors. Besides their names, there is information about their native towns (Gabrovo, Plovdiv, Tarnovo, Gorna Oryahovitsa, etc.), their birth and death dates in case they are known. If not known, this position is marked as *unknown*. Additionally, the type of giving act is mentioned: last will, charity, letter of interest, etc. It turns out that the charity documents are used more frequently in comparison to last will ones and letters of interest.

Concerning the second use case, the documents include information about the circumstances under which the beneficiary might be deprived from its grant. Such circumstances refer to cases when the student does not behave or gets low grades. Some pre-conditions for the grants might be declared in advance, such as the students to return in Bulgaria after their studies or to work in an appointed area (in education or in church, etc.).

Concerning the third case, the frequent situation is when the main support goes to the primary and secondary schools but the sum interest goes for the living expenses of poor and/or blind children who are hard working and capable in their studies.

Since in the giving act also the place of giving is important, Table 2 shows a list of ranked places that appear within the top 150 selected keywords.

Before 1919		*Between 1919 and 1930*		*After 1930*	
Name	**Rank**	**Name**	**Rank**	**Name**	**Rank**
Букурещ (Bucharest)	1.64	*Чепеларе* (Chepelare)	0.84	*Копривщица* (Koprivshtitsa)	1.45
Търново (Tarnovo)	1.06	*Габрово* (Gabrovo)	0.76	*София* (Sofia)	0.89
Габрово (Gabrovo)	0.89	*София* (Sofia)	0.71	*Шумен* (Shumen)	0.64
Свищов (Svishtov)	0.61	*Неврокоп* (Nevrokop)	0.48	*Търново* (Tarnovo)	0.55
Галац (Galats)	0.49	*Търново* (Tarnovo)	0.46	*Габрово* (Gabrovo)	0.34
Одеса (Odessa)	0.44	*Прилеп* (Prilep)	0.39	*Севлиево* (Sevlievo)	0.31
Браила (Braila)	0.39	*Севлиево* (Sevlievo)	0.30	*Пазарджик* (Pazardzik)	0.30

София (Sofia)	0.38			*Етрополе* (Etropole)	0.21
Карлово (Karlovo)	0.29				

Table 3: The words in first 9 positions (where applicable) with the highest rank, presented per period.

It can be observed that in the first period there appear more cities outside Bulgaria, such as Bucharest, Galats and Braila in Romania and Odessa in Russia. In the second period there appear also cities from South-West Bulgaria and Vardar Macedonia like Nevrokop and Prilep. In the third period the focus is only on cities that are within the boundaries of nowadays Republic of Bulgaria. Another interesting thread is the role of today's capital Sofia. In the first period its rank is 0.38. In the second period it grows to 0.71, while in the third one it is already 0.89. Also, after 1930s the most important towns happen to be the smaller province ones such as Koprivshtitsa and Etropole.

4 Named Entity Annotation and Linked Open Data

All the Named Entities have been annotated in XML format with respect to their categories: Person, Location, Organization, Date, Amount. In this way the actual charity documents have been connected to the biographies of the donors. Thus, we established a connection between the events within donors' biographies and the overall acts of giving. This information will be used in at least two ways: (1) the creation of Linked Open Data datasets interconnected with the existing datasets like DBpedia, GeoNames, etc; and (2) support for better understanding of the culture of giving, motivation for donation, etc.

For each document we manually explicated all the mentions of persons . The metadata includes: the names of the persons who donated the sum, the date of the issue of the document, the place of issue. For each person we recorded events in which they participated: birth – place, date, parents; education, working periods, marriage, etc. Most of the places mentioned in the documents were associated with one or more events of these types. Having this factual information explicitly in the text, we could find relationships between the institution or the place of education and the beneficiary of the giving document. For example, when somebody was born is some place A, but studied in place B and finally worked in place C, through the recording of this biographical information it would become clear why the donation was performed in favor of the school in place A or place B.

Additionally, the data has been encoded as RDF statements in such a way that: (1) if there is an appropriate DBpedia URI for the instance, then we use it; (2) if there is no appropriate DBpedia URI, then we create one for the corresponding entity attempting to resemble DBpedia ones. For the corresponding new instances we selected appropriate ontology classes like *dbp:Person, dbp:Politician, dbp:Village*, *dbp:Location*. If it is a location, but not represented in DBpedia, we searched for an appropriate GeoNames instance and if found, we established an *owl:sameAs statement*.

For the moment our Linked Open Data dataset is relatively small, but we consider it important with respect to the representation of people that played a crucial donor role in the Bulgarian society without having been recorded in the big datasets. The sets of documents are: 89 documents in the first period, 111 documents in the second period and 185 documents in the third period. They contain information for 461 people (some documents are related to more than one person - usually couples). For each of them we have records of information for their names, place of birth, date of birth, university, place of work, place of donation, amount of donation, currency, date and place of death. The statements are a little more that 3200. As mentioned above, we envisage to combine out approach with the micro biographies of Fokkens et al. 2018. This will ensure interoperability with other biographical datasets.

It will be interesting to compare such datasets on European level to check how many of the donors lived in different European countries and what their donating coverage was.

5 Conclusions

The paper presents the specialized corpus in the area of giving for education – CoDar, as well as the basic steps of its processing from both – linguistic and statistical points of view (including word clouds). Since Bulgarian belongs to the morphologically rich languages, the most important step was the morphosyntactic tagging. However, the error analysis showed that not only the historical periods of the language are important but also whether the texts were normalized, or not, and if yes, to which extent.

The ultimate aim of our efforts is to facilitate the extraction of appropriate content that might answer research questions and support objective generalizations within the area of socio-economic humanities. Our future work refers to: cleaning the corpus from errors added during the digitization; standardizing normalization of old and rare words; re-training of the processing modules on the cleaned and normalized data.

The project (as other similar projects before it) has impact on the planning of the developments within the Bulgarian National Interdisciplinary Research e-Infrastructure for Resources and Technologies in favor of the Bulgarian Language and Cultural Heritage, part of the EU infrastructures CLARIN and DARIAH – CLaDA-BG. From the perspective of language resources our goal is to integrate the existing ones, to extend them to a reasonable size and to construct new ones. The integration will be done through direct references and annotation according to the appropriate standards, including the LOD technologies. In addition to handling the necessary language resources, our goal will be to construct a knowledge graph of the data and tools represented within CLaDA-BG. The knowledge graph will be populated with entities extracted from existing sources like Bulgarian DBpedia, but also with entities extracted from the Bulgarian National Reference Corpus. It will be further automatically extended with documents from the Bulgarian Web Space, but also with the OCR versions of old documents, newspapers. In addition to entities from textual sources we will provide descriptions of cultural and historical artefacts. In this way the different entities will be contextualized in the sense of their co-occurrence in time and space.

With respect to language technologies behind the standard modules, mentioned here, we will work in the direction of Named Entities Recognition and Identification, Semantic Role Labeling, Event Recognition, and Coreference Resolution. Thus, we will be able not only to process the language structure of new texts, but relate them to the previously processed and represented data.

We believe that the combination of Language Resources, Language Technologies and Semantic Technologies is the only objective way to support successfully the research in the Social Sciences and Humanities.

Acknowledgements

The work reported here is done partially within the Bulgarian National Project: *Culture of giving in the sphere of education: social, institutional and personality dimensions*, Grant DFNI-K02/12. In addition, some of the work has been done within the *Bulgarian National Interdisciplinary Research e-Infrastructure for Resources and Technologies in favor of the Bulgarian Language and Cultural Heritage, part of the EU infrastructures CLARIN and DARIAH – CLaDA-BG*, Grant DO01-164/28.08.2018.

References

Anthony, L. 2014. AntConc (Version 3.4.4w) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/

Fokkens et al. 2018. Fokkens, A., Ter Braake, S., Ockeloen, N., Vossen, P., Legêne, S., Schreiber, G., De Boer, V. BiographyNet: Extracting Relations Between People and Events. At: arXiv:1801.07073 [cs.CL]

Georgiev G., Zhikov V., Simov K., Osenova P., Nakov P. 2012. Feature-Rich Part-of-speech Tagging for Morphologically Complex Languages: Application to Bulgarian. In: proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. Avignon, France. pp 492-502.

Monachesi, P., Lemnitzer, L., Simov, K. Language Technology for eLearning. In: *Innovative Approaches for Learning and Knowledge Sharing. EC-TEL 2006*. Lecture Notes in Computer Science, vol 4227. Springer, Berlin, Heidelberg, 2006, p. 667-672.

Popov, D., Simov, K. and Vidinska, S. 1998. A Dictionary of Writing, Pronunciation and Punctuation of Bulgarian Language. (in Bulgarian) Atlantis KL, Sofia, Bulgaria. 927 pages.

Popov D., Simov, K., Vidinska, S. and Osenova, P. 2003. A Spelling Dictionary of Bulgarian Language. (in Bulgarian), Nauka i Izkustvo, Sofia, Bulgaria. 808 pages.

Savkov, A., Laskova, L., Kancheva, S., Osenova, P., Simov, K. *Linguistic Analysis Processing Line for Bulgarian*. Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012), ELRA, 2012, 2959-2964

Simov, K., Osenova, P., Kolkovska, P., Balabanova, E., Doikoff, D. *A Language Resources Infrastructure for Bulgarian*. Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004), ELRA, 2004, 1685-1688.

Simova, I., Vasilev, D., Popov, A., Simov, K., Osenova P. 2014. Joint Ensemble Model for POS Tagging and Dependency Parsing. In: Proceedings of the First Joint Workshop on Statistical Parsing of Morphologically Rich Languages and Syntactic Analysis of Non-Canonical Languages. Dublin, Ireland. pp 15–25.

Staykova, K., Simov, K., Agre, G., Osenova, P. Language Technology Support for Semantic Annotation of Iconographic Descriptions. In: *Proceedings of the Workshop on Language Technologies for Digital Humanities and Cultural Heritage*, *RANLP 2011*, 2011, p. 51-56.

Zhikov, V., Georgiev, G., Simov, K., Osenova, P. *Combining POS Tagging, Dependency Parsing and Coreferential Resolution for Bulgarian*. Proceedings of RANLP, 2013, 755-762.

Operationalizing "public debates" across digitized heterogeneous mass media datasets in the development and use of the Media Suite

Berrie van der Molen
Freudenthal Institute
Utrecht University, The Netherlands
b.j.vandermolen@uu.nl

Jasmijn van Gorp
Media and Culture Studies
Utrecht University, The Netherlands
j.vangorp@uu.nl

Toine Pieters
Freudenthal Institute
Utrecht University, The Netherlands
t.pieters@uu.nl

Abstract

In this paper, we propose a methodological operationalization of "public debates" as we focus on the research process of CLARIAH research pilot Debate Research Across Media (DReAM). In this pilot, heterogeneous datasets (of digitized print and audiovisual media) were made searchable with tools of the CLARIAH Media Suite, using the leveled research approach that we coined previously (combining distant and close reading) to do historical public debate analysis. The qualitative research interest in public debates on drugs and regulation is historical, but in order to bridge the gap between distant and close reading of the combined digital datasets, a number of insights from media studies is taken into consideration. The natures of the different media, the type of analysis and focus on the source material itself, and the necessity to combine historical expertise with a sensibility towards discursive relations are all considered before we argue that the accommodation of this approach in the Media Suite helps the researcher to gain an improved understanding of historical public debates in mass media.

1 Introduction

In the research pilot Debate Research Across Media (DReAM)[1], we tested and contributed to the development of the Compare tool and related tools in the CLARIAH media research infrastructure Media Suite[2][3]. We worked to accommodate the leveled research approach that we coined earlier (Van der Molen and Pieters 2017) in the Media Suite. This explorative historical research approach assumes that a combination of distant reading techniques (keyword search, word cloud analysis and timeline graph analysis) and historical analysis (close reading) of a thematic subselection can help us to trace and understand public debates in digitized historical material. By combining relevant tools in the Media Suite, we worked to make this possible across two different datasets: the digitized newspaper dataset of the National Library of the Netherlands (KB), and the digital radio and television archive of the Netherlands Institute for Sound and Vision (NISV). The research environment is built on media studies principles; one of the main aims of DReAM was to make the Media Suite equipped for *historical* public debate research. Our historical research interest in drugs and regulation was used to test the usability of the approach. In this paper, we reach a methodological operationalization of *public debate* based on (i) theoretical reflection on the relation between the digitized datasets and public opinion and (ii) reflection on (decisions made in the development of) the research infrastructure in the CLARIAH Media Suite. This gives us a pragmatic methodological framework for use of the leveled approach to

[1] < www.clariah.nl/projecten/research-pilots/dream/dream>.
[2] The MediaSuite (<mediasuite.clariah.nl>) is CLARIAH's online media research environment accessible to all humanities researchers in the Netherlands. The infrastructure consists of different tools and datasets to be combined freely by the researcher. Our research pilot helped to make a combination of tools in this environment suitable for public debate analysis for researchers in the humanities. Some of the questions raised after the presentation of (the short version of) this paper at the CLARIN 2018 conference in Pisa (Italy) regarded the accessibility of the code for others: all of the code is open and can be found at <github.com/CLARIAH/wp5_mediasuite>.
[3] CLARIAH.nl is the Dutch infrastructure related to CLARIN.eu and DARIAH.eu.

Berrie van der Molen, Jasmijn van Gorp and Toine Pieters 2019. Operationalizing "public debates" across digitized heterogeneous mass media datasets in the development and use of the Media Suite. *Selected papers from the CLARIN Annual Conference 2018*. Linköping Electronic Conference Proceedings 159: 205–213.

research public debates in the available heterogeneous digital sources, that works in both the historical and media studies paradigm.

Keyword search has created access to large digital datasets with historical relevance to historians that would be too time-consuming to search manually (Nicholson 2013). In DReAM we wanted to benefit from this for historical public debate research by combining a number of so-called distant reading (Moretti 2013) methods and tools in the Media Suite. The most important of these tools, Compare (Comparative search), is based on a previous CLARIAH cross-media analysis tool called AVResearcherXL (Huurnink et al. 2013; Van Gorp et al. 2015). AVResearcherXL simultaneously searched the previously mentioned KB newspaper and NISV radio and television archive and offered timeline graphs, word clouds and a result viewer.

The development process was iterative: as end users, we set out by outlining our ideas and needs in a so-called Demonstration Scenario; developers then worked on this, after which we then tested the implementations and provided feedback. As such, all developer steps were based directly on our explicit research requirements. Underlying this was our ambition to enable the leveled research approach (Van der Molen et al. 2017). This research approach is based on the assumption that navigation between three levels of reading (macro, meso and micro level; see below in-text) can function as a signposting strategy to find relevant material. The leveled approach itself is based on theoretical assumptions about how the digitized source material can be understood as relating to a public debate, which will be explained in detail below. The accommodation of the leveled approach in the Media Suite, based on our researcher needs but also on pragmatic decisions made in the development process, also frames what we mean exactly when we call this approach public debate analysis.

This paper consists of two parts. In the first part, we reflect on the methodological question of the research pilot, which results in a theoretical connection between the (digitized) source material and a particular conceptualization of public debate based on Jürgen Habermas' writing on the public sphere. The second part completes the methodological operationalization of public debate, by reflecting on how implementations in the research infrastructure lead to further explication of this type of public debate analysis that highlights discursive relations in the relevant cross-media public debates. In this second part, we argue that this approach enables researchers to uncover specific discursive strands in the source material, along with relevant results in need of close reading, resulting in an improved understanding of important themes and power relations in historical debates regarding drugs and regulation in print and audiovisual mass media.

2 Theorizing "public debates"

The methodological question that we aimed to answer in the pilot is "How can public debates on drugs and regulation between 1945 and 1990 be researched across print and audiovisual datasets?". This question means that we set out to safeguard both the historical expressiveness and the methodological soundness of the research infrastructure. Before we get to describe the implementations, it is necessary to explicate the assumption that there is a relation between the relevant datasets and historical public debates.

The qualitative research interest of the research pilot is primarily historical, as it is embedded in historical research project The Imperative of Regulation, in which the postwar drug history of the Netherlands is scrutinized[4]. The historian's primary concern is the careful contextualization of events that does justice to the actors involved. Historical research has a long tradition of source criticism and awareness of the constructive and interpretative role of historians in their efforts to produce an informed understanding of the past. Historians understandably take an ambiguous stance towards digital humanities (DH) techniques. On the one hand, they are sometimes critical towards leaving part of the interpretative process to algorithms, and the quantitative component (word counts and distances) seems to be at odds with the interpretative practice of *understanding* the past. But on the other hand, they embrace the benefits of mass access to historical sources granted by digitization (e.g. Zaagsma 2013), and recent research output continues to highlight the potential of combining historical research with mass access to digital source material (e.g. Klein 2018).

[4] <www.nwo.nl/onderzoek-en-resultaten/onderzoeksprojecten/i/46/13546.html>.

Although media studies is a heterogeneous field of study (Bron et al. 2016, 1536), many strands within it depart from theoretical conceptualization in order to understand the complex role of media and their relation to both media producers and media consumers. Our claim is that drawing on conceptual insights from media studies is one way to critically bridge the gap between distant and close reading of digital media sources to reconstruct historical public debates. Bridging this gap, or fully grasping what goes on between the different distant reading methods and close reading of the material, is essential in order to make the research environment solid for historical research on public debates.

Figure 1. The digitized datasets in the context of the public sphere[5]

Assuming a relation between what is said or written about a particular topic in national or local media and "public debates" of said topic seems obvious. In order to understand public debates on drugs and regulation between 1945 and 1990, it appears natural to research the newspapers and radio/television broadcasts of this time period, as is readily implied in the methodological question of the pilot. But in order to make concrete and meaningful sense of the material in these datasets, an explicit theorization of this relation is required. How *do* these mass media relate to the public debates on a national level? Jürgen Habermas has argued that in modern societies, mass media are part of a public sphere that accommodates a 'society engaged in critical public debate' (Habermas 1989, 52). This perspective, which is rooted in critical theory, is particularly useful for our research aim because of its critical stance towards power relations in society. Habermas' conception of the public sphere and mass media means that the existence of such a public sphere could foster true democratic public opinion, but it could also be a sphere in which the bourgeois class reproduces desirable political thought (Outhwaite

[5] Figures 1 and 3 were developed in cooperation with Frank-Jan van Lunteren <collageboys.nl>.

2008, 251). This can be translated into a necessity to remain observant when it comes to who is given a voice: also on radio, television and in newspapers. This is naturally relevant when it comes to debates on drugs and regulation. Is the public opinion on a particular substance mostly defined by its users, by policymakers or by different actors such as law enforcers or medical specialists? The question is not just *how* public opinion transforms over time, but also *who* gets to be a part of this process.

From a historical perspective, this means that we need to underline that tracing this type of public opinion in mass media is a very specific type of public debate analysis that focuses on the meaning-making process in national and local print and audiovisual mass media. Looking at these statements in national mass media precludes a focus on oral history, on backdoor politics, on non-mainstream media, or even on what mass media producers and consumers *actually* thought about drugs. Those would be equally relevant areas of inquiry that are not covered by this approach.

With the relation between the two datasets and public opinion established (see Figure 1), we can move on to the implementations of the research pilot in order to reach an even more precise methodological operationalization of public debate, based not just on theoretical reflection but also on the infrastructure of the Media Suite. The embedment of the levelled approach in the research infrastructure needs to enable more than the unearthing of *what* has been said about drugs and regulation in the relevant period, it also should allow the researcher to grasp what actors were featured prominently and what actors were excluded from this type of public opinion.

3 Operationalizing "public debates" in the CLARIAH Media Suite

The CLARIAH Media Suite is an online infrastructure that provides media scholars and digital humanists access to datasets from different institutional providers for exploration and mixed-media research (Ordelman et al. 2018)[6]. In order to align concrete researcher needs with the development of this infrastructure, several research pilots comprising scholars and developers tested and contributed to parts of the Media Suite during its development . Our research pilot DReAM aimed to accommodate public debate analysis capable of answering historical research questions. Public debate analysis in the Media Suite can be done by combining the digital tools Collection Inspector, Search and Compare (Comparative search) with the Workspace. Broadly speaking, Collection Inspector is used to gain an understanding of the composition of the different datasets, while Search and Compare are subsequently used to query and analyze the inspected datasets, with the Workspace allowing analysis and annotation of the bookmarked results. Below, we will describe all these methodological steps necessary for public debate analysis with the tools of the Media Suite, thereby also describing the elements of the Media Suite we tested and/or co-developed in the research pilot.

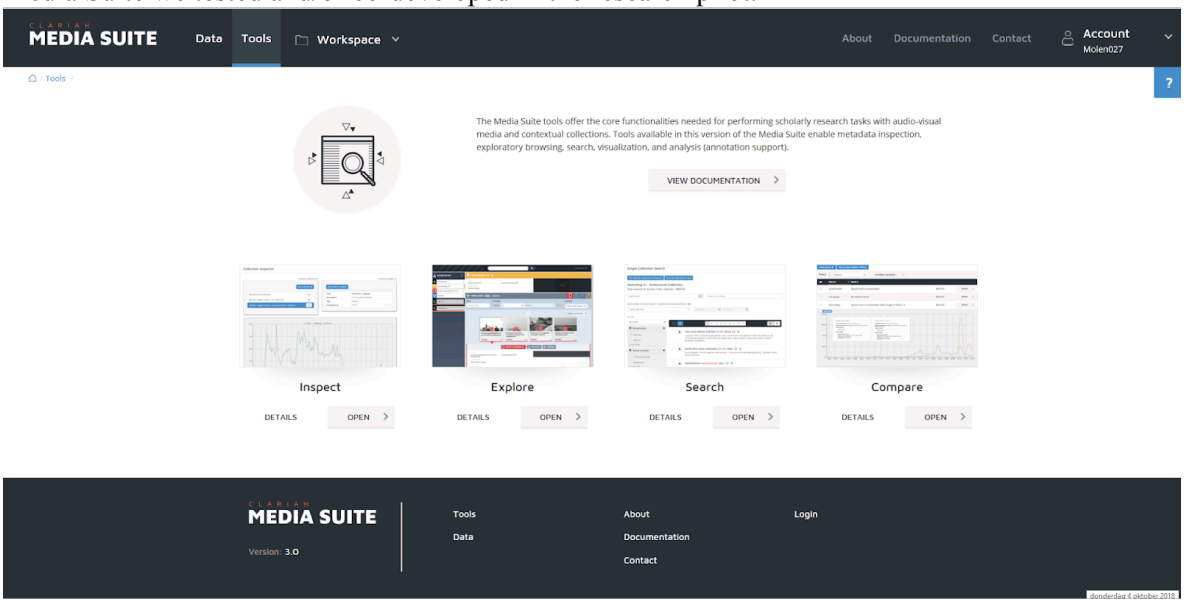


Figure 2. The tools tab of CLARIAH Media Suite Version 3.0 (October 2018)

[6] See also: contribution by Ordelman et al in this volume.

First, each dataset is loaded in the tool Collection Inspector for an assessment of metadata completeness. This allows the researcher to assess the usability of the different datasets. Historical interpretation of the data is only possible with a sufficiently complete date field for both datasets. This is a requirement, because our research question can only be answered if the data can be contextualized historically. A further requirement that needs to be checked in Collection Inspector is whether there are sufficient

a. Optical Character Recognition (OCR) metadata for the newspaper dataset
b. Automatic Speech Recognition (ASR) metadata for the radio and television datasets[7]

This ensures that both datasets are searchable on a similar (textual) level. Newspaper articles without searchable OCR metadata or audiovisual broadcasts without searchable ASR metadata cannot be found using keyword search, which is the first step of the leveled approach. The researcher is then able to send a selected dataset based on specified complete metadata to the next tool: Search[8].

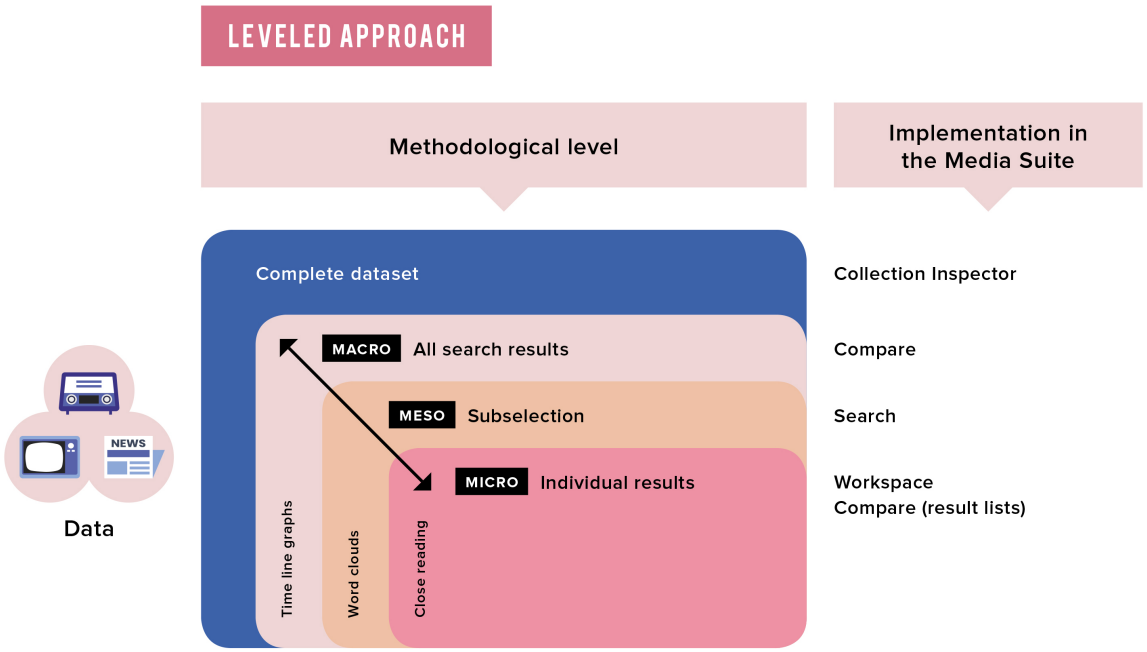


Figure 3. The leveled approach and its embedment in the Media Suite[9]

At this point, the datasets can be analyzed in the tools Search and Compare, using the leveled approach. Now that their metadata have been scrutinized in Collection Inspector, both datasets can be queried by means of specific keyword search queries (macro level) in Search. Combined queries using Boolean operators yield relevant results about particular substances. The party drug ecstasy, for in-

[7] At the time of writing (January 2019) the data integration process for the NISV data is ongoing but not completed yet. Recent status updates regarding the process can be found here: <mediasuitedata.clariah.nl/dataset/nisv-catalogue>.
[8] Depending on the type of research question and approach, the completeness of all metadata fields can be checked in Collection Inspector. The data requirements we describe in-text are necessary for the embedment of the leveled approach in the Media Suite. If, for instance, we would want to do quantitative exploration of different actors, we could search for the relevant metadata fields (e.g. actor/person/author/guest/presenter etc.) in the collections and check whether they are complete enough to proceed with such an approach.
[9] A previous version of this schematic overview appeared in the book chapter we wrote about the leveled approach before the DReAM pilot commenced (Van der Molen and Pieters 2017).

stance, is traced using the search query 'xtc OR ecsta*y OR mdma', comprising the most common spelling variations of its street name and its chemical abbreviation. This query is performed in Search for both the newspaper dataset and the audiovisual dataset, and these queries are subsequently stored in the Workspace. The Compare tool enables the researcher to display these two searches simultaneously by loading them from the relevant project in the Workspace, allowing for comparison of the different searches in the heterogeneous datasets. The search results can be further explored by means of timeline graphs (macro level) and word clouds (meso level)[10]. Furthermore, individual results are listed and can be sorted in several ways. The researcher has an option to bookmark results to the Workspace for structural analysis of the results (micro level). Although this search strategy works as a signposting strategy or funnel, the approach must be performed iteratively. Before the researcher decides to actually analyze the final subselection on the micro level, the query will most likely need to be adjusted a few times. All of these methodological steps along with their implementation location in the Media Suite have been visualized in Figure 3.

When all of these functionalities are combined in a savvy manner, they thus allow for analysis of a cross-media dataset ("public debate") that is thematically and chronologically linked. Pursuing this type of cross-media public debate analysis raises several points of reflection from a media studies perspective. First, we need to reflect on what it means to perceive combined datasets from different media types as public debates, for these media are not just neutral conveyors of messages (e.g. Derrida 1996). Any media, in this case television, radio and print media, function differently and, according to Marshal McLuhan (1964), they even *are* the message (as opposed to what we would traditionally understand as the content): what these media convey is defined to some degree by each medium. In that sense, in order to describe a historical public debate, it is necessary to understand precisely how different media can contribute to a meaningful public debate. The search results in the leveled approach remain clustered in their respective medium-specific datasets, meaning that the researcher can reflect on how the different media contribute differently to public opinion about drugs.

To complicate things more, there are two further layers/media to take into consideration: the digitization processes for both datasets, plus, most importantly, the way digitized datasets are made available and searchable in the Media Suite.[11] The textual data is searchable by means of the OCR data; the audiovisual data is searchable by means of the ASR data. Doing this on this scale is unexplored methodological territory, and it naturally forces reflection on how we can still do justice to the *visual* meanings of the television data. In other words, the distant reading steps of the leveled approach in the Media Suite are currently all based on textual metadata. On the close reading level, this is not the case: with the annotation tool the broadcasts can be annotated (based on whatever visual elements) by means of time-coded tags or comments[12].

Secondly, there are different meaningful focus points when it comes to studying media. Should a public debate analysis based on digitized newspaper, television and radio sources focus on agenda setting points (production history analysis), on what there is *in* the sources (textual analysis), or on how they were likely understood by the public back then (reception research)? All of these meanings are valid angles when it comes to researching the public sphere and public opinion, as has become clear in the previous paragraph. There are many ways to understand and account for the different levels of meaning on this continuum, for instance the encoding/decoding model that claims that audiences decode the media they consume based on their individual backgrounds, meaning that media can have

[10] At the time of writing the word cloud functionality has not been integrated yet. Since Summer 2018 this has been accommodated in a Jupyter notebook. Word cloud functionality is scheduled to be implemented in the Media Suite in April 2019 as part of CLARIAH PLUS, the follow-up to CLARIAH.

[11] It is important to be aware of *which* newspapers or television and radio shows are available in the digitized datasets too, as a reasonable sample (e.g. conservative or progressive titles or broadcasters) is necessary for the approach to yield a narrative that can truly contribute to an improved understanding of a general public opinion on drugs and regulation. A further point to be made is that the datasets primarily comprise news media, which further delineates the meaning of "public debate" here.

[12] We learned an important further lesson regarding digitization and data accessibility methods during the recent CLARIAH Summer School that was organized for Dutch researchers to test the latest version of the Media Suite with sample projects. We led a project on the representation of refugees in Dutch audiovisual media. The effect of some infrastructure design decisions on the representation of refugees was found to be considerable. Speech recognition fails at picking up non-Dutch languages, meaning that everything that has been said by refugees themselves does not become part of the searchable data, making the searchable discourse mostly defined by reporters and politicians. A number of related relevant findings from the summer school are described here: <www.beeldengeluid.nl/kennis/blog/clariah-summer-school-2018>.

different encoded (intended) and decoded (interpreted) meanings (Hall 1980). It is therefore highly problematic to assume that meaning is consistent across these different focus points. All different focus points could lead to meaningful interpretations of the relation between public opinion and mass media. Public debate analysis thus has to be explicit about where on this continuum it locates meaning and, equally importantly, which meanings are then *excluded* from this type of historical narrative. The Media Suite does not directly isolate and contextualize historical events: its reliance on distant reading techniques means that it groups historical sources based on strategies predefined by the infrastructure. The historical meaning is distilled from the digitally combined source material (found in the heterogeneous datasets) itself, which precludes a focus on production and reception. The arrows in Figure 1 could signify each of the meaningful relations (production, text, reception), but in our research approach the scope is limited to textual analysis.

The last related point of reflection is that this more or less artificial nature of public debate requires a theoretical approach. We have already established that we perceive the newspaper as one of the arenas of the public sphere in a society of mass media (as per Habermas), but we also need to be explicit about how the actual infrastructure of the Media Suite implies a particular conceptualization of historical public debates. Since the Media Suite does not isolate, group or contextualize historical events, knowledge of the historical contextualization (the different newspapers, sections, actors), along with a sensibility towards discursive relations, is necessary to signalize meaningful discursive strands within the search results. The grouped data are related in the sense that they are produced by the same society at the same time, meaning they can be understood as part of what Michel Foucault has called discourse: the culturally constructed conditions of truth. These conditions of truth are in dialogue with power over the truth in *all* relations between people (Foucault 1998, 97). This means that discursive conditions enable and restrict what can be said about ecstasy at any given time. Discourse is continuously reproduced in all relations between people, and looking at these moments of reproduction can help us understand how discourse may shift over time. Looking at actors is thus essential. Following Habermas, in modern societies, discursive shifts and its most prominent actors can be traced in the arenas of the public sphere. The digital search method makes it possible to collect and interpret a large number of articles and broadcasts mentioning any particular drug as interrelated in something that can be called a history of the public discourse of drugs in mass media. An awareness of the role of the different actors is important to understand who is most prominent in leading public opinion, which, as we have already seen, is of crucial importance to understand the role of the public sphere itself too. What needs to be researched in order to understand shifts in the discursive formation of drugs and regulation in the public sphere are all the different moments where they are discussed, which is what Foucault called looking at the techniques themselves in a search for patterns (Foucault 2004, 8). This is why the leveled approach ultimately functions as a signposting strategy: understanding historical shifts in public opinion depends on the eventual close reading of the source material.

This brings us back to the importance of researcher expertise that is crucial in the leveled approach to trace and understand the possible connections and different actors in the results. Despite plenty of noise (due to OCR issues or dual word meanings (e.g. XTC as a drug name and XTC as a band name)), sufficient historical contextual knowledge (based on historical expertise, previous research and secondary literature) allows us to recognize meaningful historical relations. The leveled research approach makes it possible to find specific relations with word clouds, and to trace these relations with targeted queries. The query '(xtc OR mdma OR ecsta*y) AND (acid OR house OR acidhouse OR dance)' yields results that relate to ecstasy's reputation of a party drug, whereas the query '(xtc OR mdma OR ecsta*y) AND (politie OR inval OR laboratorium OR onderzoek[13])' yields results about the (prosecution of the) illegal production of ecstasy. Performing the approach iteratively in the Media Suite can help us to define the most important discursive strands, that with extensive close reading can help us understand developments within public debates of or public opinion on drugs. By recognizing how a particular drug undergoes changes in the way it is framed over time (e.g. how use of the substance is either normalized or "othered") across the different results, very specific nuances can be applied to our historical understanding of the socio-cultural context of the drug, or any topic with historical relevance.

[13] The second half of this query translates as 'police OR raid OR laboratory OR investigation'.

4 Conclusion

In this paper, we proposed a methodological operationalization of "public debates" based on theoretical reflection and the resulting pragmatic development decisions we made. This approach is not aimed towards re-constructing particular debates as they happened; instead, it focuses on discursive processes and is a result of critical reflection on the CLARIAH Media Suite infrastructure, grounded in historical research and safeguarded with media studies sensibilities. By searching and analyzing the relevant datasets with the leveled approach in the Media Suite, it is possible to become aware of shifts in the discursive formation of particular topics. Although this is a fundamentally constructive exercise, reliance on historical contextual expertise makes it possible to improve our understanding of historical relations and discursive dynamics of public debates across media and the roles of the different media in this process. For our qualitative research interest in drugs and regulation, this means that tracing and following different substances in the national print and audiovisual media enables us to answer historical questions about the dynamics of public debates in mass media and about the interaction between regulation and public debates, based on fine-grained reading of the digitized source material.

References

[Bron et al 2016] Marc Bron, Jasmijn van Gorp, Maarten de Rijke. 2016. "Media studies in the data-driven age. How research questions evolve." *Journal of the Association for Information Science and Technology*. 67(7), 1535-1554.

[Derrida 1996] Jacques Derrida. 1996. *Archive fever. A Freudian impression*. Chicago: University of Chicago Press.

[Foucault 1998] Michel Foucault. 1998. *The will to knowledge. The history of sexuality: 1*. London: Penguin Books.

[Foucault 2004] Michel Foucault. 2004. *Security, Territory, Population. Lectures at the Collège de France 1977-1978*. New York: Picador.

[Habermas 1989] Jürgen Habermas. 1989. *The structural transformation of the public sphere. An inquiry into a category of bourgeois society*. Cambridge: MIT Press.

[Hall 1980] Stuart Hall. 1980. "Encoding/decoding." In: Stuart Hall, Dorothy Hobson, Andrew Love and Paul Willis (eds.) *Culture, Media Language*. London: Hutchinson.

[Huurnink et al 2013] B. Huurnink, A. Bronner, M. Bron, J. van Gorp, B. de Goede, J. van Wees. 2013. AVResearcher: Exploring Audiovisual Metadata. DIR 2013: Dutch-Belgian Information Retrieval Conference Delft: DIR.

[Klein 2018] Wouter Klein. 2018. *New Drugs for the Dutch Republic. The Commodification of Fever Remedies in the Netherlands (c. 1650-1800)*. Utrecht: Freudenthal Institute, FI Scientific Library no. 101.

[McLuhan 1964] Marshal McLuhan. 1964. *Understanding Media*. London: Routledge.

[Moretti 2013] Franco Moretti. 2013. *Distant Reading*. London: Verso.

[Nicholson 2013] Bob Nicholson. 2013. "The digital turn. Exploring the methodological possibilities of digital newspaper archives" *Media History* 19.1

[Ordelman et al 2018] Roeland Ordelman, Liliana Melgar, Carlos Martinez-Ortiz, Julia Noordegraaf. (2018) "Media Suite. Unlocking archives for mixed media scholarly research." In: Inguna Skadina, Maria Eskevich (eds.). *CLARIN Annual Conference 2018. Proceedings.* <office.clarin.eu/v/CE-2018-1292-CLARIN2018_ConferenceProceedings.pdf>. 21-25.

[Outhwaite 2008] William Outhwaite. 2008. "Jurgen Habermas". In: Rob Stones (ed.) *Key Sociological Thinkers*. Second Edition. New York: Palgrave MacMillan, (251-260): 254.

[Van Gorp et al 2015] J. van Gorp, J.S. de Leeuw, J. van Wees, B. Huurnink. 2015. "Digital media archaeology. Digging into the digital tool AVResearcherXL." *VIEW. Journal of European Television History and Culture*. 4.7, 38-53.

[Van der Molen et al 2017] Berrie van der Molen, Lars Buitinck, Toine Pieters. 2017. "The leveled approach. Using and evaluating text mining tools AVResearcherXL and Texcavator for historical research on public perceptions of drugs." arXiv:1701.00487.

[Van der Molen and Pieters 2017] Berrie van der Molen, Toine Pieters. 2017. "Distant and close reading of Dutch drug debates in historical newspapers. Possibilities and challenges of big data research in historical public debate research." In: Arun K. Somani, Ganesh Chandra Deka (eds.). *Big Data Analytics. Tools and Technology for Effective Planning*. Boca Raton: CRC Press, 373-390.

[Zaagsma 2013] Gerben Zaagsma. 2013. "On Digital History." *BMGN - Low Countries Historical Review*, 128.4, 3–29.

LaMachine: A meta-distribution for NLP software

Maarten van Gompel and Iris Hendrickx
Centre for Language and Speech Technology (CLST)
Radboud University, Nijmegen, the Netherlands
proycon@anaproy.nl, i.hendrickx@let.ru.nl
https://proycon.github.io/LaMachine

Abstract

We introduce LaMachine, a unified Natural Language Processing (NLP) open-source software distribution to facilitate the installation and deployment of a large amount of software projects that have been developed in the scope of the CLARIN-NL project and its current successor CLARIAH. Special attention is paid to encouragement of good software development practices and reuse of established infrastructure in the scientific and open-source software development community. We explain what LaMachine is, how it can be used and the technical details. We also compare LaMachine to alternative software distributions and discuss its advantages and limitations. We illustrate how LaMachine can be used in two case studies, one in an exploratory text mining project at the Dutch Health Inspectorate where LaMachine was applied to create a research environment for automatic text analysis for health care quality monitoring, and a second case where LaMachine was used to create a workspace for a one-week, intense collaboration by a diverse research team.

1 Introduction

Software is a key deliverable and a vital component for research in projects such as those under the CLARIN umbrella. Software can be seen as an invaluable instrument that enables researchers to perform their studies. It is CLARIN's core mission to make digital language resources, including software, available to the wider research community.

We see that NLP software often takes on complex forms such as processing pipelines invoking various individual components, which in turn rely on various dependencies. Add dedicated web-interfaces on top of that and you obtain a suite of interconnected software that is often non-trivial to install, configure, and deploy. This is where LaMachine comes in; as a *distribution of software*.

Software is a broad but well-ingrained notion in society nowadays, referring to any form of computer program. This can manifest in various forms, whether it is an app on a phone, a graphical desktop application, a command line tool, or a web-based application. These different *interfaces* generally address different *audiences*; the data scientist will feel at home at the command line and in scripting environments and uses these fairly low-level interfaces, whereas researchers in the humanities demand higher-level graphical interfaces. There is often a power trade-off between lower and higher-level interfaces, with the former providing maximum flexibility at the cost of a steeper learning curve, technical ability, and a do-it-yourself mentality. Higher level interfaces, on the other hand, expose certain functionality of the software in an easy and accessible way, but in doing so often can not expose the full power of the software. The cost trade-off is also apparent in the construction of the interfaces, where high-level interfaces are typically far more costly to build.

LaMachine distributes software providing different types of interfaces which, as seen above, typically address different audiences. Whilst we attempt to accommodate both technical[1] and less-technical audiences[2]. There is a natural bias towards the former as lower-level interfaces are often a prerequisite to build higher-level interfaces on. Depending on the *flavour* (more on this later) of LaMachine chosen, it

[1] Data scientists, DevOps, system administrators, developers.
[2] The wider researcher community, particularly the Humanities; also educational settings.

Maarten van Gompel and Iris Hendrickx 2019. LaMachine: A meta-distribution for NLP software. *Selected papers from the CLARIN Annual Conference 2018*. Linköping Electronic Conference Proceedings 159: 214–226.

makes a good virtual research environment for a data scientist, whether on a personal computer or on a computing cluster, a good development environment for a developer or a good deployment method to deliver to production servers such as CLARIN centres.

This paper provides a detailed description of LaMachine, its purpose and functionality and the technical details. We discuss the advantages and limitations of LaMachine compared to alternative software distributions in Section 6. We demonstrate how LaMachine can create a fully functioning and standalone research environment for text mining and NLP for Dutch texts in two use cases in Section 7.

2 What is LaMachine?

LaMachine is an open-source NLP software distribution. LaMachine facilitates the installation, distribution and configuration of software. It does not fork, modify or appropriate the participating software in any way, nor does it provide a hosting place or repository for software. We classify LaMachine as a *meta distribution* as it can be installed in various contexts. The heart of LaMachine consists of a set of machine-parsable instructions on how to obtain, build (e.g. compile from source), install and configure software. This is notably different from the more classical notion of Linux distributions, which generally provide their own repositories with (often binary) software packages. LaMachine builds on this already established infrastructure by taking these repositories as a foundation where possible. Similarly, as implied in point five above, there are different programming-language-specific ecosystems providing their own repositories, such as the Python Package Index[3] for Python, CRAN[4] for R, CPAN[5] for Perl, Maven Central[6] for Java. LaMachine relies on those to obtain and install software. In doing so, we compel participating software projects to adhere to well-established distribution standards and ensure the software is more sustainable (van Gompel et al., 2016). Moreover, we ensure that LaMachine never becomes a prerequisite for the software but merely a courtesy or convenience, and attempt to limit any amount of duplication in packaging and distribution efforts.

LaMachine lives in an open-source ecosystem and therefore builds on Unix-like platforms; this primarily means Linux, as well as BSD and, with some restrictions, macOS. This by definition excludes certain software for different platforms, such as mobile platforms (Android/iOS/etc), native Windows/mac desktop applications, or certain interface types in general such as classical desktop GUI applications or mobile 'apps', all of which fall beyond our scope. Cygwin[7] is not tested or supported either. However, virtualisation technology enables deployment on a wider range of platforms, including Windows.

LaMachine exists since May 2015 and has been installed extensively ever since by numerous users. Quantifying its actual use is not trivial despite having some tracking solutions in place; our statistics show that LaMachine was installed at least over a thousand times, on machines in over 25 different countries, since its inception[8]. In early 2018 version 2 was released which was a significant redesign, powered by Ansible[9]. We ourselves use LaMachine as a means by to make our full software stack available to end-users, whether those be individual researchers, research groups, or hosting providers like CLARIN centres. LaMachine has in fact been adopted in all of those areas. We also see a regular influx of questions on our issue tracker, indicating community interest.

The software included in LaMachine has to adhere to the following prerequisites:

1. The software must have a recognised (i.e. OSI-approved) open-source license. Proprietary (closed-source) software is explicitly excluded.
2. The source code must be hosted in a public version controlled repository.[10]
3. There must be a build process to compile the source, if applicable, and install the program or library.

[3] `https://pypi.org`
[4] `https://cran.r-project.org/`
[5] `https://www.cpan.org`
[6] `https://search.maven.org`
[7] A unix environment on Windows
[8] https://applejack.science.ru.nl/lamastats/lamachinestats.html
[9] `https://www.ansible.com`
[10] e.g. Github, Gitlab, Bitbucket, provided the repository is public.

4. There must be some kind of release protocol (adhering to semantic versioning) that publishes software using the proper technology-specific channels. We will elaborate on this later.
5. All software that is incorporated in LaMachine must bear at least some relevance to the field of Natural Language Processing.
6. Participating software must be actively maintained (i.e. not outdated or abandoned) and not place any demands on outdated dependencies.

3 Included Software

In LaMachine, software is grouped into various "packages", each package[11] groups one or multiple programs that have some kind of relation. The installation manifest lists all packages that will be installed, at the user's discretion. After the initial installation, the user can always add more packages using `lamachine-add` or by editing the installation manifest directly.

LaMachine was initially conceived as the primary means of distribution of the software stack developed at the Language Machines Research Group and the Centre of Language and Speech Technology, Radboud University Nijmegen. The majority of this sofware was either fully or partially developed under the auspices of CLARIN-NL or successor CLARIAH. Some software by other CLARIN-NL/CLARIAH partners is also included. LaMachine is not limited to one research group and is explicitly open to participation by other software providers, especially those also in CLARIAH.

We list a selection of the most important software included in LaMachine, grouped by research institute:

- by the Language Machines Research Group and the Centre of Language and Speech Technology, Radboud University, Nijmegen:[12]
  - **Timbl** – A memory-based machine learning toolkit, and **Mbt**, a memory-based tagger based on timbl. Python bindings included as well.
  - **Ucto** – A multilingual rule-based tokeniser. Python binding included as well.
  - **Frog** – An integration of various memory-based natural language processing (NLP) modules developed for Dutch. It can do Part-of-Speech tagging, lemmatisation, named entity recogniton, shallow parsing, dependency parsing and morphological analysis. Also included in LaMachine; Python bindings for Frog and **Toad**, Trainer Of All Data, training tools for Frog.
  - **Wopr** – Memory-based Word Predictor.
  - **CLAM** – Quickly build RESTful webservices, powers many webservices offered by LaMachine.
  - **FoLiA** – Format for Linguistic Annotation (van Gompel and Reynaert, 2013), with tools and libraries in/for Python and $C++$.
  - **FLAT** – FoLiA Linguistic Annotation Tool: a web-based linguistic annotation tool.
  - **PyNLPl** – Python Natural Language Processing Library.
  - **Colibri Core** – Colibri core is an NLP tool as well as a C++ and Python library for working with basic linguistic constructions such as n-grams and skipgrams (i.e patterns with one or more gaps, either of fixed or dynamic size) in a quick and memory-efficient way.
  - **Gecco** – Generic Environment for Context-Aware Correction of Orthography, an NLP pipeline for spelling correction, and **Valkuil.net**, an instantiation thereof for Dutch.
  - **PICCL** – A set of workflows (NLP pipeline) for corpus building through OCR, post-correction (through **TICCL**) and Natural Language Processing.

[11] For those familiar with Ansible, a package in LaMachine is an Ansible role

[12] Links: `https://languagemachines.github.io/timbl`, `https://languagemachines.github.io/mbt`, `https://languagemachines.github.io/ucto`, `https://languagemachines.github.io/frog`, `http://ilk.uvt.nl/wopr`, `https://proycon.github.io/folia`, `https://proycon.github.io/clam`, `https://github.com/proycon/flat`, `https://proycon.anaproy.nl/pynlpl`, `https://proycon.github.io/colibri-core/`, `https://github.com/proycon/gecco`, `https://github.com/proycon/valkuil-gecco`, `https://github.com/LanguageMachines/PICCL`, `https://github.com/LanguageMachines/ticcltools`, `https://github.com/proycon/labirinto`, `https://github.com/proycon/oersetter-webservice`

  - **Labirinto** – A web-based portal listing all available tools in LaMachine, an ideal starting point for LaMachine.
  - **Oersetter** – A Frisian-Dutch Machine Translation system in collaboration with the Fryske Akademy.
- by the University of Groningen: **Alpino**[13] – A dependency parser and tagger for Dutch.
- by the Vrije Universiteit Amsterdam: **KafNafParserPy**[14] – a python module to parse NAF files, another format for linguistic annotation; and **NafFoLiApy**[15] – converters between FoLiA and NAF.
- by Utrecht University: **T-scan**[16] – a Dutch text analytics tool for readability prediction (initially developed at TiCC, Tilburg University, and in collaboration with Radboud University, Nijmegen)
- by the Meertens Instituut: **Python Course for the Humanities**[17] – interactive tutorial and introduction into programming with Python for the humanities.

In addition to the above listed specific software, LaMachine also incorporates a large number of renowned tools by external international parties, offering most notably a mature Python environment with scientific modules. The following list gives an impression and is not exhaustive: [18]

- **Python**: Numpy, Scipy, Matplotlib, Scikit-Learn, IPython, Jupyter, ...
  - **Jupyter Lab** The successor of the popular Jupyter Notebooks, offers notebooks, a web-based IDE, terminals. An ideal entry point to get started with LaMachine and all it contains!
  - **PyTorch** - Deep-learning library for Python
  - **NLTK** - Natural Language Toolkit for Python
  - **Spacy** - Industrial-Strength NLP in Python
- **R**, **Perl**, **Java**:
  - **NextFlow** - A system and language for writing parallel and scalable pipelines in a portable manner.
  - **Stanford CoreNLP** - Various types of linguistic enrichment
- **Tesseract** - Open Source Optical Character Recognition (OCR)
- **Tensorflow** - Open-source machine learning framework
- **Kaldi** - Speech Recognition Framework (ASR)
- **Moses** - Statistical Machine Translation system

4 Architecture

In this section we present the technical design choices that were made and lay out the architecture of LaMachine.

4.1 Flavours

LaMachine provides ample flexibility that allows it to be deployable in different contexts. First of all there is flexibility regarding the installation form, which we call *flavours*:

1. **Local installation** – This installs the bulk of LaMachine in a separate local virtual environment (a seperate directory[19]) that has to be explicitly activated to be used. This also allows for multiple different installations of LaMachine on the same host system (e.g. for different users or with different software configurations). This local installation still actively relies on various global dependencies that are available through the package manager of your distribution.

[13] `http://www.let.rug.nl/vannoord/alp/Alpino/`
[14] `https://github.com/cltl/KafNafParserPy`
[15] `https://github.com/cltl/NafFoLiApy`
[16] `https://github.com/proycon/tscan`
[17] `http://www.karsdorp.io/python-course/`
[18] Links: `https://jupyterlab.readthedocs.io/en/stable/`, `https://pytorch.org`, `http://www.nltk.org`, `https://spacy.io`, `http://www.nextflow.io`, `https://stanfordnlp.github.io/CoreNLP/`, `https://github.com/tesseract-ocr/tesseract`, `https://tensorflow.org`, `http://kaldi-asr.org`, `http://www.statmt.org/moses`
[19] Python users should know we just use `virtualenv` for this, with some additions of our own

2. **Global installation** – This flavour is used for a host that is fully dedicated to LaMachine. Everything will be installed globally so there is only one installation possible, multiple users will all make use of the same installation. Unlike the local installation, we do not make use of a Python Virtual Environment in this flavour.
3. **Docker container** – Installs LaMachine in a container. Containerisation separates the entire runtime environment from the host system, only the Linux kernel is shared. It is a lighter option than full virtualisation[20]. Support for two other forms of containerisation, i.e. Singularity[21] and LXC/LXD[22], are currently under development.
4. **Virtual Machine** — Installs LaMachine as a virtual machine, i.e. through full virtualisation. This allows deployment of LaMachine on hosts which would otherwise not support it, such as Windows. Virtualisation for LaMachine is achieved through Vagrant and VirtualBox[23].
5. **Remote provisioning** – Installs LaMachine on a remote dedicated server. This option is most suited for hosting centres and directly uses Ansible's remote provisioning abilities.

The different flavours all offer a different degree of separation from the host OS, where Virtual Machines are completely virtualised, containers still share the kernel with the host OS, and the two native installation flavours, local and global, actually compile against the machine's distribution itself and thus offer the least amount of overhead.

The two native options support a variety of major GNU/Linux distributions: Debian, Ubuntu, Arch Linux, CentOS, Fedora and, to a more limited degree, we also support macOS, powered in part by the homebrew package manager[24]. Support for macOS is limited because not all participating software supports it natively[25]. Certain Linux Distributions that are derivatives of the aforementioned distributions *may* also work, such as RedHat Enterprise Linux (CentOS) and Linux Mint (Ubuntu).

For the containerisation and virtualisation solutions, the default distribution we supply is Debian[26]. It is, however, still possible to build your own container or virtual machine based on any of the other supported distributions. In fact, containers, virtual machines and remote provisioning can all be considered special wrapped forms of the global installation flavour.

Pre-built docker containers and virtual machine images with a limited selection of participating software are uploaded to the Docker Hub[27] and Vagrant Cloud[28], respectively, for each LaMachine release.

4.2 Versions

LaMachine offers three distinct *versions*, regardless of the flavour:

1. **stable** - This is the default and recommended for most situations. It installs the latest stable versions of all included software.
2. **development** - This installs the latest development versions of the included software. In practise, this usually means that software is pulled directly from the version-controlled repository is and compiled and installed from source. Due to the experimental nature, the development version of LaMachine may at times break and not install successfully.
3. **custom** - This installs custom versions of all included software, i.e. the user explicitly specifies which versions to install for each software package. Such a version list can for instance be exported from another LaMachine installation, and then allows to rebuild a similar environment from scratch, providing a limited level of reproducibility. We say limited, because packages provided by the underlying distribution are not a part of this scheme.

[20] Note that Docker on other platforms such as Windows and macOS does do use full virtualisation.
[21] `https://www.sylabs.io/singularity/`
[22] `https://linuxcontainers.org`
[23] `https://vagrant.org`, `https://www.virtualbox.org`
[24] `https://brew.sh`; alternatives such as macports are not supported
[25] One can always still fall back to the virtualised flavour of LaMachine
[26] The latest stable release
[27] `https://hub.docker.com/r/proycon/lamachine/`
[28] `https://app.vagrantup.com/proycon/boxes/lamachine`

The nomenclature is admittedly a bit confusing, but the notion of version discussed in this section refers to the versions of the various software packages inside LaMachine. It should not be confused with the actual *version release number* of LaMachine as a whole, which is the version number assigned to the collective of installation scripts LaMachine provides, and which marks LaMachine releases.

4.3 Bootstrapping

Installation of LaMachine begins with a single *bootstrap* command[29] executed on the command line. It can interactively query the user for her software preferences *(stored as the host configuration)*, e.g. the flavour of LaMachine, as well as the set of software to install, *the installation manifest*. This set is never static but can be customised by the user. The bootstrap procedure, a screenshot of which is shown in Figure 1, detects and installs the necessary prerequisites automatically and eventually invokes Ansible to perform the bulk of the work, unless a pre-published container or VM is selected. Figure 2 provides a schematic view of the LaMachine architecture.

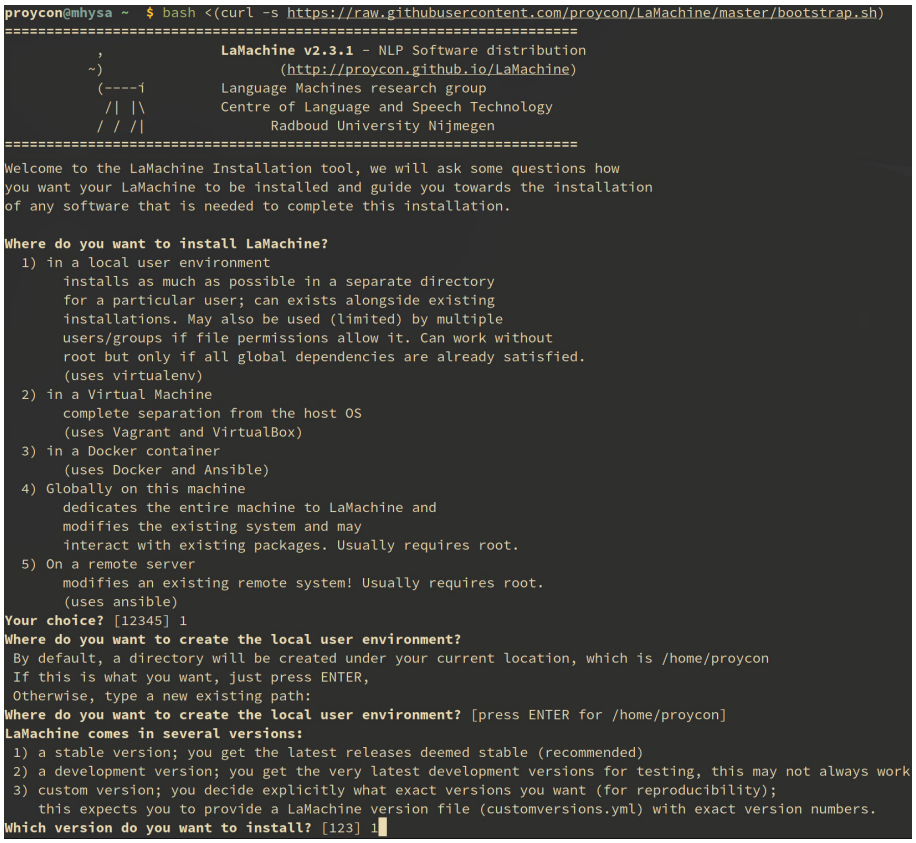


Figure 1: A screenshot of the bootstrap procedure

Once LaMachine is installed, in any of its flavours, it can be updated from inside by running `lamachine-update`. This updates all of the software managed by LaMachine.

4.4 Metadata

LaMachine aims to harmonise the metadata of all installed software. This is accomplished by converting metadata from upstream repositories, i.e. the repositories where tool providers deposit their software, to a common yet simple standard called CodeMeta [30] (Jones et al., 2016; Boettiger, 2017) where possible, or encouraging software developers to provide their codemeta metadata inside their source code repositories and using that directly. CodeMeta is a linked data initiative that provides a mapping from/to various commonly used software metadata standards[31]. All this metadata in LaMachine in turn enables other

[29] See `https://proycon.github.io/LaMachine`

[30] `https://codemeta.github.io/`, described in JSON-LD

[31] such as DOAP, Github API, Debian packages, Python distutils, R packages, Ruby gems, Maven metadata, DataCite, WikiData

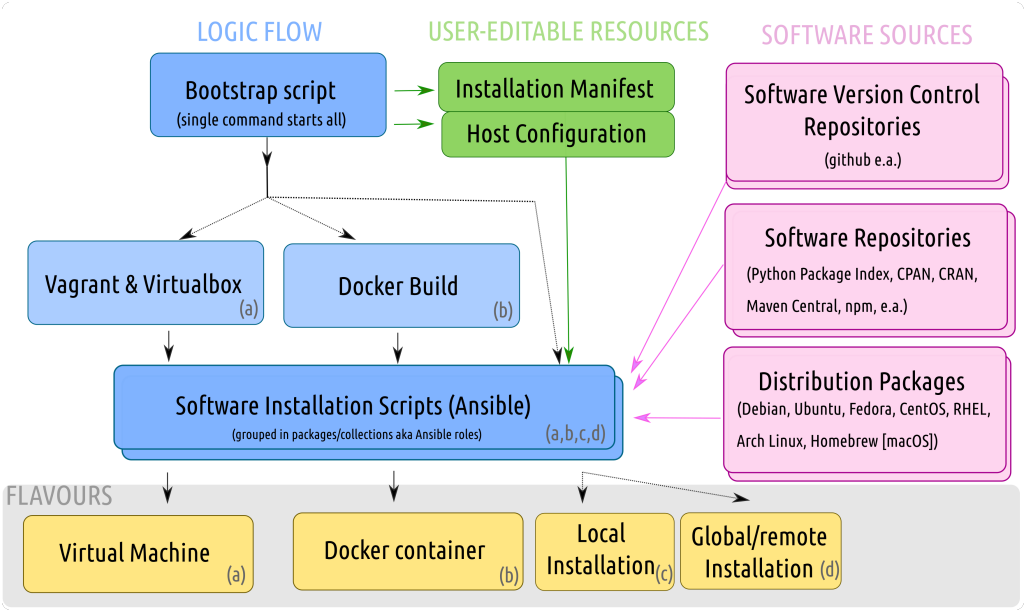


Figure 2: A schematic representation of the LaMachine architecture

tools to do proper service discovery and provenance logging.

The software metadata discussed here is of a generic and practical nature, it most importantly contains information regarding the version, source, licensing, and authors of the software. The underlying principle is to obtain sofware metadata from as close to the source as possible. This can be contrasted with other metadata efforts, also within CLARIN, to establish a more centralised metadata store[32] with more specialised and research-oriented metadata. There is merit to both approaches as they serve distinct functions. But in order to best serve everybody's different metadata needs, these bottom-up and top-down processes need to come to some kind of synthesis, e.g. by having component registries do automatic harvesting of CodeMeta (or other) metadata, similar to what LaMachine does.

5 Interfaces and Audiences

We already addressed the need for different interfaces for different audiences in Section 1. The challenge we face, and for which LaMachine offers a solution, is one of software maintainability, distribution and deployment. How do we maintain, distribute, and deploy software that is often highly complex and consists of multiple interconnected components, considering that this software is used differently by different audiences? A key aspect here is the reusability and accessibility of individual software components. The philosophy we subscribe to encourages the development and distribution of software in a layered or modular fashion, allowing each building block to serve as the foundation for another more high-level interface, without sacrificing the usability of the foundation as such. This is in line with the UNIX philosophy of developing tools that do *"one thing only, and do it well"* and is contrasted with monolithic software solutions that are limiting because they either provide only high-level interfaces but do not expose their inner components to build upon, or they are as a swiss-army knife full of needless but inseparable components.

5.1 Low-level Interfaces

In any of its flavours, LaMachine offers low-level shell access, i.e. a command-line interface accessed through a terminal. In flavours that are separated from the host system by a network, this is accomplished over `ssh`. After accessing the LaMachine environment through a terminal, as shown in Figure 3, the user has the liberty to do whatever she wants and can invoke any of the tools that offer a text interface, including text editors such as vim, emacs or nano, version control systems such as git and interpreters such as Python or R (or compilers for that matter). This allow users to use one of the many specialised

[32] Such as CLARIN's CMDI Component Registry; `https://www.clarin.eu/content/component-metadata`

programming libraries included in LaMachine to build their own tools. All this makes LaMachine ideally suited as a development environment.

```
proycon@mhysa ~  $ source lamachine-activate
================================================================================
          ,             LaMachine v2.4.0 - Unified NLP Software distribution
        ~)    '                (https://proycon.github.io/LaMachine)
         (----)
          /| |\         Centre for Language and Speech Technology
         / / /|                  Radboud University Nijmegen
================================================================================
 Build Name: dev
 Version: development      Build time: 2018-08-31 12:13:49 CEST
 Maintainer: proycon@mhysa.anaproy.nl

Welcome to LaMachine!
  - run lamachine-list to see a verbose list of all installed software
     add -s for a short sorted view
  - run lamachine-add to add extra software collections to this installation
     add the --list flag to see a list of installable packages
  - run lamachine-update to update and test your LaMachine installation
     add the --edit flag to edit settings and/or the installation manifest directly
  - run lamachine-start-webserver to (re)start the webserver

proycon@mhysa ~ (dev) $
```

Figure 3: Activation of a local LaMachine environment on the command-line

5.2 High-level Interfaces: Web-based access

LaMachine comes with a webserver[33]. This webserver serves various web-capable tools that are incorporated in LaMachine. One of these tools is a portal website that provides an overview of all installed software, and acts as a point of access to all its web services and web applications. This portal, as shown in Figure 4, leverages the metadata registry compiled in each LaMachine installation (see section 4.4). A live example of the portal to our own LaMachine installation on our production server is publicly available[34].

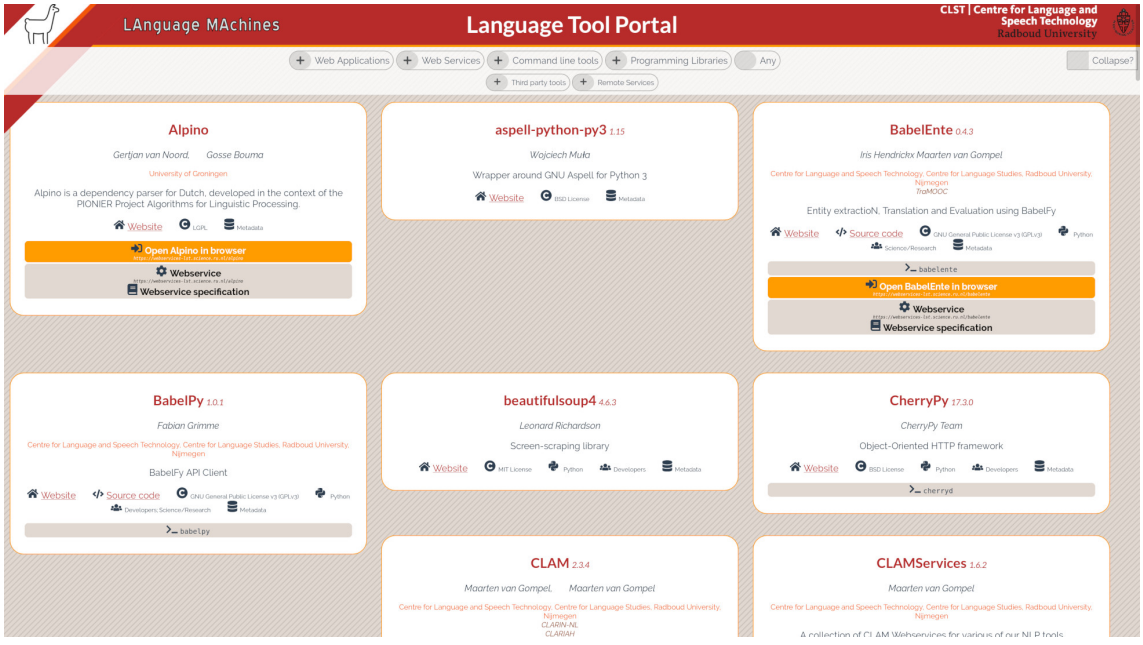


Figure 4: A screenshot of the LaMachine portal, powered by Labirinto: `https://github.com/proycon/labirinto`

LaMachine also comes with a Jupyter Lab[35] installation which provides a web-based Integrated De-

[33] In situations where web interfaces are not needed or desired, the user has the ability to opt-out of this.

[34] `https://webservices-lst.science.ru.nl`, note that we expose only certain high-level interfaces so this is just demonstrates one facet of LaMachine rather than the full LaMachine experience

[35] `https://jupyter.org/`

velopment Environment (IDE) for scripting in Python and R, web-based terminal access, and so-called *notebooks* which mix text, code and data output and have gained great popularity in the data science community. This is shown in Figure 5. This type of virtual laboratory provides a powerful interface that provides an alternative to regular shell access and may have a larger appeal for those parts of the audience that do not feel completely comfortable in the terminal.

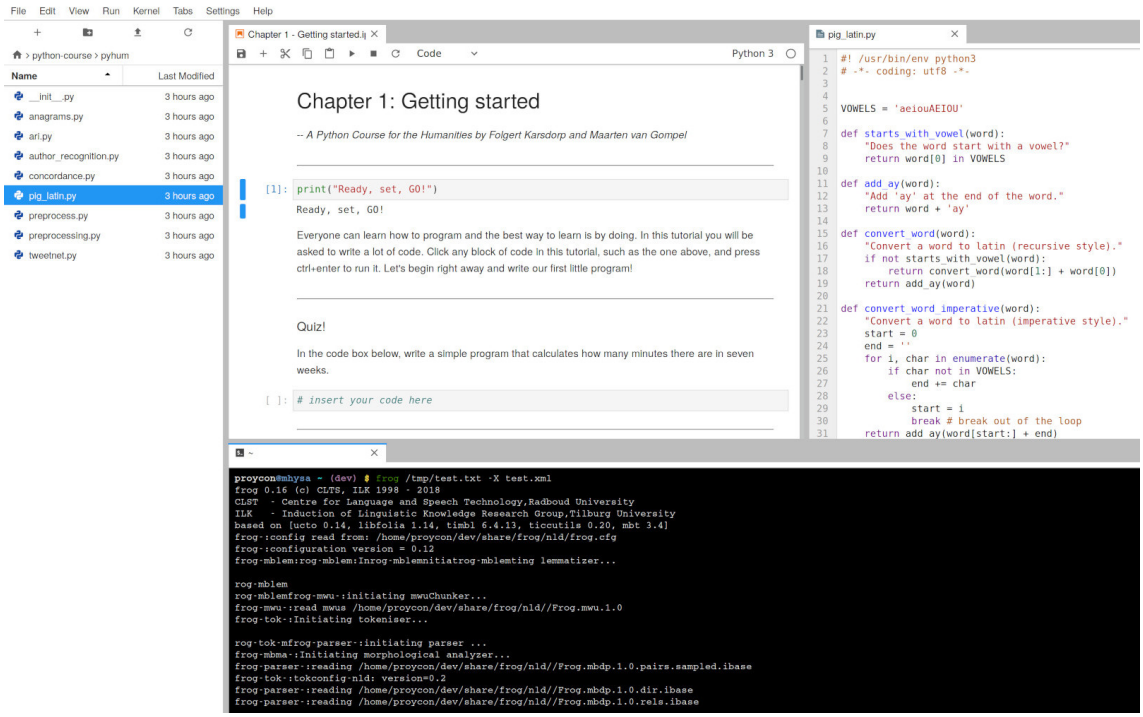


Figure 5: A screenshot of the Jupyter Lab Environment

LaMachine includes various CLAM-based webservices. CLAM is a long-time CLARIN project that allows developers to turn their command-line NLP tool into a RESTful webservice with relatively little effort (van Gompel and Reynaert, 2014), and which in addition also provides a generic web-interface for human end-users. The portal website in LaMachine links to these. This again addresses an audience that does not require a high degree of technical expertise, but is suitable for a wider audience of researchers, including those in the Humanities.

6 Comparison to alternative distributions

How does LaMachine compare to alternatives? To answer this question we must first look at what may be considered alternatives to LaMachine. At this point, it is important to emphasise once more that LaMachine is offered as convenience. LaMachine always seeks to use installation channels that the user can also access directly. For instance, is the user only interested in a single Python package that LaMachine offers? Then she can just as well install it directly using `pip install`. Is the user only interested in a single package that the underlying Linux distribution (or macOS's homebrew) already provides? Then, by all means, she should go for that.

LaMachine starts to fill a need when the user is interested in complex software, such as complex NLP pipelines with multiple components, the individual installation (sometimes involving compilation from source) and configuration of which would not be trivial. Moreover, LaMachine shines when the user wants higher-level interfaces on top of that, something she can use right out of the box. It also fulfills a role in case the user simply does not know what she wants or needs yet, and needs an environment where she can explore a variety of pre-installed tools and experiment.

Within the scope of CLARIN, one might be inclined to draw comparisons with the CLARIN Language Resource Switchboard (Zinn, 2016), which offers access to NLP tools through a web interface where users can upload their texts for NLP processing, providing easy access and automated tool discov-

ery. This, however, is notably different from what LaMachine is. LaMachine is a software distribution; the CLARIN switchboard is not. Vice versa, LaMachine does not currently provide any switchboard functionality, it merely offers a portal interface. The switchboard allows access to tools hosted elsewhere, LaMachine installs and configures tools and only provides access to those it installed. Unlike the switchboard, the audience for LaMachine is in the first place the data scientist and researcher who has programming experience and wants to work with complex NLP software, on her own machine and unencumbered by high-level interfaces if she so desires.

There are also other CLARIN initiatives that provide not just a switchboard, but aim at an even more ambitious high-level virtual research environment (VRE) with graphical interfaces to accommodate the researcher and allow execution and visualisation of various NLP pipeline. The Dutch CLARIAH WP3 project[36] is such an initiative: LaMachine is currently developed in close collaboration with this project, in a role of technology supplier.

Comparisons could be drawn between LaMachine as a kind of Linux distribution and other Linux distributions. That too, however, would not be appropriate as we characterise LaMachine more as a meta-distribution which simply builds on the foundation of several existing major Linux distributions. As a meta-distribution that runs in a variety of flavours and on a variety of platforms, it offers a great deal of flexibility that normal distributions do not have. At the same time, the scope of LaMachine is more narrowly defined than that of a generic Linux distribution, including even specialised Linux distributions that focus on scientific computing, such as Fedora Scientific[37].

The best comparison can perhaps be drawn with Anaconda[38], which we could also qualify as a meta-distribution under our definition, and which enjoys great popularity in the data science community. Anaconda has a focus on Python and R. It is much wider in scope than LaMachine, which has a strong NLP and CLARIN/CLARIAH focus. When it comes to Python, which Anaconda was initially aimed at, there is considerable overlap in the modules that LaMachine and Anaconda offer. Unlike LaMachine, Anaconda does not focus on providing different flavours such as a VM or a Docker Container[39], nor separate stable and development versions of the included software.

At the start of the development cycle for LaMachine v2, we investigated whether Anaconda and its ecosystem would provide a sufficient foundation for us to build upon. We concluded this was not the case for a number of reasons: Anaconda introduces more overhead than we desired for our purposes and conflicted with certain technology choices we made. We wanted our native flavours (local environment and global environment) to be as close to the underlying distribution as possible and to reuse existing technologies and packages, such as compilers (e.g. gcc) and interpreters (e.g. python), from the distribution (or from homebrew for macOS). Anaconda, on the other hand, choses to provide its own packages and builds on those. It does so using its own package manager (conda) and package repositories (conda-forge). This would require repackaging certain software for the anaconda ecosystem.

For the distribution and deployment of complex software setups such as NLP pipelines, containers are a common solution nowadays. The Docker flavour of LaMachine provides something similar, but rather than providing a single static Docker recipe that builds a single kind of container, LaMachine offers a high degree of flexibility for the construction of different containers, i.e. containing different software or having different configurations, based on the user's needs. It is quite feasible to instantiate a variety of LaMachine containers and use a container orchestration system such Kubernetes or Docker Swarm for automated deployment, scaling and management thereof. Such orchestration, however, is beyond the scope of LaMachine itself, which aims to provide a singular environment, i.e. everything installed in a LaMachine instance shares the same userspace and duplication is strongly minimised.

The flexibility LaMachine offers, with various flavours and versions, makes it more accessible and deployable in multiple contexts, but this does come at a cost. Maintaining LaMachine itself is a non-

[36] `https://github.com/meertensinstituut/clariah-wp3-vre`

[37] `https://labs.fedoraproject.org/en/scientific/`

[38] `https://www.anaconda.com/`

[39] This is not a technical limitation but simply a matter of different objectives, one could easily create a VM or a container and install Anaconda in it.

trivial task that requires constant maintenance and testing[40]; software exists in an ever moving ecosystem. Any of the distributions we target, or software that participates, *might* at any time introduce a change that requires us to adapt LaMachine accordingly. Similarly, underlying Linux distribution releases that are currently up-to-date and supported, will eventually be deprecated and replaced by newer releases, a development LaMachine is obliged to follow by design.

7 Case studies

LaMachine is most often used by a researcher or data scientist who installs LaMachine on his own computer (in a local environment, Docker container or VM). Here we discuss two case studies where LaMachine was used under circumstances that go beyond this most common usage of LaMachine: using LaMachine on a secure server without internet connection, and using LaMachine to provide a work environment for a group of researchers. These case studies illustrate the advantages and limitations of LaMachine in practice.

7.1 Text Mining for Health Inspection

We participated in a small Dutch national project titled *"Text mining for Inspection: an exploratory study on automatic analysis of health care complaints"* [41] led by IQhealthcare[42], the scientific centre for healthcare quality of RadboudUMC hospital. This project took place at the Dutch Health Inspectorate and aimed to apply text mining techniques to health care complaints that have been registered at the national contact point for health care (Landelijk Meldpunt Zorg[43]) We investigated the usefulness of text mining to categorise and cluster complaints, to automatically determine the severity of incoming complaints, to extract patterns and to identify risk cases. This project turned out to be a good test case of the applicability and usefulness of LaMachine as a standalone research environment. As the complaint data is highly sensitive, it could not leave the secure servers of the health inspectorate and was stored in an environment without internet access. We needed to bring the software to the data via a shared folder.

We used a virtual machine (VM) image of LaMachine and we ran this 64-bits Linux-based VM inside another VM with Windows Server 2012, provided to us by the health inspectorate for this project, in which we did have administrative rights but no internet access. In terms of hardware we ran on a machine with 8 cores and 32GB internal memory available. Note that as we had no internet access available on the working server, we prepared a custom-made VM image of LaMachine that included an additional editor and some extra Python libraries. LaMachine provided a fully functional research environment and we ran all our experiments within LaMachine. We interacted with LaMachine both through the command line, which offers a standard shell and enables access to all lower-level tools and programming languages; and through the (offline) webbrowser to use the Jupyter Notebook environment.

LaMachine comes with some simple data sharing facilities that allowed us to access the sensitive complaint data via a single shared dataspace between host and the VM. Extensive data search and management functions are deliberately beyond the scope of LaMachine, and left to more high-level tooling.

We used many of the available tools in LaMachine within this project: Frog for linguistic annotation of the textual content of the complaint and the scikit-learn Python package for classification, T-scan for feature extraction in the form of text characteristics and colibri-core for n-gram analysis.

7.2 Workshop: Cataloguing of Textual Cultural Heritage Objects

The ICT-Research Platform Netherlands and NWO organise a yearly one-week workshop 'ICT with Industry' [44] to stimulate collaboration between industry and academia. The industrial partner provides a problem and a team of researchers from different backgrounds and universities collaborate to come up with solutions. We participated in the 2019 edition on the case study by the Dutch Royal Library

[40] We run automated tested on a continuous integration platform to this end
[41] `https://bit.ly/2N2AICS`
[42] `http://www.iqhealthcare.nl/nl/`
[43] `https://www.landelijkmeldpuntzorg.nl`
[44] https://ict-research.nl/ict-with-industry/ictwi2019/

who wanted to investigate automatic methods for cataloguing of textual cultural heritage objects, in this particular case a large collection of digital dissertations.

For this workshop, computing power was purchased at SURFsara, a collaborative organization for ICT in Dutch education and research. Subsequently, we had the ability to create Virtual Machines on their hosting platform. For this workshop we had twelve participants and decided to create a single multi-core and high memory VM to share amongst the participants to create one common digital workspace.

As recommended by SURFsara, we opted for one of their default Linux images as a basis for the VM, based on Ubuntu 18.04, instead of providing a LaMachine VM image directly like we did in 7. This was recommended because their image was already preconfigured to integrate nicely with their cloud environment. Inside this VM we simply bootstrapped the local installation flavour of LaMachine. Here we benefit from the flexibility LaMachine offers because of its various flavours, and regardless of the flavour, the resulting installations are always functionally equivalent. The local environment flavour had the added bonus of not being in anyone's way in case any of the twelve workshop participants did not want to use LaMachine, considering it needs to be explicitly activated prior to usage.

LaMachine offered a convenient platform for a range of different explorations and experiments in the area of NLP and text mining. However, for some situations LaMachine, or rather Linux in general, was not a good fit for the audience of the workshop: for team members who did not have experience with a non-Windows environment, LaMachine was not a suitable or useful tool. The limit of LaMachine was also reached for members who wanted to use desktop text editors with a graphical user interface as this is not offered by LaMachine. Moreover, we did not manage to get X-forwarding working in the Ubuntu Linux VM and after a few attempts the team gave up on resolving this issue due to time pressure. This, also demonstrates that fine-tuning the configuration of certain aspects of LaMachine, but especially beyond LaMachine, is beyond the reach of a data scientist without system administration skills. This certainly also applies also to the installation as a whole in the SURFsara context, which involved things like the partitioning, formatting and mounting of (virtual) drives and setting up user accounts on the shared VM, all of which require some system administration skills and are too context-specific to be within the scope of LaMachine. LaMachine was convenient and speeded up writing code as the most common scientific data-related packages are already present in LaMachine.

8 Conclusion & Future work

LaMachine provides a flexible solution for the installation of a variety of NLP software, resulting in a kind of virtual laboratory that, through various interfaces, can be employed by a variety of people, from developers and data scientists to the wider research community that CLARIN explicitly addresses.

Most users of LaMachine are installing the software distribution in their own machine, either locally, via Docker or VM. However, we described two case studies in section 7 where we explore the possibilities under more constrained, complex, and demanding circumstances, which contributes to thorough testing and running of LaMachine.

Aside from the incorporation of new relevant software, the main objectives for the future are to provide greater *interoperability* between the included tools through better *high-level interfaces* for the researcher. We see this as a bottom-up process and have now established a firm foundation to build upon. Note that such proposed interfaces, including the current portal application in LaMachine, are always considered separate independent software projects, which may be deployed by/in/for LaMachine, but also in other contexts. LaMachine remains 'just' a software distribution at heart.

An important part of our future focus will be on interoperability with the higher-level tools emerging from the CLARIAH WP3 VRE effort, but also with other parts of the CLARIN infrastructure; single-sign on authentication being a notable example here.

LaMachine v2 is open to outside contribution. Contributor documentation has been written, and at this stage, we greatly welcome external participants to join in.

Acknowledgements

This research was funded by NWO CLARIN-NL, CLARIAH and the ZonMw project *Tekstmining in het toezicht: een exploratieve studie naar de automatische verwerking van klachten ingediend bij het Landelijk Meldpunt Zorg*, project number 516004614. We thank the Dutch Health Inspectorate, IQhealthcare, and Tim Voets for their valuable contributions and help in the ZonMw project, and NWO, the ICT Research Platform Nederland, the Lorentz Centre, the Royal Library and all our team members for making the ICT with Industry workshop a success.

References

[Boettiger2017] C. Boettiger. 2017. Generating CodeMeta Metadata for R packages. *The Journal of Open Source Software*, 2:454.

[Jones et al.2016] MB. Jones, C. Boettiger, A. Cabunoc Mayes, A. Smith, P. Slaughter, K. Niemeyer, Y. Gil, M. Fenner, K. Nowak, M. Hahnel, et al. 2016. CodeMeta: an exchange schema for software metadata. *KNB Data Repository*.

[van Gompel and Reynaert2013] M. van Gompel and M. Reynaert. 2013. FoLiA: A practical XML format for linguistic annotation - a descriptive and comparative study. *Computational Linguistics in the Netherlands Journal*, 3.

[van Gompel and Reynaert2014] M. van Gompel and M. Reynaert. 2014. CLAM: Quickly deploy nlp commandline tools on the web. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, pages 71–75. Dublin City University and Association for Computational Linguistics.

[van Gompel et al.2016] M. van Gompel, J. Noordzij, R. de Valk, and A. Scharnhorst. 2016. Guidelines for Software Quality. CLARIAH Task 54.100.

[Zinn2016] C. Zinn. 2016. The CLARIN Language Resource Switchboard. *Proceedings of the CLARIN Annual Conference. CLARIN ERIC*.

SVALA: Annotation of Second-Language Learner Text Based on Mostly Automatic Alignment of Parallel Corpora

Mats Wirén
Department of Linguistics
Stockholm University, Sweden
mats.wiren@ling.su.se

Arild Matsson
Språkbanken, Department of Swedish
University of Gothenburg, Sweden
arild.matsson@gu.se

Dan Rosén
Språkbanken, Department of Swedish
University of Gothenburg, Sweden
dan.rosen@svenska.gu.se

Elena Volodina
Språkbanken, Department of Swedish
University of Gothenburg, Sweden
elena.volodina@svenska.gu.se

Abstract

Annotation of second-language learner text is a cumbersome manual task which in turn requires interpretation to postulate the intended meaning of the learner's language. This paper describes SVALA, a tool which separates the logical steps in this process while providing rich visual support for each of them. The first step is to pseudonymize the learner text to fulfil the legal and ethical requirements for a distributable learner corpus. The second step is to correct the text, which is carried out in the simplest possible way by text editing. During the editing, SVALA automatically maintains a parallel corpus with alignments between words in the learner source text and corrected text, while the annotator may repair inconsistent word alignments. Finally, the actual labelling of the corrections (the postulated errors) is performed. We describe the objectives, design and workflow of SVALA, and our plans for further development.

1 Introduction

Corpus annotation, whether manual or automatic, is typically performed in a pipeline that includes tokenization, morphosyntactic tagging, lemmatization and syntactic parsing. Because of the deviations from the standard language, however, learner data puts special demands on annotation. Automatic tools trained on the language of native speakers can sometimes be applied with more or less satisfactory results even to learner language. Where available, spell and grammar checking tools providing suggestions can be used to approximate a corrected version of the text. More commonly, however, an additional manual step is added before applying a standard annotation pipeline, namely, normalization, which means changing the original learner text to a grammatically correct version. A problem with this is that there is seldom a unique grammatically correct version, and, related to this, that the agreement between different annotators is often low. For this reason, Lüdeling (2008) argues for making the corrections — in other words, construction of the *target hypotheses* — explicit so as to factor the problem of correction and the ensuing step, which is the actual error annotation, that is, labelling the type of correction that has been made to the learner text according to a taxonomy.

The aim of this paper is to describe SVALA[1], a tool for pseudonymization[2], normalization and correction annotation of learner texts. The work is part of the SweLL project (Volodina et al., 2018), whose

[1]"Svala" in Swedish means "swallow" (noun, in the sense of the bird), but is here an acronym which is a concatenation of SVA (*SV*enska som *A*ndraspråk; in English: *Swedish as a Second Language*) and LA (*L*inking and *A*nnotation).

[2]Pseudonymization of personal data means that it can be restored to its original state (re-identified) by keeping keys that are stored securely. In contrast, anonymization means that no such keys are kept, and that the data can therefore never be restored (unanonymized). For more discussion of this in the context of the project, see Megyesi et al. (2018).

Mats Wirén, Arild Matsson, Dan Rosén, and Elena Volodina 2019. SVALA: Annotation of Second-Language Learner Text Based on Mostly Automatic Alignment of Parallel Corpora. *Selected papers from the CLARIN Annual Conference 2018*. Linköping Electronic Conference Proceedings 159: 227–239.

aim is to create a platform for collecting, digitizing, pseudonymizing, normalizing and annotating learner texts, and to use this to construct a corpus of 600 texts written by learners of Swedish to be made available for research.[3] In addition, our intent is to make the methods and tools from the project as generally applicable as possible. To this end, SVALA is free software under the MIT license.[4]

In setting out this work, our objective was to arrive at a tool with the following characteristics:

1. *Uniform environment.* The tool should provide a single environment for pseudonymization, normalization and correction annotation, with a uniform representation of data to avoid problems with conversions between different formats.

2. *Lightweight data format.* The data format should be easy to work with programmatically and to convert to formats used by other systems.

3. *Intuitive interface.* From the point of view of usability, the tool should primarily target researchers in Learner Corpus Research (LCR) and Second Language Acquisition (SLA). In particular, the interface should have separate components that mirror the conceptual tasks of the problem (Ellis, 1994; Granger, 2008, page 266): detecting sensitive information for the purpose of pseudonymization, detecting learner deviations from the standard language, correcting the deviations (normalization), and finally annotating them.[5] It should be possible for an annotator to perform the different types of tasks separately, and it should even be possible to assign the tasks to annotators with different skills. The tool should provide rich visual support for the different tasks, specifically:

   (a) *Target hypotheses via text editing.* The annotator should be able to construct the target hypotheses in the simplest possible way by using text editing.

   (b) *Parallel corpus.* To make the differences between the layers of the text (pseudonymization, normalization, correction annotation) fully explicit, they should be visualized and represented as word-aligned texts in a parallel corpus (see Figure 1). The word alignments should be constructed automatically from the editing operations, but it should be possible for the annotator to correct them.

4. *Administration and scheduling.* The tool should be designed so as to make it straightforward to link it to facilities for administration and scheduling of the work of multiple annotators, as well as statistical analyses of the results of their work, including inter-annotator agreement.

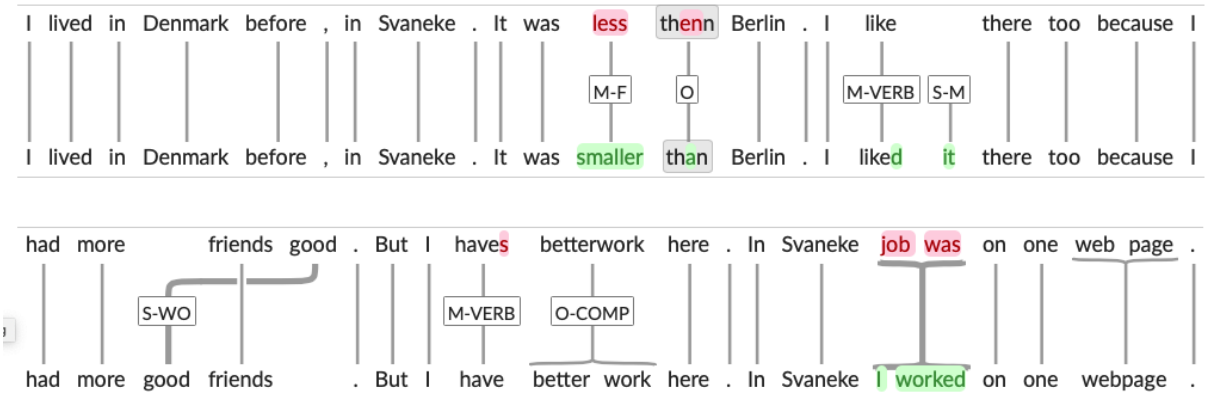


Figure 1: Example learner sentence (top rows), corrected version representing the target hypotheses (bottom rows), and correction annotation on the edges aligning the words (middle areas).

[3] https://spraakbanken.gu.se/eng/swell_infra.

[4] https://github.com/spraakbanken/swell-editor.

[5] Deviations may be viewed as errors relative to the standard language (Ellis, 1994; Granger, 2008). However, we annotate the corrections rather than the postulated errors, which is a slightly different notion (see Section 3.5).

The rest of this paper is organized as follows: We provide an overview of related work (Section 2), introduce the way in which the system works from the user perspective (Section 3), describe the technical solutions and data format (Section 4), and discuss what we have achieved and our plans for future work (Section 5).

2 Related work

Annotation of second-language learner texts is typically carried out using a purpose-built editing tool (or a pipeline of tools) with a hierarchical set of error categories (Granger, 2008), for example, the XML editor Oxygen in the ASK project (Tenfjord et al., 2006) and the TEITOK editor in the COPLE2 project (Mendes et al., 2016). However, the conceptually different tasks referred to in item 3 in the list in Section 1 are not always clearly separated. For example, in ASK (Tenfjord et al., 2006), the Oxygen editor used is customized with pop-up menus that reflect the chosen set of error categories. Specifically, annotation of an error involves inserting an XML *sic* tag with a *type* attribute encoding the error category, a *desc* attribute encoding the error subcategory, and a *corr* attribute encoding the correct (normalised) form of the token or tokens. By running a set of XSLT scripts, the XML file can be translated to an HTML format that can be viewed in a web browser, by which the annotation and corrections can be proofread.

In contrast, there are three learner corpora for which normalization and error annotation have been separated. The first two are Falko, a learner corpus for German (Reznicek et al., 2012), and MERLIN, a learner corpus for Czech, German and Italian (Boyd et al., 2014). Both of these use tools that support annotation in a tabular stand-off format with separate layers for transcription, target hypotheses and error tags. Furthermore, they both use the guidelines developed for Falko, in which two types of target hypotheses are distinguished: the Minimal Target Hypothesis, in which corrections are restricted to orthography, morphology and syntax, and the Extended Target Hypothesis, in which semantics, pragmatics and style are corrected. These hypotheses are representend in different layers in the stand-off format.

The third is CzeSL, a learner corpus for Czech (Rosen et al., 2014), for which the graphical annotation tool Feat has been used (Hana et al., 2012). This again includes two levels of target hypotheses, though with different scopes compared to Falko and MERLIN. At the lower level, only the orthography of isolated tokens is corrected, whereas at the higher level, ”the grammatical and lexical aspects of the learner’s language in a narrow sense” are corrected (Rosen et al., 2014, page 75). However, this level may include corrections of words that do not fit the context, and may thus affect semantics; in this sense it goes a bit further than the Minimal Target Hypothesis in Falko and MERLIN.[6] Feat displays three levels: the tokenized text (called level 0) and the two levels of target hypothesis (levels 1 and 2), with edges between the words at the different levels. Changes are made between the levels, and the annotator can join, split, add, delete and move words, as well as insert error labels for these. Feat saves the three levels of text and annotation as three word-aligned files. The texts are downloadable in that format, but are searchable only with SeLaQ, a tool which does not display the edges between levels.[7]

As for normalization, none of the systems mentioned above let the annotator correct the learner data using text editing. There is one such tool that we are aware of, however, namely, QAWI which was used in creating the Qatar Arabic Language Bank (QALB), a corpus of Arabic text with manual corrections (Obeid et al., 2013). Here, the annotator can insert, delete and move words while correcting orthographic, lexical and syntactic errors. No parallel corpus is constructed, however.

There has recently been some work on manual construction of parallel treebanks for learner and normalized texts. This work has not made use of any special-purpose (graphical) tools, however. For example, Berzak et al. (2016) added part-of-speech tags and dependency analyses to CoNLL-based textual templates for 5,124 learner sentences from the Cambridge First Certificate in English (FCE) corpus. By then adding dependency analyses and error codes to a corrected version of each of these sentences, they obtained a parallel treebank based on the pairs of learner and corrected sentences. Li and Lee (2018) took a somewhat similar approach for a corpus of sentences written by learners of Chinese. However,

[6] Alexandr Rosen, personal communication, 1 February 2019.
[7] Alexandr Rosen, personal communication, 1 February 2019.

their corpus also has word alignments between the learner and corrected texts and does not require strict pairing at the sentence level, thus comprising 600 learner sentences and 697 corrected sentences.

In translation and related fields, several tools for manual or semi-automatic word alignment of parallel corpora have been developed, for example, Interactive Linker (Ahrenberg et al., 2002; Merkel et al., 2003), ICA (Tiedemann, 2006), the Hypal4MUST interface, adapted in 2016 from the Hypal system (Obrusník, 2012) for the MUST project (Granger and Lefer, 2018), and, for multiparallel corpora, Hierarchical Alignment Tool (Graën, 2018). An assumption in our system is that the differences between the source and target texts are relatively small, which makes it possible for an automatic aligner to rely only on orthographic differences. Such differences, for example, the longest common subsequence ratio by Melamed (1999), have previously been used for the alignment of parallel corpora, but not for the alignment of learner corpora as far as we know. Also, we are not aware of any tool which includes all the tasks of pseudonymization, normalization and correction annotation in a uniform environment.

3 Functionality and user interface of SVALA

3.1 Prestudy

Before embarking on the design of SVALA, we studied existing tools and discussed them with SLA researchers to see if there were any possibilities of re-use. First, we wished to avoid linking different tools in a pipeline, as in the MERLIN project (Boyd et al., 2014) (item 1 in Section 1). This would have required conversions between the formats, which have been reported to be a source of multiple problems, including loss of information.[8] Secondly, we wanted a tool which allowed annotators to work on the different tasks independently as far as possible, supported by visualization (item 3 in Section 1), as in CzeSL (Hana et al., 2012). Thirdly, we wanted a tool that allowed normalization of a text in a traditional text editing mode (item 4), as in QAWI (Obeid et al., 2013). Finally, to make the differences explicit (particularly word order changes), we wanted a tool that represented and visualized the different layers of the text as a parallel corpus (item 5). It should also be mentioned that we deemed the formats used by other tools for text annotation, such as Brat[9] and WebAnno (Eckart de Castilho et al., 2016) not to be suitable for the intended users. Motivated by these considerations, we decided to design our own tool to fulfil the objectives described in Section 1.

As a first step, we made a proof-of-concept implementation based on some of the desiderata in Section 1, which confirmed the suitability for SLA researchers of normalization based on text editing (Hultin, 2017).[10] We then began a full-scale implementation, specifically including the incremental construction and visualization of the different layers of the text as a parallel corpus, resulting in SVALA. The tool has then undergone several design iterations based on annotation experiments with participants from the project team, resulting in gradual improvements of the functionality.

3.2 Overview of the user interface

In the normalization and correction annotation modes, the SVALA interface includes three panels related to the texts (see Figure 2): from the top, a display of the learner text (*Source text*), an editable display of the version of the text being corrected (*Target text*), and a graphic rendering of the parallel corpus, featuring word alignments of the source and target texts as well as annotations (which we refer to as the *spaghetti area*). Initially, the target text is the same as the source text, and each word is aligned to its copy, thus forming a trivial parallel corpus. To correct a text, the annotator uses the editable text panel *Target text* in Figure 2, which works like a standard display editor. Changes (that is, editing operations) can be made in any order, and upon each change, the system immediately updates the word alignments between the two texts, thereby maintaining the parallel corpus. Figure 2 shows the effects of several changes; note, in particular, the edges corresponding to the insertion of ”it”, the move of ”good” and the

[8] Adriane Boyd (2017), presentation at the Workshop on Interoperability of L2 Resources and Tools, https://sweclarin.se/sites/sweclarin.se/files/event_atachements/L2wsh_Boyd_slides.pdf.

[9] http://brat.nlplab.org/.

[10] Originally a student project at Stockholm University supervised by Mats Wirén and Robert Östling.

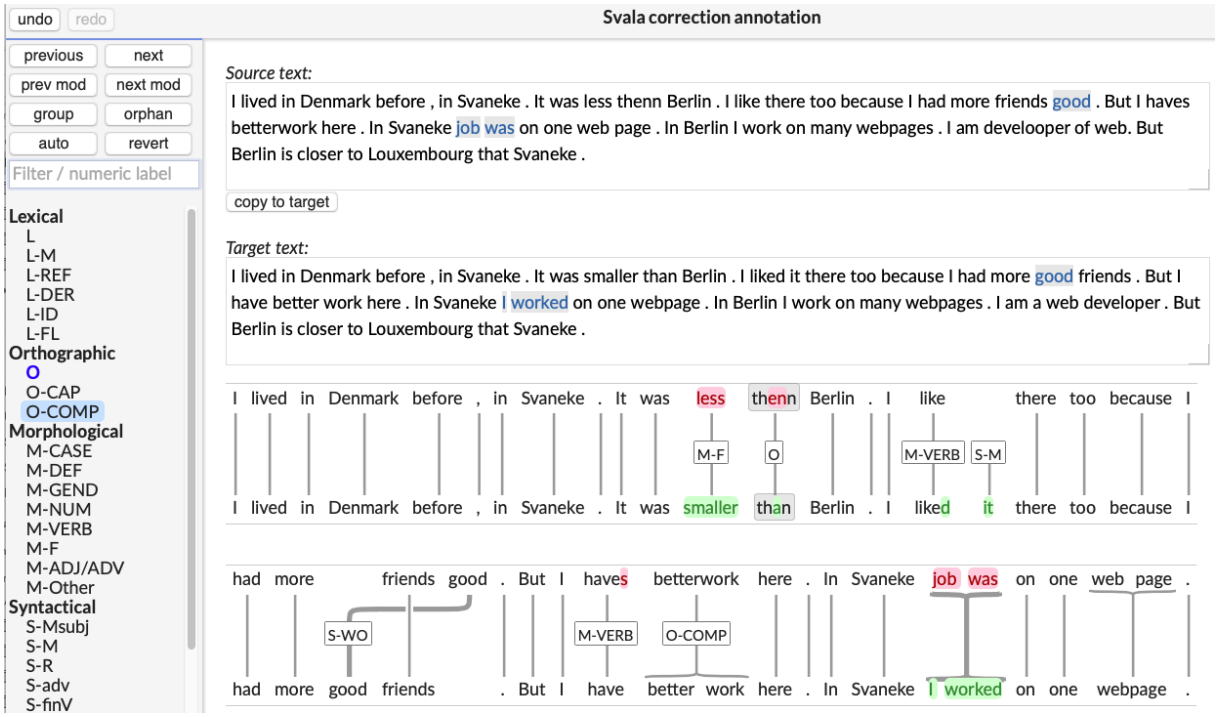


Figure 2: The SVALA graphical user interface (correction annotation mode).

replacement of ”job was” with ”I worked”. Correction annotation is made by clicking on a word or edge in the spaghetti area, and choosing a relevant label that appears in the menu to the left.

On the right of the interface, shown in Figure 4, there is a drop-down menu to switch between the different modes, open relevant guidelines, and inspect the underlying representation of the annotated data. To facilitate effective work, users can leave comments for selected sections of the text, and can browse these if the need for dicussion comes up. The comments may eventually be saved for the final version to facilitate the work for future users of the corpus.

3.3 Pseudonymization

Pseudonymization starts in a special environment that we call *Kiosk*. This is a computer with a pre-installed database and minimal support for administration of the project tasks. The Kiosk supports work on the non-pseudonymized original learner data, and is encrypted to prevent intruders from accessing learner texts. In the Kiosk, two main preprocessing steps are performed, namely, transcription and pseudonymization, the latter of which is carried out using SVALA. Once these steps have been made, hand-written and digital non-pseudonymized texts are sent to a secure storage, and the pseudonymized texts are committed to a server with which regular communication is possible.

The (sequences of) tokens that users label with pseudonymization categories are highlighted in both the *Source text* and *Target text* areas, and the pseudonyms are shown in both the spaghetti area and the *Target text* area. Edges in the spaghetti area that contain pseudonymization labels are given the status ”manually manipulated” in the data representation. This secures their highlighting in the next steps of text processing. Each pseudonymization segment which is labelled in the text is assigned a running index number, which is incremented as new entities are being pseudonymized. A display of these is provided in the area to the right; see Figure 3, where, for example, index 2 corresponds to all occurrences of the *Svaneke*.

We plan to experiment with automatic linguistic annotation of non-pseudonymized texts for two purposes: detecting candidates for pseudonymization using Named Entity Recognition (NER), and projecting morphological and grammatical information to the pseudonymized segments (for example, retaining genitive case when replacing ”John’s” with ”Adam’s”).

SVALA is pre-installed in its pseudonymization mode in the Kiosks, and works as in the usual (on-line)

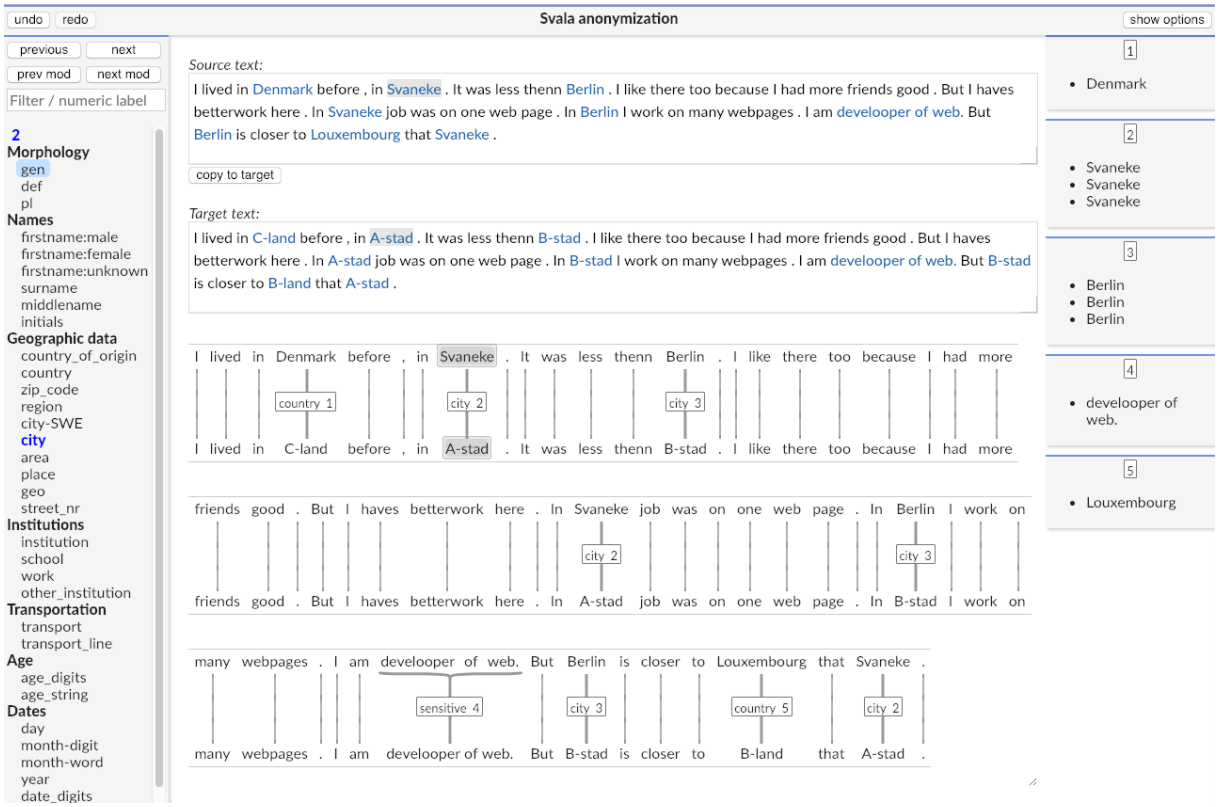


Figure 3: The SVALA pseudonymization mode.

environment in spite of the security precautions. In contrast, normalization and correction annotation are performed in the on-line version of SVALA. During the normalization step, annotators may discover errors in the pseudonymization, and because of this the pseudonymization mode is also available in the on-line version of SVALA.

A detailed description of the pseudonymization taxonomy in SVALA, the legal background of pseudonymization, and the data management in SweLL for the purpose of complying with ethical and legal demands is provided by Megyesi et al. (2018).

3.4 Normalization

The aim of normalization in SweLL is to render the text in a version which remains as close as possible to the original, but which is correct with respect to orthography, morphology and syntax, as well as semantically and stylistically coherent. To unify these incompatible requirements, the tolerance with respect to deviations from the standard language is relatively high. Correction of orthography, morphology and syntax is similar to the Minimal Target Hypothesis in the German learner corpus Falko (Reznicek et al. 2012, page 42 ff.). In contrast, the latter points go beyond this by allowing changes that affect meaning, mainly for the purpose of correcting sentences that are semantically anomalous in the context and maintaining a stylistic level which on the whole is consistent.

The purpose of normalization is twofold:

1. To render the text in a version which is amenable to automatic linguistic annotation. For Swedish, two standard pipelines for this are Sparv (Borin et al., 2016) and efselab.[11]

2. To obtain an explicit representation of the deviations relative to a postulated standard variant of the language (the target hypotheses), based on which correction annotation can be performed.

In SVALA, normalization is carried out as editing of what is initially a copy of the learner source text which appears in the *Target text* area (see Figure 4). Pseudonymized segments are highlighted to make

[11] https://github.com/robertostling/efselab.

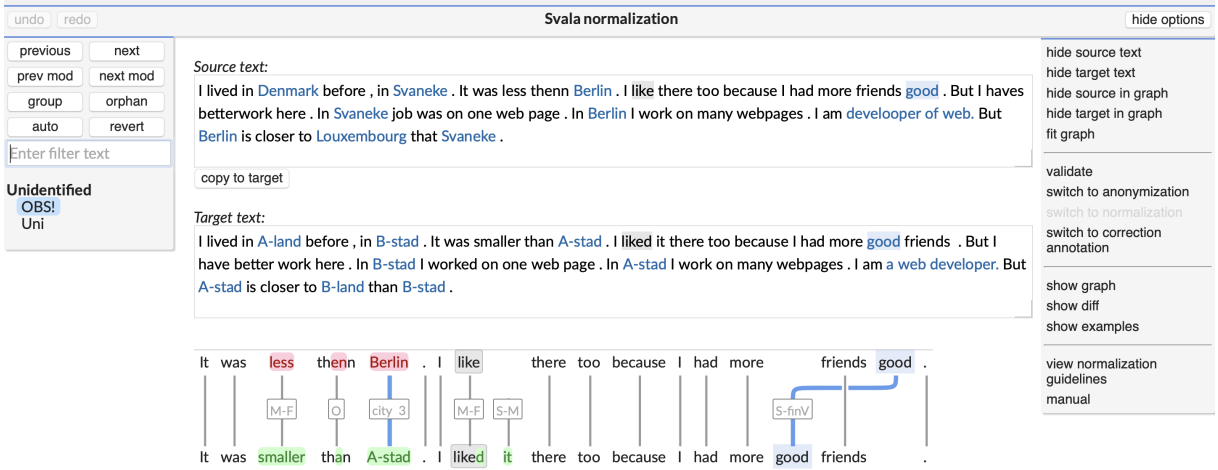


Figure 4: The SVALA normalization mode.

it clear that they were not written as such by learners. The spaghetti area displays the sentence on which the annotator is working. Differences between the source and target texts are colour-coded, creating a kind of colour map of the learner language. For usability purposes, hovering over a segment in any of the panels (source, target, spaghetti) causes the corresponding segment to be highlighted in the other areas as well, as for the word ”good” in Figure 4.

3.5 Correction annotation

By ”correction annotation” we specifically mean that we annotate the *changes* made to the target text rather than the postulated errors that occur in the learner text. The distinction is subtle, and typically there are no differences between the two notions. What it means, however, is that when a correction involves a replacement, we do not annotate errors that appear solely in the replaced material, but only the replacement as such. For example, if ”She go to the park” is replaced by ”She walks to the park”, what we annotate is the lexical replacement, but not the agreement error of ”go”, since this token does not occur in the corrected text.

Correction annotation is carried out by selecting one or several (not necessarily contiguous) tokens of the learner and corrected texts, or the edges that connect them, in the spaghetti panel. A pop-up menu is displayed to the left (see Figures 2 and 4), which contains correction labels as well as some function buttons for purposes such as grouping of tokens in the spaghetti panel. The annotator selects one or more correction labels, each of which will be displayed on the corresponding edges. The rationale for labelling the edges is that we view correction labels as relations between the learner and corrected versions of the text; because of this, the annotation is directly accessible from both. For ergonomics, and particularly for annotators who use the tool for extended periods, we provide keyboard bindings (shortcuts) for labelling and navigating between edges.

Our correction taxonomy is inspired by the error categories of ASK (Tenfjord et al., 2006), with two hierarchical levels consisting of main categories and subcategories. However, following several rounds of testing of the taxonomy in pilots, the original ASK-like taxonomy has been modified substantially. To facilitate experimentation, and to allow SVALA to be used in other projects, the error taxonomy is fully customizable.

3.6 SweLL workflow

To put SVALA into the larger context of the intended SweLL workflow, the overall text preparation and annotation steps are outlined below.

1. *Essay collection*. The original learner texts, consisting of hand-written and born-digital material, respectively, are collected. Hand-written material is scanned and transcribed.

2. *Pseudonymization*. Names, locations, institutions, etc., are annotated and automatically replaced with placeholder tokens (Megyesi et al., 2018).

3. *Normalization*. The learner text is edited for the purpose of providing a corrected version. Based on this, the system automatically constructs a parallel corpus of the learner and normalized texts, displayed in a third panel with word alignments that may need to be corrected by the annotator.

4. *Correction annotation*.

   (a) Misaligned edges between words are corrected.

   (b) Labels describing the deviations of the original text relative to the corrected version are added.

5. *Linguistic annotation*. Automatic annotation is made by way of part-of-speech tagging, lemmatization, dependency parsing, word sense disambiguation, etc. The annotation is added to the normalized version of the text, but we plan to experiment with automatic annotation of the learner version as well with the goal of obtaining a parallel treebank (Berzak et al., 2016; Li and Lee, 2018).

6. *Import to a corpus search interface.*

Work on steps 5 and 6 above have not yet begun, but for step 1, we have currently (January 2019) collected 388 essays from speakers of more than 30 native languages with different Swedish-language proficiency levels. To guide the design of SVALA, we have made several iterations (pilots) in which each time we have annotated around 30–40 essays through steps 1–4 of the workflow outlined above. Each time, we measured inter-annotator agreement (IAA) as a way of evaluating reliability and quality of the annotations, and assessed the consistency of our guidelines, correction label taxonomy and the tool functionality.

These experiments brought into focus the fact that merging normalization and correction annotation into one step produces a lot of uncertainty as to what is measured by IAA: whether annotators agree on a) which token sequence to change, and/or b) how to change it, and/or c) which correction label to assign. Similar conclusions have been drawn by Rosen et al. (2014) and Boyd (2018). As a consequence, pseudonymization, normalization and correction annotation will be assessed separately. In particular, for normalization, we need to know to what extent annotators agree on the token-level changes (target hypotheses), and for correction annotation, we need to know to what extent annotators agree on a correction label given a particular normalization.

This separation of components is also reflected in the construction of the SweLL corpus. To begin with, we allow only one master version of an pseudonymized text (see Figure 5). Thus, while pseudonymization of a text may be assigned to several assistants, only one version, the *master*, is saved for use in the successive normalization step. If it is discovered during normalization that corrections of the pseudonymization are needed, these are made in the pseudonymized master version.

The same principle is used in the normalization step. While several independent normalizations will be produced for the purpose of comparison, only one consensus annotation will be saved as the master version for use in correction annotation, and as part of the SweLL corpus. No further normalization changes will be possible once this version has been assigned for correction labelling.

We argue that normalization is possibly the most critical step and needs to be performed by highly competent staff with SLA training. This is also why step 4 a (correction of misaligned edges) is not done as part of the normalization, even though that is possible: the annotators need to concentrate on the text as such. As a basis of normalization, we have developed thorough guidelines and have had at least two people with an SLA profile discuss the normalization decisions for the same texts until they reach consensus. Only then each annotator works individually with normalization. Because of the amount interpretation involved, we believe that previous training in SLA theory as well as practical work with second-language learner texts reduce the subjectivity. Further discussions of a number of texts are a prerequisite for agreement on target hypotheses. This way we are also decreasing the complexity level of the correction annotation step.

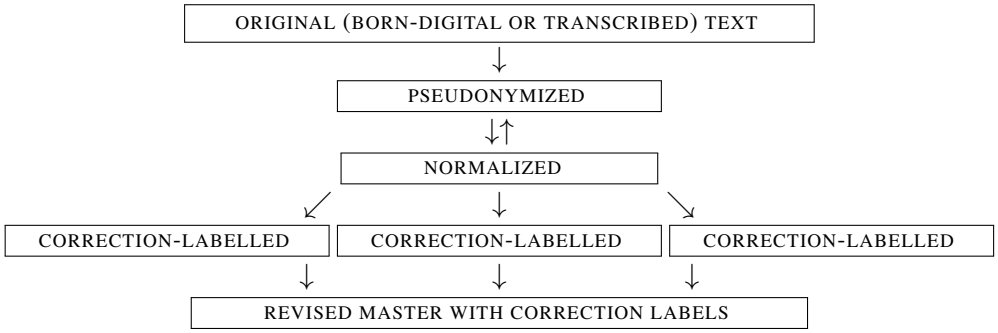


Figure 5: Versions of processed texts stored in SVALA.

Unlike in steps 2 and 3, we see a need for several versions of the same text in step 4 (correction annotation) as a prerequisite to secure consistent annotation in the whole corpus. Not all of the texts need to have several correction labelled versions, however, but for a least 20% of the corpus this will be done in order to report inter-annotator agreement.

To summarize, the above described decisions and considerations of reliability and consistency of the SweLL corpus annotation have influenced the design of the SVALA tool.

4 System design

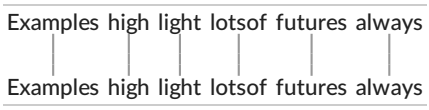
4.1 Overview

The tool is developed as a web page in TypeScript, a backwards-compatible statically typed version of JavaScript. This allows the tool to be run on all major operating systems using a browser without any installation on the user side. The interface is built using the modern and commonplace library React. The parallel corpus editing state is internally stored as a plain old JavaScript object which can be trivially serialised to and from the JavaScript object notation format JSON. Although it is conventional to use XML as a representation format in corpora, we argue that the difference between XML and JSON as a structured format is mostly superficial as there are libraries for lossless conversion between the two. Note that users with already labelled corpora can use our tool to visualize their corpus and edit it further by exporting it to the lightweight format outlined below. We are considering providing our format in a corpus conversion tool such as Pepper (Zipser and Romary, 2010).

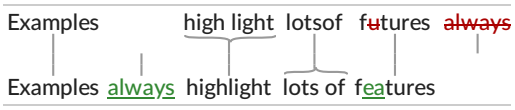
For illustration of the format, here is a small contrived example of a source (learner) text and a target text with one token each in our format:

```
{ "source": [{"id": "s0", "text": "Example "}],
  "target": [{"id": "t0", "text": "Token "}],
  "edges": {
    "e-s0-t0": {
      "id": "e-s0-t0",
      "ids": ["s0", "t0"],
      "labels": ["L"],
      "manual": false
    }
  }
}
```

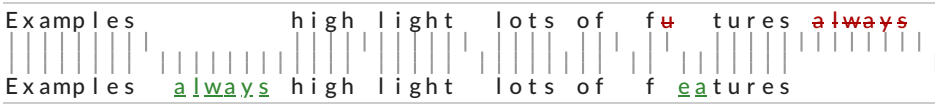
Each text is an array of tokens under their respective key (`source` and `target`). Each token consists of its text (including whitespace) and a unique identifier. Edges are stored in an unordered object since they have no order. These refer to the tokens they connect by their identifiers in the `ids` field. Annotations are put on the edges (not on tokens themselves) and each edge can have multiple labels so they are represented as an array of strings. Here the user has used the label `L`. The `manual` field is used to tell the linker to not consider it for automatic linking. This is explained in Section 4.3. Edges are strictly speaking multi-edges since they may link multiple tokens together (not just two). Edges have an identifier which is by default derived automatically from its token identifiers. This is not strictly needed as nothing in the format ever refers to this identifier, and can be considered an implementation convenience (as well as it



(a) State before editing.



(b) State after editing.



(c) Character based diff: this is lifted to the token level in Figure 6b. (not shown to the user)

Figure 6: Before and after editing. The automatic aligner gets most words right and the user will later manually connect the unlinked words.

being reflected twice). The rest of this section describes how the mostly automatic alignment procedure works.

4.2 Alignment procedure

Our system builds a corrected version aligned with the learner text using the tokens, the groups of edges between these, and any labels attached to the groups. How is this alignment calculated? We start with a standard diff edit script on the *character level*. Internally, the representation in Figure 6c is generated, which is calculated using Myers's diff algorithm (Myers, 1986) provided by the diff-match-patch[12] library. Each character is associated with the token it originates from. Next, these character-level alignments are lifted to the token level. Spaces are not used for alignment to avoid giving rise to too many false positives. We can now read off from this representation which tokens should be aligned. For each pair of matching characters, we add an edge to their corresponding tokens. For example, since there is an edge between the *h* in *high* and *highlight*, these two words are linked. Furthermore, there is an edge between the *l* in *light* to this target word too, so all these three words should be linked. There are no other characters linked to characters from these tokens, so exactly these three will become a group. The other tokens are connected analogously.

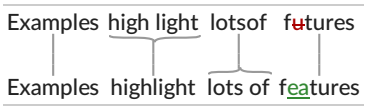
4.3 Manual alignments: word order changes and inconsistent edges

In Figure 6b, the two occurrences of the word *always* are not aligned. The user can correct this error by selecting both occurrences of *always* and clicking the *group* button (not shown here). After this grouping we are in a state where the parallel structure has one manual alignment pertaining to the word *always*, with all other words being candidates for automatic (re-)alignment. To (re-)align these we carry out the same procedure as before but excluding the manually aligned *always*: We first *remove* manually aligned words, align the rest of the text automatically (see Figure 7a), and then *insert* the manually aligned words again in their correct position (Figure 7b). Here the correct position is where they were removed from the respective texts. The editor indicates that this edge is manually aligned by colouring it blue and (for the purposes of this article) making it dotted. These edges interact with automatically aligned (grey) edges differently. How this works is explained in the next section.

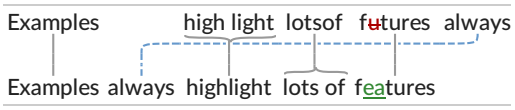
4.4 Editing in the presence of both manual and automatic alignments

The tokens that have been manually aligned are remembered by the editor. The user may now go on editing the target hypothesis. Things are straightforward as long as the edit takes place wholly in either an automatically aligned passage or in a manually aligned passage. When editing across these boundaries, the manual segment is contagious in the sense that it extends as much as needed. For example, if we select *always highlight* in Figure 7b and replace it with *alwayXighligt* the state becomes as shown in Figure 7c. However, if we cross the boundary of *of features* in the same starting state of Figure 7b to *oXeatures* we get the state of Figure 7d. Here the edit is not contagious: the automatic alignment decided to not make

[12] `https://github.com/google/diff-match-patch`.

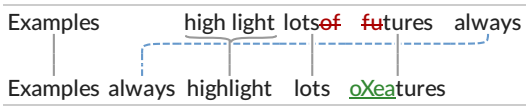


(a) Diff without *always*.



(b) Diff with *always* correctly inserted from Figure 7a.

(c) Editing from Figure 7b between *always* and *highlight*.



(d) Editing from Figure 7b between *of* and *features*.

Figure 7: Aligning and editing in the presence of a manually aligned word (here *always*).

a big component, instead it chose to split to align the words independently. In case the manual aligner has absorbed too much, our editor provides a way of removing the manual alignment tag. The editor will then fall back to the automatic aligner for those tokens.

5 Discussion

We have described SVALA, a tool for pseudonymization, normalization and correction annotation of second-language learner text. In the SVALA interface and workflow, these tasks are separate but interdependent, with rich visual support for each of them. Normalization is carried out in the simplest possible way by text editing, in the course of which the tool automatically maintains a parallel corpus with alignments between words in the learner source text and corrected text.

SVALA is being used in the SweLL project, which aims at constructing a corpus of 600 texts from learners of Swedish. However, given the relative lack of established infrastructure for research in second-language learning in CLARIN, it is also targeted as a general utility. To this end, SVALA is free software under the MIT license, and the taxonomy of correction labels is customizable. Alternatively, users of corpora that are annotated with different labels could export their format to the lightweight format of SVALA in order to make use of the system.

Several avenues for further development of SVALA are possible. In particular, the notion of parallel corpus is useful for several reasons. First, it would be straightforward to extend SVALA with additional levels of target hypotheses, for example, a purely orthographic level as in the Feat tool (Hana et al., 2012) or something akin to the Extended Target Hypothesis in Falko (Reznicek et al., 2012, page 51). Secondly, the parallel corpus is relevant for automatic linguistic annotation (step 5 in the SweLL workflow as outlined in Section 3.6). Although annotation of the corrected text is likely to be most straightforward, the alignments of the parallel corpus open up the possibility of using annotation transfer to project relevant labels (also) to the learner source text. Thirdly, we expect that a parallel corpus of learner and corrected texts will be an independently valuable resource for later development of automatic tools for identification of learner errors.

As for correction annotation, it is possible to infer some of this automatically from the editing operations and the differences between the learner and corrected texts, for example, changes of orthography, missing or redundant words, and word order. Specifically, in an agglutinating language the system might suggest relevant morphosyntactic labels upon detecting changes in suffixes for definiteness, number or tense. We experimented with automatic correction annotation in the pilot system by Hultin (2017), and the Feat tool (Rosen et al., 2014, Section 5.5) provides it for several cases.

We have argued that SVALA, with its intuitive interface, uniform environment, lightweight data format and mostly automatic word alignment, provides a rational tool and methodology for pseudonymization, normalization and correction annotation of second-language learner text. By virtue of this, we also expect it to provide a useful starting-point for the construction of parallel treebanks for learner language, in which each of the learner and normalized texts are linguistically analysed.

Acknowledgements

This work has been supported by Riksbankens Jubileumsfond through the SweLL project (grant IN16-0464:1). We want to thank the three reviewers for helpful comments, and Robert Östling, Markus Forsberg, Lars Borin and our co-workers in the SweLL project for valuable feedback.

References

[Ahrenberg et al.2002] Lars Ahrenberg, Mikael Andersson, and Magnus Merkel. 2002. A system for incremental and interactive word linking. In *LREC'02*, pages 485–490.

[Berzak et al.2016] Yevgeni Berzak, Jessica Kenney, Carolyn Spadine, Jing Xian Wang, Lucia Lam, Keiko Sophie Mori, Sebastian Garza, and Boris Katz. 2016. Universal Dependencies for Learner English. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 737–746, Berlin, Germany. Association for Computational Linguistics.

[Borin et al.2016] Lars Borin, Markus Forsberg, Martin Hammarstedt, Dan Rosén, Roland Schäfer, and Anne Schumacher. 2016. Sparv: Språkbanken's corpus annotation pipeline infrastructure. In *SLTC 2016. The Sixth Swedish Language Technology Conference, Umeå University, 17-18 November, 2016*.

[Boyd et al.2014] Adriane Boyd, Jirka Hana, Lionel Nicolas, Detmar Meurers, Katrin Wisniewski, Andrea Abel, Karin Schöne, Barbora Štindlová, and Chiara Vettori. 2014. The MERLIN corpus: Learner Language and the CEFR. In *LREC'14*, Reykjavik, Iceland. European Language Resources Association (ELRA).

[Boyd2018] Adriane Boyd. 2018. Normalization in Context: Inter-Annotator Agreement for Meaning-Based Target Hypothesis Annotation. In *Proceedings of Natural Language Processing for Computer-Assisted Language Learning (NLP4CALL)*, Stockholm, Sweden.

[Eckart de Castilho et al.2016] Richard Eckart de Castilho, Éva Mújdricza-Maydt, Seid Muhie Yimam, Silvana Hartmann, Iryna Gurevych, Anette Frank, and Chris Biemann. 2016. A web-based tool for the integrated annotation of semantic and syntactic structures. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 76–84. The COLING 2016 Organizing Committee.

[Ellis1994] Rod Ellis. 1994. *The Study of Second Language Acquisition*. Oxford University Press, Oxford.

[Granger and Lefer2018] Sylviane Granger and Marie-Aude Lefer. 2018. The Translation-oriented Annotation System: A tripartite annotation system for translation research. In *International Symposium on Parallel Corpora (ECETT — PaCor)*, pages 61–63. Instituto Universitario de Lenguas Modernas y Traductores, Facultad de Filología, Universidad Complutense de Madrid, Spain.

[Granger2008] Sylviane Granger. 2008. Learner corpora. In Anke Lüdeling and Merja Kytö, editors, *Corpus Linguistics. An International Handbook*, volume 1, chapter 15, pages 259–275. Mouton de Gruyter, Berlin.

[Graën2018] Johannes Graën. 2018. *Exploiting Alignment in Multiparallel Corpora for Applications in Linguistics and Language Learning*. Ph.D. thesis, University of Zurich.

[Hana et al.2012] Jirka Hana, Alexandr Rosen, Barbora Štindlová, and Petr Jäger. 2012. Building a learner corpus. In *LREC'12*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

[Hultin2017] Felix Hultin. 2017. Correct-Annotator: An Annotation Tool for Learner Corpora. CLARIN Annual Conference 2017 in Budapest, Hungary.

[Li and Lee2018] Keying Li and John Lee. 2018. L1–L2 Parallel Treebank of Learner Chinese: Overused and Underused Syntactic Structures. In *Proceedings of the 11th Language Resources and Evaluation Conference*, Miyazaki, Japan. European Language Resource Association.

[Lüdeling2008] Anke Lüdeling. 2008. Mehrdeutigkeiten und Kategorisierung: Probleme bei der Annotation von Lernerkorpora. In M. Walter and P. Grommes, editors, *Fortgeschrittene Lernervaritäten: Korpuslinguistik und Zweitspracherwerbsforschung*, pages 119–140. Max Niemeyer, Tübingen, Germany.

[Megyesi et al.2018] Beáta Megyesi, Lena Granstedt, Sofia Johansson, Dan Rosén, Carl-Johan Schenström, Gunlög Sundberg, Mats Wirén, and Elena Volodina. 2018. Learner Corpus Anonymization in the Age of GDPR: Insights from the Creation of a Learner Corpus of Swedish. In *Proceedings of Natural Language Processing for Computer-Assisted Language Learning (NLP4CALL)*, Stockholm, Sweden.

[Melamed1999] I. Dan Melamed. 1999. Bitext maps and alignment via pattern recognition. *Computational Linguistics*, 25(1):107–130, March.

[Mendes et al.2016] Amália Mendes, Sandra Antunes, Maarten Janssen, and Anabela Gonçalves. 2016. The COPLE2 corpus: a learner corpus for Portuguese. In *LREC'16*.

[Merkel et al.2003] Magnus Merkel, Michael Petterstedt, and Lars Ahrenberg. 2003. Interactive word alignment for corpus linguistics. In *Proc. Corpus Linguistics 2003*.

[Myers1986] Eugene W. Myers. 1986. An $O(ND)$ difference algorithm and its variations. *Algorithmica*, 1(1):251–266.

[Obeid et al.2013] Ossama Obeid, Wajdi Zaghouani, Behrang Mohit, Nizar Habash, Kemal Oflazer, and Nadi Tomeh. 2013. A Web-based Annotation Framework For Large-Scale Text Correction. In *The Companion Volume of the Proceedings of IJCNLP 2013: System Demonstrations*, pages 1–4. Asian Federation of Natural Language Processing.

[Obrusník2012] Adam Obrusník. 2012. A hybrid approach to parallel text alignment. Masaryk University, Faculty of Arts, Department of English and American Studies, Brno, Czech Republic. Bachelor's Diploma Thesis.

[Reznicek et al.2012] M. Reznicek, A. Lüdeling, C. Krummes, and F. Schwantuschke. 2012. Das Falko-Handbuch. Korpusaufbau und Annotationen Version 2.0. Humboldt-Universität zu Berlin, Berlin, Germany.

[Rosen et al.2014] Alexandr Rosen, Jirka Hana, Barbora Štindlová, and Anna Feldman. 2014. Evaluating and automating the annotation of a learner corpus. *Lang. Resour. Eval.*, 48(1):65–92, March.

[Tenfjord et al.2006] Kari Tenfjord, Paul Meurer, and Knut Hofland. 2006. The ASK corpus: A language learner corpus of Norwegian as a second language. In *LREC'06*, pages 1821–1824.

[Tiedemann2006] Jörg Tiedemann. 2006. ISA & ICA—two web interfaces for interactive alignment of bitexts. In *LREC 2006*.

[Volodina et al.2018] Elena Volodina, Lena Granstedt, Beáta Megyesi, Julia Prentice, Dan Rosén, Carl-Johan Schenström, Gunlög Sundberg, and Mats Wirén. 2018. Annotation of learner corpora: first SweLL insights. In *Abstracts of the Swedish Language Technology Conference (SLTC) 2018*, Stockholm, Sweden.

[Zipser and Romary2010] Florian Zipser and Laurent Romary. 2010. A model oriented approach to the mapping of annotation formats using standards. In *Workshop on Language Resource and Language Technology Standards, LREC 2010*, La Valette, Malta, May.

Common Language Resources and Technology Infrastructure

Linköping Electronic Conference Proceedings 159
eISSN 1650-3740 (Online) • ISSN 1650-3686 (Print) 2019
ISBN 978-91-7685-034-3

Front cover illustration:
picture composition by CLARIN ERIC