Literary Exploration Machine A Web-Based Application for Textual Scholars

Maciej Maryl
Institute of Literary Research of the Polish Academy of Sciences
Warsaw, Poland
maciej.maryl@ibl.waw.pl

Maciej Piasecki
Faculty of Computer Science and Management
Wrocław University of Science and Technology
maciej.piasecki@pwr.edu.pl

Tomasz Walkowiak
Faculty of Electronics

Wrocław University of Science and Technology
tomasz.walkowiak@pwr.edu.pl

Abstract

This paper presents a design of a web-based application for textual scholars. The goal of this project is to create a complex and stable research environment allowing scholars to upload the texts they analyse and either explore them with a suite of dedicated tools or transform them into a different format (e.g. text, table, list, spreadsheet). The latter functionality is especially important for research focusing on Polish texts (due to the rich morphology and weakly constrained word order of Polish) because it allows for their further processing with tools built for English. This project utilises the existing CLARIN-PL applications and supplements them with new functionalities.

1 Challenge

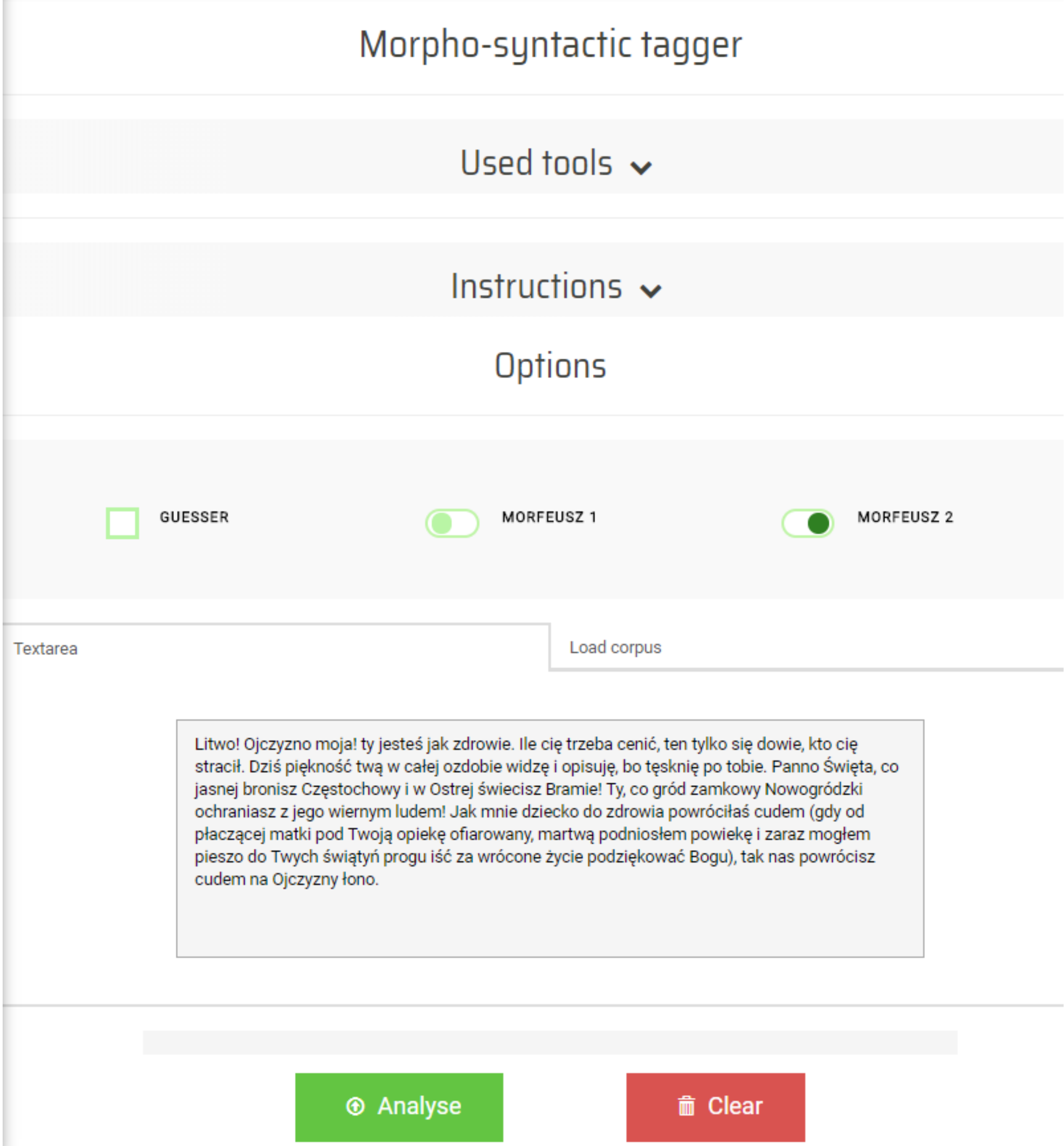
Digital literary studies seem to be among the most rapidly developing strands of Digital Humanities. The main obstacle to the development of the field lies in the users' lack of programming skills and insufficient knowledge of how to use digital methods and operate the existing tools. The authors of this paper were involved in various educational activities as organisers and instructors of CLARIN-PL workshops dedicated to endowing Polish scholars with digital methods of textual scholarship.[1] The main lesson learned from these endeavours is that although there is a genuine interest in computational literary criticism, the learning curve remains steep and workshop participants do not eventually incorporate those tools into their research workflows because they find them too complicated.

For instance, all basic language tools developed by CLARIN-PL are offered through web applications, which are seemingly easy to use from the perspective of their designers.[2] For instance, potential users of a morpho-syntactic tagger (Figure 1) need to simply upload a file to the web application to have their text tagged and morpho-syntactically disambiguated. Although the result contains complete information about word forms (lemmas, Part of Speech (henceforth PoS) and values of grammatical attributes), it is available only in an XML file, i.e. the CCL format (Marcińczuk, & Radziszewski, 2013) native for the KPWr Corpus (Broda et al., 2012). Deciphering the output may be challenging for non-professional users. For instance, let us take a look at the representation of a simple phrase *ciemny włos jej* `her dark hair' presented in Listing 1, below, where *base* tag specifies a lemma and a positional tag *ctag* encodes PoS and values of grammatical attributes.

[1] Materials from the workshops are accessible at http://clarin-pl.eu/en/mediateka-2/

[2] http://ws.clarin-pl.eu/

Selected papers from the CLARIN Annual Conference 2017, Budapest, 18–20 September 2017.
Conference Proceedings published by Linköping University Electronic Press
at www.ep.liu.se/ecp/contents.asp?issue=147. © The Author(s).



□ GUESSER ○ MORFEUSZ 1 ● MORFEUSZ 2

Litwo! Ojczyzno moja! ty jesteś jak zdrowie. Ile cię trzeba cenić, ten tylko się dowie, kto cię stracił. Dziś piękność twą w całej ozdobie widzę i opisuję, bo tęsknię po tobie. Panno Święta, co jasnej bronisz Częstochowy i w Ostrej świecisz Bramie! Ty, co gród zamkowy Nowogródzki ochraniasz z jego wiernym ludem! Jak mnie dziecko do zdrowia powróciłaś cudem (gdy od płaczącej matki pod Twoją opiekę ofiarowany, martwą podniosłem powiekę i zaraz mogłem pieszo do Twych świątyń progu iść za wrócone życie podziękować Bogu), tak nas powrócisz cudem na Ojczyzny łono.

Figure 1. Web-based User Interface for the morpho-syntactic tagger.

```
<chunkList> <chunk id="ch1" type="p">

<sentence id="s1">

<tok> <orth>Ciemne</orth>

    <lex disamb="1"> <base>ciemny</base>

    <ctag>adj:sg:nom:n:pos</ctag></lex> </tok>

<tok> <orth>włosy</orth>
```

```
        <lex disamb="1"> <base>włos</base>

        <ctag>subst:pl:nom:m3</ctag></lex>   </tok>

    <tok> <orth>jej</orth>

        <lex disamb="1"> <base>on</base>

        <ctag>ppron3:sg:gen:f:ter:akc:npraep</ctag> </lex>

    </tok>
```

Listing 1. An example of the XML-based format for the output of a tagger from Figure 1.

Despite such rich description, this output is useless for those users who are not familiar with XML or are unable to convert this input to a desirable form, e.g. a list of lemmas. Users from Humanities and Social Sciences, without the background in Computer Science, do not possess such skills, which dramatically reduces usefulness of this service as a research tool and effectively excludes it from their workflows. Furthermore, the results of the 2015 survey into digital methods and practices, conducted by the DARIAH DiMPO Working Group, seem to corroborate these observations: scholars clearly articulate a need for technological support and guidance concerning the existing tools (Dallas et al., 2017: 7).

To address these challenges we have developed a web-based system, called *Literary Exploration Machine* (LEM),[3] which:

- does not require installation,
- aggregates existing language tools for Polish,
- has a modular architecture and can be expanded through the addition of new components;
- allows for the preprocessing of texts to make them compatible with tools developed for other languages, e.g., CLUTO (Zhao & Karypis, 2005), or Mallet (McCallum, 2002),
- offers a simple workflow not requiring programming skills,
- provides elaborated descriptions of tools, outputs, and parameters, and aims at supplementing them with rich use-case descriptions.

This approach makes LEM similar to other 'one-stop-shop' initiatives which make sophisticated tools accessible to less experienced users. A good example is DARIAH-DE/Topics[4] package, which provides interface and tailored workflows for topic modeling. However, LEM offers a wider variety of tools and techniques. Popular *Voyant*[5] allows for quick analysis of the word forms and their relative frequencies across texts, supplemented by a range of NLP tools based on the Stanford CoreNLP, e.g. Proper Nouns recognition. However, *Voyant* is dedicated to texts written in English, so its applications remain limited. Nevertheless, *Voyant* visualisation methods and a friendly GUI remains a gold standard for LEM.

2 Design of the system

LEM was developed in a user-driven paradigm through interdisciplinary cooperation of computer scientists, linguists, literary scholars, and sociologists. All functionalities, options, and output formats were defined and described in a series of case studies, dedicated to particular research problems in Digital Humanities:

1. *Literary studies:*

[3] http://ws.clarin-pl.eu/lem.shtml

[4] https://github.com/DARIAH-DE/Topics

[5] Voyant: http://docs.voyant-tools.org, CoreNLP: https://nlp.stanford.edu/software/

a. research on secondary literature, i.e. academic journal articles. A comparative study of the transformation in the Polish literary scholarship 1989-2014 (Maryl 2016a).

b. Text preprocessing for advanced stylometric analysis of authorship, genre, and chronology in a corpus of novels (Rybicki, 2017).

c. analysis of an online genre and its evolution overtime on the example of a web portal (Maryl 2016b).

2. *Sociology*: multi-feature classification of teachers' attitudes on the basis of students' comments. Complex preprocessing (e.g., counting co-occurrences of selected lemmas and PoS) served as an input for statistical data analysis, e.g. (Bryda & Tomanek, 2015), cf (Brosz et al., 2017).
3. *Social psychology*: sentiment analysis in the studies of depression and emotions in Polish texts, inspired by a workflow of Linguistic Inquiry and Word Count (LIWC) (Tausczik & Pennebaker, 2010),[6] e.g. (Rohnka et al., 2015).

Each of these case studies produced the following content:

a. research questions which could be answered through a computational analysis;

b. relevant textual source material;

c. design of tools capable of answering those questions;

d. development of actual tools;

e. rich description of a tool with a hands-on guide for other scholars.

LEM is built on the processing engine WebSty,[7] an open stylometric system, which provides tools for the statistical description of text corpora, texts and subcorpora comparison. However, LEM is designed for a variety of research uses, so in addition to WebSty it allows for an analysis of other linguistic levels:

- basic description of words
  - segments (e.g., length of documents, paragraphs, and sentences),
  - morphology (word forms, punctuation marks, pseudo-suffixes, and lemmas),
- combined morpho-syntactic information
  - grammatical classes and categories (based on the Polish National Corpus tagset (Przepiórkowski et al., 2012)), as well as their n-grams,
- lexical semantics
  - proper names and their semantic categories,
  - word senses and their selected lexico-semantic relations like synonymy or hypernymy.

All of the LEM **processing paradigms** are designed to fit into this general workflow:

1. *Uploading* a corpus of documents together with their metadata (thanks to the compatibility with CLARIN repositories, Component Metadata format is supported, but also simple metadata can be read from the file names).
2. *Text extraction* and cleaning. OCR-ed documents usually contain many language errors which can be corrected at this stage.
3. *Selection of features* for the description of documents is done manually by users or available by default, depending on the processing paradigm. This step is based on the hands-on guide.

[6] https://liwc.wpengine.com/

[7] WebSty: http://websty.clarin-pl.eu/, Mallet: http://mallet.cs.umass.edu/

Users are not expected to have advanced knowledge of Natural Language Engineering or Data Mining.

4. *Setup of parameters* for processing is also customised, but some default settings of parameters are provided. More advanced users will be able to tune the tool to their needs.
5. *Text preprocessing* with language tools provided by CLARIN-PL. Each text is analysed by a PoS tagger, e.g. *WCRFT2* (Radziszewski, 2013), and eventually piped to a Name Entity Recognizer, i.e. *Liner2* (Marcińczuk et al., 2013), a temporal expression recognizer and a word-sense recogniser, e.g., *WoSeDon* (Kędzia et al., 2015), etc.
6. *Feature values calculation*. Selected features of the preprocessed texts are extracted together with their frequencies and annotations by comparing patterns defining the features with every position in a document.
7. *Filtering and/or transforming* the original feature values. Most filtering and transformation functions are provided by WebSty and its components. Further data-analysis features allowing for advanced comparison of corpora will be added.
8. *Data mining*. Several processing paradigms are employed to allow for gathering more complex information about the data, namely *topic modelling* (representing a document in terms of subsets based on word co-occurrences), *unsupervised clustering* (grouping documents on the basis of the document-feature vectors similarity, and *supervised classification* (a prototypical application of Machine Learning based on *Weka* (Witten et al., 2017),[8] *scikit-learn* (Pedregosa et al., 2017), and *SciPy*[9] packages (Jones et al., 2018), trained on documents manually classified by users).
9. *Presentation of results.* The results are presented either as interactive visualisation, or as downloadable files in formats compatible with external exploratory tools and programs (e.g., spreadsheets or *Gephi*).

3 Current functionality of LEM

From the user's perspective, the complex workflow described above could be translated into a simple, three-step procedure:

1. upload texts to be processed,
2. choose the task and its parameters,
3. browse or download the results.

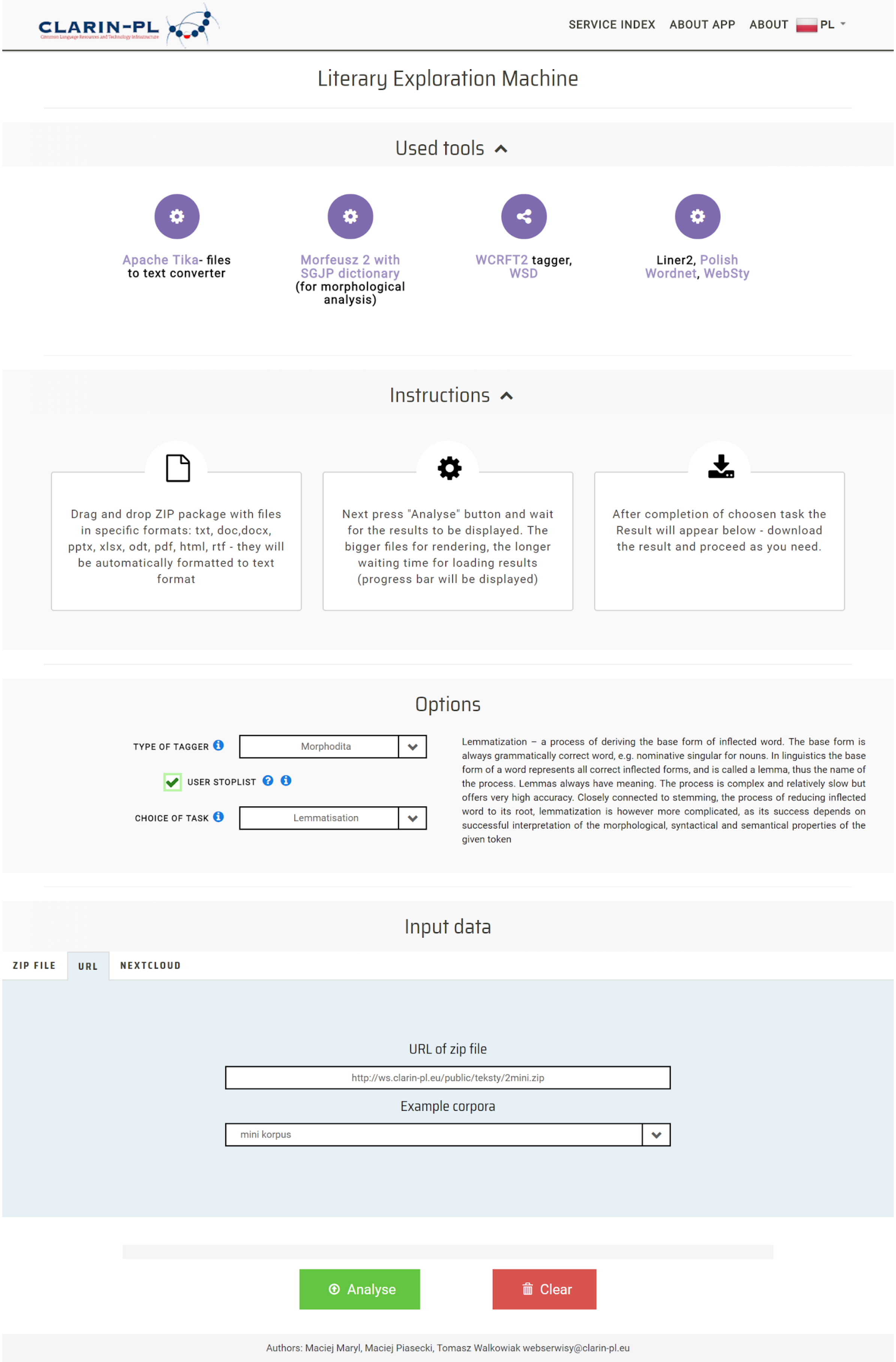
In the first step, users need to prepare a ZIP archive with texts they want to analyse. LEM accepts most of the popular formats: TXT, RTF, DOC, DOCX, ODT, XLSLX, PDF. Files could be uploaded directly from the hard drive, from the URL, or from CLARIN Cloud storage (based on the NextCloud technology and maintained by CLARIN-PL)[10]. LEM was designed for efficient processing of large volumes of data. However, the size and the number of files to be processed is limited for common users, due to processing workload and limitations of some of the output formats (e.g., XLSX). Larger datasets can be processed with the assistance of the project team.

The results of processing depend on the quality of the corpus. Optionally, users can provide a 'stoplist', i.e. a list of words or characters which should be excluded from the analyses. It is especially convenient when users want to filter out OCR mistakes or words overrepresented in the corpus. For instance, in the corpus of academic articles we used for a case study discussed below, the numerals were overrepresented because of the footnote numbers. Knowing that we could easily exclude them from further processing.

[8] Weka: http://www.cs.waikato.ac.nz/ml/weka/, scikit-learn: http://scikit-learn.org/stable/, SciPy: https://www.scipy.org, Gephi: https://gephi.org

[9] https://www.scipy.org

[10] https://nextcloud.clarin-pl.eu/



Literary Exploration Machine

Used tools

Apache Tika- files to text converter

Morfeusz 2 with SGJP dictionary (for morphological analysis)

WCRFT2 tagger, WSD

Liner2, Polish Wordnet, WebSty

Instructions

Drag and drop ZIP package with files in specific formats: txt, doc,docx, pptx, xlsx, odt, pdf, html, rtf - they will be automatically formatted to text format

Next press "Analyse" button and wait for the results to be displayed. The bigger files for rendering, the longer waiting time for loading results (progress bar will be displayed)

After completion of choosen task the Result will appear below - download the result and proceed as you need.

Options

TYPE OF TAGGER: Morphodita

☑ USER STOPLIST

CHOICE OF TASK: Lemmatisation

Lemmatization – a process of deriving the base form of inflected word. The base form is always grammatically correct word, e.g. nominative singular for nouns. In linguistics the base form of a word represents all correct inflected forms, and is called a lemma, thus the name of the process. Lemmas always have meaning. The process is complex and relatively slow but offers very high accuracy. Closely connected to stemming, the process of reducing inflected word to its root, lemmatization is however more complicated, as its success depends on successful interpretation of the morphological, syntactical and semantical properties of the given token

Input data

URL of zip file: http://ws.clarin-pl.eu/public/teksty/2mini.zip

Example corpora: mini korpus

Authors: Maciej Maryl, Maciej Piasecki, Tomasz Walkowiak webserwisy@clarin-pl.eu

Figure 2. LEM web-based User Interface – the main screen.

Users can also choose between different morphological analysers: two versions of the morphological analyser *Morfeusz*[11] (Woliński, 2014) combined with the WCRFT tager and a new MorphoDiTa-pl morphosyntactic tagger (Piasecki and Walentynowicz, 2017) including a morphological analyser. *Morfeusz 1* is the older of the two and its dictionary is smaller, and the words which do not appear in it are left unlemmatized. *Morfeusz 2*, on the other hand, has a significantly larger built-in dictionary, and provides some additional features (e.g., classification of proper names).[12] The third tagger, MorphoDiTa, was developed by CLARIN-PL and is based on neural networks. Its model is frequently modified through a constant learning process, hence the processing results for the same text may differ over time. The use of Morfeusz 1 can result in a higher accuracy of the tagger's output, Morfeusz 2 recognises more word forms on the basis of its dictionary, but the tagger may still make some mistakes when choosing the correct form, while MorphoDiTa-pl was trained already on the NCP automatically transformed to the annotation compatible with Morfeusz 2 and uses its own morphological analyser and guesser based on SGJP dictionary (Saloni et al., 2015). Researchers can select the tagger which suits their research questions or source materials best. MorphoDiTa-pl should produce the best results in most cases for the cost of lower efficiency. However, the effect of the lower efficiency can be visible only for a larger amount of text to be processed.

In the second step, users choose one of the processing tasks, which are described in the following subsections in more detail. Once the version of a morphological analyser and a task are selected, users click the "Process" button to perform an analysis and may observe the progress of the analysis on the percentage bar. When the processing is done, users can access the results through an interactive visualisation or a downloadable output file. What follows is a detailed description of tasks.

3.1 Lemmatisation

Lemmatisation is a task of converting inflected word forms – words as they appear in sentences – to *lemmas* (basic morphological forms, dictionary forms). It is closely connected to *stemming*, which was introduced in Information Retrieval as a process of reducing inflected words to their pseudo-roots (not always proper words), cf Manning et al. (2008). Lemmatisation is, however, more complicated, as its success depends on the successful interpretation of the morphological, syntactic and semantic properties of a given token. Manning et al. (2008, p. 29) provide the following example:

> "If confronted with the token *saw*, stemming might return just *s*, whereas lemmatization would attempt to return either *see* or *saw* depending on whether the use of the token was as a verb or a noun".

The task is especially significant in highly inflected languages, such as Polish that has, for instance, more than 100 possible word forms for an adjective. Proper lemmatisation is a necessary prerequisite for almost any further textual analysis, even as basic as counting word frequencies.

Lemmatisation is performed by a selected morpho-syntactic tagger, one of the three possible configurations. Word forms are either assigned by a morphological analyser or guessed. Contextually appropriate lemmas are selected, and disambiguated lemmas are assigned by the selected tagger.

For each text file from the corpus, LEM returns its lemmatized version, in which each word is replaced by its lemma, i.e., different word forms of the same meaning are replaced by the same basic morphological form representing them. Due to the reduced complexity, such files can be used as the input to many statistical tools that work on the level of words and were originally designed for English.

For instance, for the input text (from the novel *Szczęśliwa* by Eliza Orzeszkowa):

Wysoka, kształtna, z twarzą myślącą, zimną nieco, ale pięknie zarysowaną i bardzo świeżą, w stroju pełnym smaku i powagi, siedzi pod rozłożystemi drzewami wspaniałego parku i myśli o tem, jaki ten park jest piękny, jaki ten dzień letni jest pogodny i jaka ona sama jest szczęśliwa.

LEM returns:

[11] http://sgjp.pl/morfeusz/morfeusz.html.en

[12] Nevertheless the existence of the two analysers and the necessity of the choice is an additional burden put on the user that should be removed in the future by exchanging the morpho-syntactic tagger with a new one, better compatible with Morfeusz 2, e.g. a new CLARIN-PL tagger MorphoDiTa-pl with a guessing module (http://ws.clarin-pl.eu/morphoDiTa.shtml) or even a better one (in development). This transition has not been fully completed yet, due to the lower efficiency of MorphoDiTa-pl in comparison to WCRFT and on-going works in CLARIN-PL on taggers that should be even better.

wysoki , kształtny , z twarz myśleć , zimny nieco , ale pięknie zarysować i bardzo świeży , w strój pełny smak i powaga , siedzieć pod rozłożystemi drzewo wspaniały park i myśleć o tema , jaki ten park być piękny, jaki ten dzień letni być pogodny i jaki on sam być szczęśliwy .

A couple of lemmatisation errors, that can be noticed (e.g. *rozłożystemi* instead of *rozłożysty*) are caused by the archaic word forms in the original text written in the XIXth century, while the tagger was trained on the contemporary texts and is using a contemporary morphological analyser.

3.2 Part of Speech Tagging

PoS tagging refers to both manual and automatic attribution of tokens to word classes. In a simplified version, PoS tagging is taught at school under the form of attribution of word classes such as nouns, verbs, adjectives etc. to given morphological forms. As in the case of lemmatization, successful tagging depends on the correct interpretation of morphologic and syntactic properties of a word, as the same word form may represent different classes, e.g. `drink', depending on the context, can function as a noun and as a verb.

LEM encodes the PoS Tagging results with the tagset developed for the National Corpus of Polish (NCP, cf. Przepiórkowski et al., 2012). For each file in the corpus, LEM returns a CSV file with columns containing tokens, their lemmas and tags[13]. Table 1 contains an encoded excerpt from *Szczęśliwa* by Eliza Orzeszkowa.

WORD	LEMMA	PoS
Nie `not'	nie	qub
była `was'	być	praet
już `already'	już	qub
młodą `young'	młody	adj
,	,	interp
lecz `but'	lecz	conj
twarz `face'	twarz	subst
jej `her'	on	ppron3
zachowała `had kept'	zachować	praet
delikatność `delicacy'	delikatność	subst
rysów `of countenance'	rys	subst
i `and'	i	conj
cery `complexion'	cer	subst
,	,	interp
kibić `figure'	kibić	subst

Table 1. LEM Part-of-Speech tagging. English translations added by the authors.

[13] Simplified tagset table is available here: http://nkjp.pl/poliqarp/help/ense2.html

3.3 Verb characteristics

LEM allows for the further exploration of the corpus through the verb analysis. Verb characteristics feature returns numeric data about the occurrences of the verbs depending on their tense, number, person and gender, using the grammatical categories from the NCP tagset. The resulting table is delivered in an XLSX[14] file and contains the number of tokens and verbs in the corpus, together with aggregated counts for the following verb forms: infinitive; 1st, 2nd, 3rd person singular; 1st, 2nd, 3rd person plural. Table 2 contains sample results.

		SINGULAR						PLURAL					
Tokens	Verbs	1Pers	2pers	3pers	3pers_m	3pers_f	3pers_n	1Pers	2pers	3pers	3pers_m	3pers-nm	inf
11242	1299	100	100	84	151	465	0	0	0	0	0	0	150

Table 2. LEM verb form characteristics for an excerpt from Eliza Orzeszkowa's novella *Kto winien*

3.4 Lemmas and PoS statistics

Basic descriptive statistics of lemmas and PoS occurrence in the corpus is aggregated on the basis of lemmatisation and PoS tagging, described in A and B above. Resulting tables, delivered in an XLSX spreadsheet, contain absolute counts of respective lemmas and tags, together with their share in the entire corpus. Tables 3 and 4 contain sample results.

być \`to be'	201	1.7673%
który \`which/who'	97	0.8529%
to \`this'	92	0.8089%
oko\`eye'	35	0.3077%
ręka \`hand/an arm'	32	0.2814%
czy \`whether'	31	0.2726%

Table 3. LEM lemma statistics for Eliza Orzeszkowa's novella *Kto winien*

interp	2607	22.9227%
qub	673	5.9175%
conj	497	4.3700%
adv	254	2.2334%
praet:sg:f:imperf	211	1.8553%
prep:gen:nwok	189	1.6618%
subst:sg:gen:f	172	1.5124%

[14] The data here have a more complex form, and that is why we exchanged CSV to XLSX. This was done after we observed that our users had had a lot of problems with importing a CSV encoded in UTF-8 in a proper way into Microsoft Excel.

adv:pos	163	1.4332%

Table 4. LEM frequencies of the top most frequent morpho-syntactic tags in Eliza Orzeszkowa’s novella *Kto winien*

3.5 Named-Entity Recognition and Statistics

Corpora usually contain proper names, which may be relevant for a given research question (e.g. names of scholars quoted in a collection of academic papers). The Named Entity Recognition feature extracts Named Entities from the corpus and returns a sorted XSLX table with all proper names and their frequencies (See Table 5.).

NAMED ENTITY	LEMMA	FREQ
Rzym	*Rzym* `Rome’	19
Palatynie	*Palatyn* ‘Palatinus’	13
Kapitolu	*Kapitol* `Capitol’	7
Forum	*forum* `forum’	6
Konstantyna	*Konstantyn* `Constantine’	4
Koloseum	*Koloseum* `Colosseum’	3
Piotra	*Piotr* `Peter’	3
Słońce	*słońce* `Sun’	3
Via Sacra	*via sacrum*	3
Baedeker	*Baedeker*	2
Grecji	*Grecja* `Greece’	2
Kastora	*Kastor* `Castor’	2
Marka Aureljusza	*Marek aureljusza* `Marcus Aurelius’	2

Table 5. LEM Named-Entities recognized in Jerzy Żuławski’s short story *Veneri et Romae*. English translations provided by the authors.

Proper names are extracted with *Liner2*, which was built through Machine Learning[15] performed on *KPWr Corpus*, manually annotated with more than 28,000 proper name occurrences (Broda et al., 2012, Marcińczuk et al., 2016). *Liner2* can recognise the beginning and end of a proper name occurrence in text and also classify it semantically into up to 82 classes organised in a shallow hierarchy. *Liner2* expresses very high accuracy concerning the recognition of the proper name occurrences (above 90%) and very good accuracy of their classification. In the example above, one can notice that *Liner2* had problems with generating proper lemmas for multi-word proper names. Although the problem posed by proper names consisting of common words was almost solved by Marcińczuk (2017) and his solution will be implemented to LEM, recognition of multi-word proper names containing foreign words not listed in the dictionary still poses a challenge. The components of proper names do not provide clues for its internal morpho-syntactic structure, and it is hard to recognise components that should be inflected.

[15] Conditional Random Fields algorithm was used to learn contextual probabilities (depending on the left and right contexts) of tokens starting, occurring in the middle and outside proper names. Contexts are represented by more than 30 features referring to, e.g., word forms, dictionaries of characteristic words and semantic properties of words acquired from a very large wordnet of Polish, namely plWordNet (http://plwordnet.pwr.edu.pl).

For instance, *Marka Aureljusza* in Table 5 is the genitive form and was not properly processed due to the old-fashioned orthography of the second component. The only solution here would be to first determine the nominative form of a proper name on the basis of the tagger output (or its most frequent form), and then recognise its inflected forms.

3.6 Disambiguated Word-Senses and their Relations

Word-sense disambiguation (WSD) is the task of identifying the meaning of a semantically ambiguous word used in the text. Successful disambiguation is necessary for a semantic analysis of a text, especially if the ambiguities are strictly semantic (not morpho-syntactic): e.g., the word “paper” can mean either the material, a scientific article, or a newspaper.

WSD in LEM is performed with *WoSeDon* tool, which identifies the most characteristic word senses for a given text by first mapping the tokens onto the *plWordNet*[16] (Polish name: Słowosieć; Maziarz et al., 2016), a very large and comprehensive wordnet, expanded with some other connected knowledge resources like *SUMO* ontology[17] (Pease, 2011). WSD is automatically preceded with necessary lemmatisation and PoS tagging (see sections A and B above) and returns CSV files with columns containing the token, its lemma, PoS, and representation of its meaning in the form of a synset. A synset is a set of synonyms or near-synonyms, i.e. words that share the same selection of lexico-semantic relations (called constitutive relations) and, thus, possess exactly the same semantic description according to a given wordnet. Such words can be considered as semantically equivalent, and can be interchanged in certain linguistic contexts (cf. Maziarz et al., 2013).

WORD	**LEMMA**	**PoS**	**plWordNet 3.0 SYNSET**
niespokojny	*niespokojny* `uneasy’	adj	niespokojny.3(42:jak)
sen	*sen* `sleep’	subst	spoczynek.2(23:st), sen.1(23:st)

jakiejś	*jakiś* `some’	adj	jakowyś.1(42:jak); który.1(42:jak)
			jaki.1(42:jak);
			jakiś.1(42:jak) jakowy.1(42:jak);
			któryś.2(42:jak)

jednej	*jeden* `one’	adj	pewien.1(42:jak) jeden.3(42:jak)
nocy	*noc* `night’	subst	noc.2(25:czas)
jesiennej	*jesienny* `autumnal’	adj	jesienny.1(43:rel)

Table 6. LEM WSD-based analysis of Jerzy Żuławski’s *Veneri et Romae*. English translations provided by the authors.

3.7 Hypernyms & hyponyms

This feature identifies hypernyms and hyponyms of the disambiguated word senses, described in the previous section. Information about hypernyms and hyponyms sheds more light on the senses identified in the previous step and may help in the interpretation by enriching text representation with micro semantic fields (e.g., hypernyms and hyponyms can be shared by some words in the corpus).

While synonymy, the basic relation in plWordNet, is the relation of semantic equivalence, hypernymy and hyponymy entail meanings which are, respectively, more general or more specific. For instance, the hypernym of ‘plant’ is ‘organism’, while ‘flower’ is its hyponym. In plWordNet, the relations of hypernymy/hyponymy are among constitutive relations that shape a hierarchical structure of the lexicon (Piasecki et al., 2009, Maziarz et al., 2013).

[16] http://plwordnet.pwr.wroc.pl/wordnet/

[17] This also opens a possibility of collecting statistics of the concepts matching the given text.

For each file in the corpus, the feature returns a CSV file similar to the one returned by the WSD (presented in the previous section), containing columns for the token, its lemma, PoS, synset (semantic representation) and the lemma's hypernym(s)[18] and hyponym(s).

3.8 Stylometric analysis

Stylometric analysis is "the study of measurable features of (literary) style, such as sentence length, vocabulary richness and various frequencies (of words, word lengths, word forms, etc.)" (Eder and Rybicki, 2012). The best-known use of computational stylometric analysis is the authorship attribution, which provides "taxonomies of features to quantify the writing style, the so-called style markers, under different labels and criteria" (Stamatatos, 2009) to identify the author of a text. It also allows for discovering patterns in the corpus by grouping texts according to their stylistic features.

LEM provides a simplified interface to the CLARIN-PL WebSty[19] and allows for various visualizations of its results. The detailed description of the tool could be found in (Eder et al., 2017).

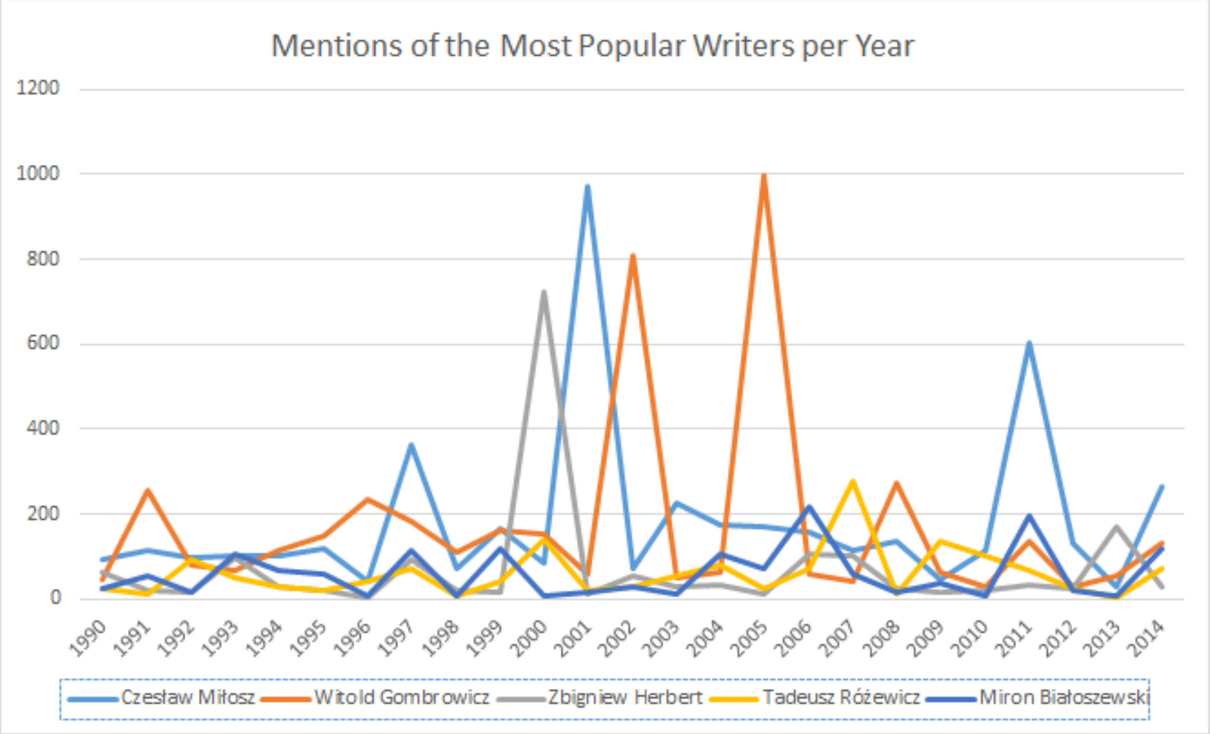


Figure 3. Patterns of interest in Polish writers based on "Teksty Drugie".

3.9 Topic modelling

This feature allows to discover of abstract `topics' that occur in a given corpus. LEM uses Latent Dirichlet Allocation (LDA) (Blei et al., 2003) method to generate word collections that are significant and specific for a given corpus on the basis of word frequency and the scope of their presence in samples. In order to control the overrepresentation of certain words in particular texts, LEM uses only lemmatized nouns that appeared at least in 80% of the documents (a fraction of the total corpus size). Users can choose the number of topics they wish to explore (the default is set at 20) and whether they want to split input files into chunks of roughly 20,000 bytes each, to ensure better processing of longer texts.

The results (i.e. words in particular topics and topic distribution across the text) could be explored through the interactive interface, or downloaded in XLSX or JSON format.

[18] As hypernymy is defined in plWordNet in a linguistic way, there can be more than one hypernym for a word sense.
19 http://ws.clarin-pl.eu/websty.shtm

4 Use Case

LEM prototype was developed by the team working with a particular textual corpus of 2,553 Polish texts, published in *Teksty Drugie*,[20] an academic journal dedicated to literary studies. The corpus consists of the two parts: OCR-ed scans (1990 – 1998) and digital-born files (1999 – 2014). Given the aim of this paper (software presentation), we will treat the results only as examples of the method, without getting into too much detail. For a more extensive interpretation of the results of *Teksty Drugie* analysis with LEM see Maryl (2016a) and Maryl and Eder (2017).

The work on the prototype was divided into several stages, conceived as feedback loops for the developing team: on every stage, a new service was added to the application, and the test run was performed. After the analysis of its result, either the step was repeated, or the team moved to the next phase.

Phase 1. The OCR-ed corpus was cleaned (e.g., word breaks and headers were removed).

Phase 2. The corpus was lemmatized and PoS-tagged. Frequency lists were created, which enabled searching for patterns in the textual output. This allowed for a simple discovery of interest patterns in the journal over time. For instance, Figure 3. shows a pattern of interest in most popular Polish writers based on the number of times their names were mentioned per year (the figure contains only the authors who reached the threshold of 1,000 total mentions).

Another temporal insight to be gained from a lemmatised corpus concerns the uptake of critical approaches in literary studies. Figure 4 presents the reception of postcolonial studies in the Polish academia on the example of mentions of words related to "postcolonial" (i.e., *postkolonializm* 'postcolonialism', *postkolonialny* 'postcolonial', *postkolonialność* 'postcolonialness', *postkolonialista* 'postcolonialist') in a literary-studies journal.

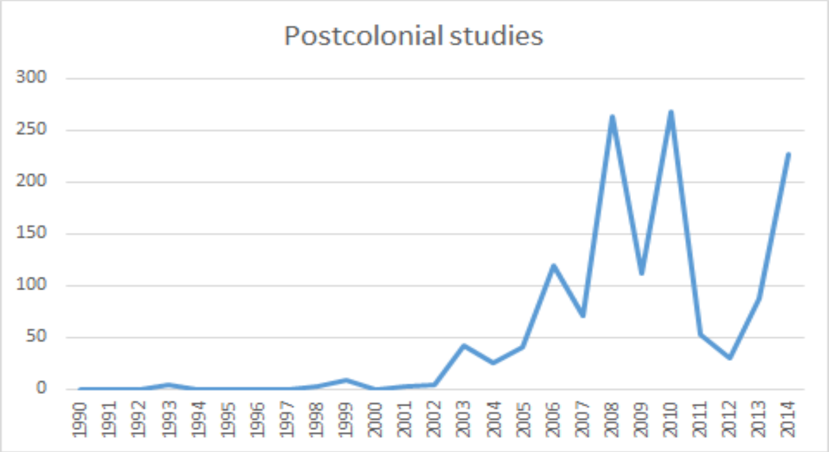


Figure 4. Mentions of words related to `postcolonial' in "Teksty Drugie" per year.

Phase 3. The analysis of word frequencies, especially the ML-based training of classifiers revealed problems with the word list, e.g., occurrences of numbers, years and city names, which were preserved in the bibliographic references. This lead to the addition of a user-defined stopword list feature. The exclusion of corpus-specific problematic words and general function words (e.g., Polish words corresponding to 'this' *to, ta, ten*, 'that' *tamto, tamta, tamten*, 'if' *jeśli, jeżeli*) allowed for the visualisation of the most frequent words in *Teksty Drugie* (Fig. 5)

[20] http://tekstydrugie.pl/en/

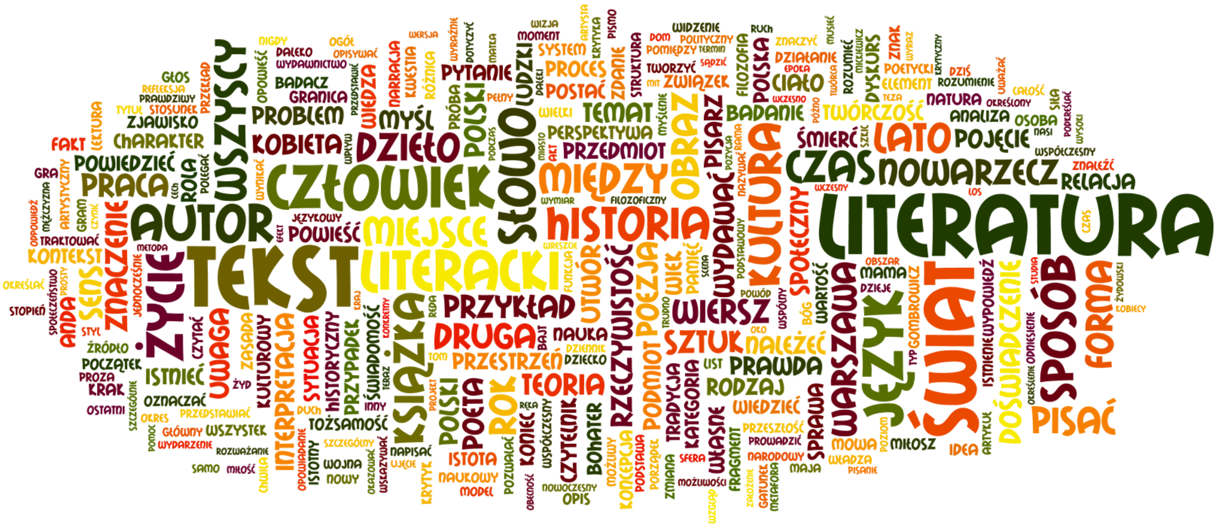


Figure 5. 300 most frequent words from *Teksty Drugie* (1990-2014) (meaningless words excluded) visualised with *Wordle* (selected glosses: autor `author', *czas* `time', *człowiek* `human', *dzieło* `creative work', *język* `language', *kultura* `culture', *literatura* `literature', *nowa* `new', *tekst* `text', *świat* `world', *wszyscy* `all', *życie* `life').

Phase 4. The texts were next grouped into clusters of 20, 50 and 100 in a series of experiments with *WebSty*. Each grouping revealed a somewhat different level of generalization about the semantic structure of the texts. By choosing the level of granularity (20, 50 or 100 clusters), one may analyse diverse patterns of discursive similarities between texts. Table 7 shows the differences in clustering of the same sample. The first option (20) shows the similarity between texts on a rather general level, which could be described as stylistic or genre similarity (e.g., formal vocabulary). Other options allow for more detailed exploration of general research approach (50) or particular topics analysed in articles (100). The semantics of clusters is described by the identified characteristic features.

Number of clusters	100	50	20
Cluster size (mean)	25.33	50.66	56.65
Cluster size (median)	24.00	47.00	51.50
Smallest cluster size	13.00	25.00	2.00
Largest cluster size	51.00	91.00	96.00

Table 7. Differences between the clustering options (numbers reflect the number of texts assigned to particular cluster)

Researchers may explore all options and analyse the vocabulary responsible for classifying particular texts into a certain group by virtue of being over- or under-represented in comparison to the entire sample.

Phase 5. The corpus was then analysed by Maryl and Eder (2017) with the 'Mallet' package for topic modelling. The resulting topics were analysed and categorised on the basis of dominant words into literary theory, poetics, methodological approaches history of literature, cross-cutting research themes, what allowed for the better visualisation and understanding of the relations between the topics. Visualisation of topics overtime allowed for a more refined results in comparison to the ones achieved in Phase 2. The results of this study served as a basis for LEM's topic modelling task.

5 Further Development

Currently, LEM's GUI is being continuously developed in cooperation with users: mostly literary scholars working on various types of texts (fiction, journal articles, blog posts), sociologists and social psychologists.

LEM prototype was fully implemented and made available as a web application[21] to the scholarly audience working on texts in Polish. Next, it will be extended with tools for other languages (e.g. English and German), in a similar way to WebSty. Thanks to LEM's modular architecture, it will require mostly linking new processing Web Services and adding converters. LEM is available on an open licence, and the authors will be happy to share their tools, code and *know-how*. Export options to other formats will be added, so researchers can easily create the output in a particular format (list, text, table) and upload it to other applications for further processing.

References

[Blei et al., 2003] Blei, D.M., Ng, A.Y. and Jordan, M.I. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*. 3 (4–5): pp. 993–1022.

[Broda et al., 2012] Broda, B., Marcińczuk, M., Maziarz, M., Radziszewski. A., and Wardyński, A. 2012. KPWr: Towards a Free Corpus of Polish. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugûr Dogân, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey, May 2012*. European Language Resources Association (ELRA)., pp. 3218–3222.

[Bryda and Tomanek, 2015] Bryda G., Tomanek K. 2015. Odkrywanie wiedzy w wypowiedziach tekstowych. Metoda budowy słownika klasyfikacyjnego. In Niedbalski J. (Ed.) *Metody i techniki odkrywania wiedzy. Narzędzia CAQDAS w procesie analizy danych jakościowych*, Wydawnictwo UŁ, pp. 51–81.

[Brosz et al., 2017] Brosz M., Bryda G., Siuda P. 2017. Od redaktorów: Big Data i CAQDAS a procedury badawcze w polu socjologii jakościowej. *Przegląd Socjologii Jakościowej* [Big Data, CAQDAS and research procedure in the field of qualitative research], t. 13, nr 2, pp. 6–23 [Access 30.01.2018, URL: www. przegladsocjologiijakosciowej.org].

[Calle-Martin and Miranda-Garcia, 2012] Calle-Martin, J. and Miranda-Garcia, A. 2012. Stylometry and Authorship Attribution: Introduction to the Special Issue. *English Studies*, 3(93): 251–258.

[Calzolari et al., 2014] Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S. (eds.) 2014. *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014. Reykjavík, Iceland*, ELRA.

[Dallas et al., 2017] Dallas, C., Chatzidiakou, N., Benardou, A., Bender, M., Berra, A., Clivaz, C., … Zebec, T. 2017. European Survey on Scholarly Practices and Digital Needs in the Arts and Humanities – Highlights Report. Zenodo.

[Eder et al., 2017] Eder, M., Piasecki, M. and Walkowiak, T. 2017. An Open Stylometric System Based on Multilevel Text Analysis. *Cognitive Studies | Études cognitives*, No. 17, https://doi.org/10.11649/cs.1430

[Eder and Rybicki, 2012] Eder, M. and Rybicki, J. 2012. Introduction to Stylometric Analysis using R. *Digital Humanities 2012 Conference. Hamburg.*

[Jones et al., 2018] Jones E, Oliphant E, Peterson P, et al. 2018. SciPy: Open Source Scientific Tools for Python. http://www.scipy.org/ [Online; accessed 2018-03-27].

[Kędzia et al., 2015] Kędzia, P., Piasecki, M. and Orlińska, M. J. 2015. Word Sense Disambiguation Based on Large Scale Polish CLARIN Heterogeneous Lexical Resources. *Cognitive Studies | Études cognitives*, (15), 269–292.

[21] http://ws.clarin-pl.eu/lem.shtml

[Manning et al., 2008] Manning, C., Prabhakar, R. and Schütze, H. 2008. *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.

[Marcińczuk, 2017] Marcińczuk, M. 2017. Lemmatization of Multi-word Common Noun Phrases and Named Entities in Polish. In (ed.) Ruslan Mitkov and Galia Angelova *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017 Varna, Bulgaria, Sep 4–6 2017*, INCOMA Ltd., pp. 483–491, https://doi.org/10.26615/978-954-452-049-6_064

[Marcińczuk et al., 2016] Marcińczuk, M., Oleksy, M., Maziarz, M., Wieczorek, J., Fikus, D., Turek, A., Wolski, M., Bernaś, T., Kocoń, J., Kędzia, P. 2016. Polish Corpus of Wrocław University of Technology 1.2, CLARIN-PL digital repository, http://hdl.handle.net/11321/270

[Marcińczuk et al., 2013] Marcińczuk, M., Kocoń, J. and Janicki, M. 2013. Liner2 – A Customizable Framework for Proper Names Recognition for Polish. In Bembenik, Robert and Skonieczny, Lukasz and Rybinski, Henryk and Kryszkiewicz, Marzena and Niezgodka, Marek, *Intelligent Tools for Building a Scientific Information Platform: Advanced Architectures and Solutions*. Springer, vol. 467, pp. 231–253.

[Marcińczuk and Radziszewski, 2013] Marcińczuk, M. & Radziszewski, A 2013. WCCL Match – A Language for Text Annotation. In Kłopotek, A., M., Koronacki, Jacek, Marciniak, Małgorzata et al (editors), *Language Processing and Intelligent Information Systems*, pages 131–144. Springer Berlin Heidelberg.

[Maryl, 2016a] Maryl, M. 2016a. Tekstów świat. Przyczynek do makroanalitycznej monografii czasopisma literaturoznawczego [World of Texts. Take on a Macroanalytical Monograph of a Scholarly Journal] In Nasiłowska, A. & Łapiński, Z. (Eds.), *Projekt na daleką metę. Prace ofiarowane Ryszardowi Nyczowi*, Warszawa: Wyd. IBL, pp. 443–462.

[Maryl, 2016b] Maryl, M. 2016b. Cyberwspólnota sądów żalu w perspektywie makroanalitycznej [Cybercommunity of regret statements in the macroanalytical perspective]. In *3rd Congress of the Polish Society for Cultural Studies, Adam Mickiewicz University of Poznań, 21–23 September 2016.*

[Maziarz et al., 2016] Maziarz, M., Piasecki, M., Rudnicka, E., Szpakowicz, S., and Kędzia, P. 2016. plWordNet 3.0 – a Comprehensive Lexical-Semantic Resource. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, The COLING 2016 Organizing Committee pp. 2259–2268, 2016, http://www.aclweb.org/anthology/C16-1213

[Maziarz et al., 2013] Maziarz, M., Piasecki, M., and Szpakowicz, S. 2013. The Chicken-and-egg Problem in Wordnet Design: Synonymy, Synsets and Constitutive Relations. *Language Resources and Evaluation*, 47(3):769–796.

[McCallum, 2002] McCallum, A.K. 2002. *MALLET: A Machine Learning for Language Toolkit*. Web page of the system. URL: http://mallet.cs.umass.edu.

[Pease, 2011] Pease, A. 2011. *Ontology: A Practical Guide*. Articulate Software Press, Angwin, CA,.

[Pedregosa et al., 2011] Pedregosa, F. and Varoquaux, G. and Gramfort, A. and Michel, V. and Thirion, B. and Grisel, O. and Blondel, M. and Prettenhofer, P. and Weiss, R. and Dubourg, V. and Vanderplas, J. and Passos, A. and Cournapeau, D. and Brucher, M. and Perrot, M. and Duchesnay, E. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12, pp.2825–2830.

[Piasecki et al., 2009] Piasecki, M., Szpakowicz, S. and Broda, B. 2009. *A WordNet from the Ground Up*. Wrocław: Oficyna Wydawnicza Politechniki Wrocławskiej, http://www.dbc.wroc.pl/dlibra/docmetadata?id=4220&from=publication

[Piasecki and Walentynowicz, 2017] Piasecki, M. and Walentynowicz, W. 2017. MorphoDiTa-based Tagger Adapted to the Polish Language Technology. In Z. Vetulani and P. Paroubek, editors, *Proceedings of Human Language Technologies as a Challenge for Computer Science and Linguistics*, pp. 377–381, Poznań, 2017. Fundacja Uniwersytetu im. Adama Mickiewicza w Poznaniu.

[Piasecki et al., 2016] Piasecki, M.; Walkowiak, T. & Eder, M. 2016. WebSty –- an Open Web-based System for Exploring Stylometric Structures in Document Collections. In Eder, M. & Rybicki, J. (Eds.) *Digital Humanities 2016 Conference Abstracts*, Jagiellonian University and Pedagogical University, 2016, pp. 859–861.

[Przepiórkowski et al., 2012] Przepiórkowski, A., Bańko, M., Górski, R. L. and Lewandowska-Tomaszczyk, B. (eds) 2012. *Narodowy Korpus Języka Polskiego*. Warszawa: PWN.

[Radziszewski, 2013] Radziszewski, A. 2013. A Tiered CRF Tagger for Polish. In Bembenik, Robert and Skonieczny, Lukasz and Rybinski, Henryk and Kryszkiewicz, Marzena and Niezgodka, Marek, *Intelligent Tools for Building a Scientific Information Platform: Advanced Architectures and Solutions*. Berlin: Springer, vol. 467, pp. 215–230.

[Rohnka et al., 2015] Rohnka, N., Szymczyk, B., Rusanowska, M., Holas, P., Krejtz, I., Nezlek, J. 2015. Właściwości języka osób cierpiących na zaburzenia emocjonalne i osobowości - analiza treści opisów codziennych wydarzeń [Language characteristics of individuals with emotional, and personality disorders: content analysis of daily events]. *Psychiatria i Psychoterapia*, Vol. 11, No. 3, pp. 3–20.

[Rybicki, 2017] Rybicki, J. 2017. Reading Novels with Statistics: What Numbers of Words Tell Us about Authorship, Genre, or Chronology. In J. A. Dobelman (Ed.) *Models and Reality: Festschrift For James Robert Thompson*, Chicago: T&NO Company, pp. 207–224.

[Saloni et al., 2015] Saloni, Z., Woliński, M. Wołosz, R., Gruszczyński, W., and Skowrońska, D. 2015. *Słownik gramatyczny języka polskiego*. [Grammatical dictionary of Polish]. SGJP, 3rd edition.

[Stamatatos, 2009] Stamatatos, E. 2009. A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for Information Science and Technology*, 3(60): 538–556.

[Tausczik and Pennebaker, 2010] Tausczik, Y.R., and Pennebaker, J.W. 2010. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, 29, 24–54.

[Witten et al., 2017] Witten, I.H., Frank, E., Hall, M.A., Pal, C.J. 2017. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufman.

[Woliński, 2014] Woliński, M. 2014. Morfeusz Reloaded. In (Calzolari et al., 2014), pages 1106–1111.

[Zhao and Karypis, 2005] Zhao, Y. and Karypis, G. 2005. Hierarchical Clustering Algorithms for Document Datasets. *Data Mining and Knowledge Discovery*, 10(2): 1.