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.

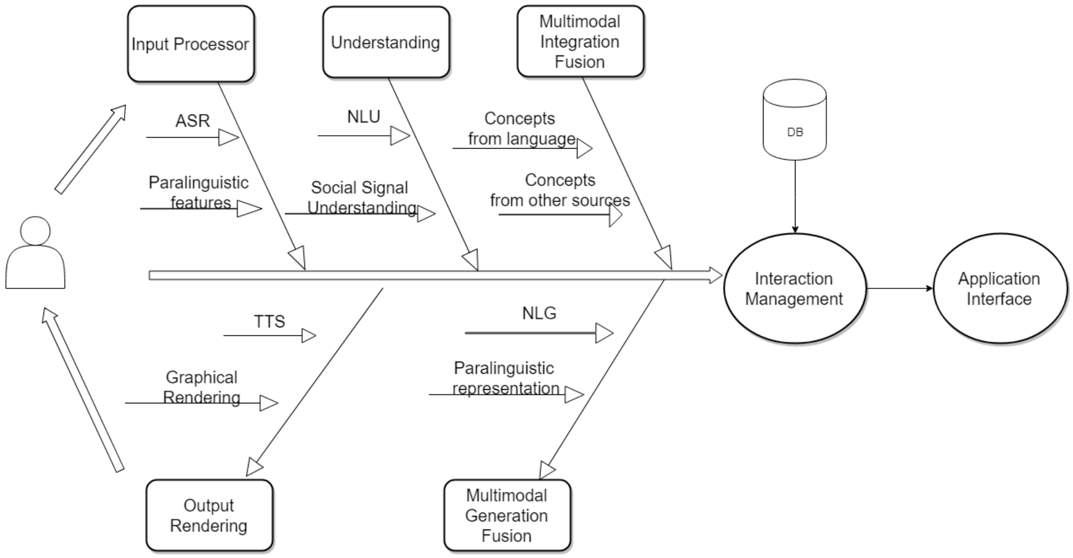


Figure 1: Multimodal Dialogue System Architecture

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.

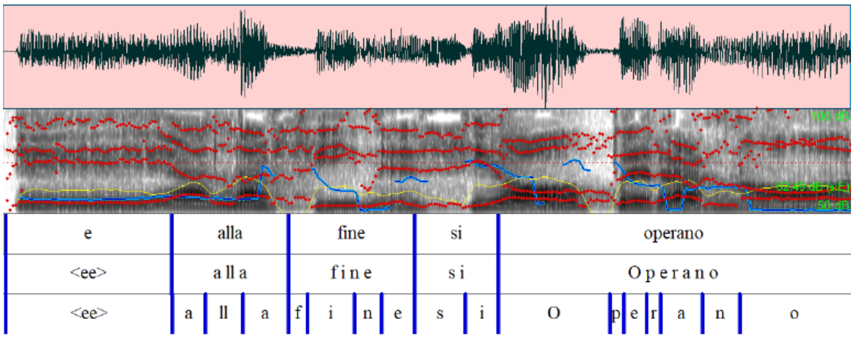


Figure 2: Resulting TextGrid file of a MAUS forced Alignment in Praat

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.

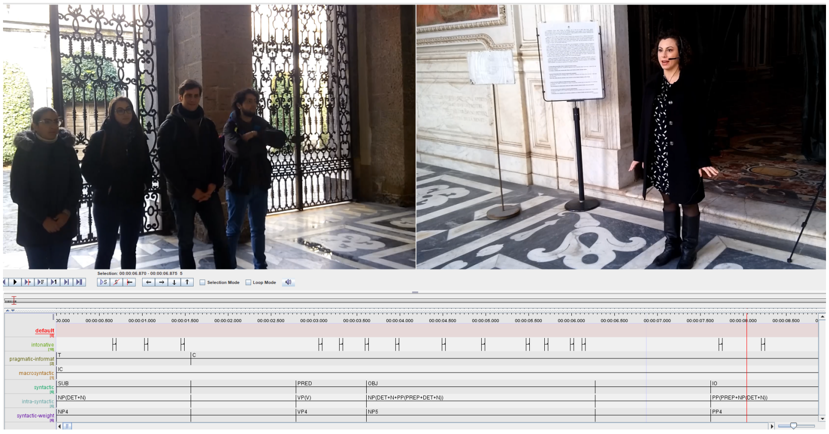
[9] SpIt-MDb Corpus (Spoken Italian - Multilevel Database). Archived at the University of Salerno. Published in 2006. http://www.parlaritaliano.it/index.php/it/corpora-di-parlato/644-spit-mdb-spoken-italian-multilevel-database

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.



Selection: 00:00:06.870 - 00:00:06.875 5

default
intonative
pragmatic-informat: T | C
macrosyntactic: IC
syntactic: SUB | PRED | OBJ | IO
intra-syntactic: NP(DET+N) | VP(V) | NP(DET+N+PP(PREP+DET+N)) | PP(PREP+NP(DET+N))
syntactic-weight: NP4 | VP4 | NP5 | PP4

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 preferences 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.