CLARIN Annual Conference 2018

PROCEEDINGS

Edited by

Inguna Skadiņa, Maria Eskevich

8-10 October 2018
Pisa, Italy

Programme Committee

Chair:

- Inguna Skadiņa, Institute of Mathematics and Computer Science, University of Latvia & Tilde (LV)

Members:

- Lars Borin, Språkbanken, University of Gothenburg (SE)
- António Branco, Universidade de Lisboa (PT)
- Koenraad De Smedt, University of Bergen (NO)
- Griet Depoorter, Institute for the Dutch Language (NL/Vlanders)
- Jens Edlund, KTH Royal Institute of Technology (SE)
- Tomaž Erjavec, Dept. of Knowledge Technologies, Jožef Stefan Institute (SI)
- Francesca Frontini, University of Montpellier (FR)
- Eva Hajičová, Charles University (CZ)
- Erhard Hinrichs, University of Tübingen (DE)
- Nicolas Larrousse, Huma-Num (FR)
- Krister Lindén, University of Helsinki (FI)
- Bente Maegaard, University of Copenhagen (DK)
- Karlheinz Mörth, Institute for Corpus Linguistics and Text Technology, Austrian Academy of Sciences (AT)
- Monica Monachini, Institute of Computational Linguistics “A. Zampolli" (IT)
- Costanza Navarretta, University of Copenhagen (DK)
- Jan Odijk, Utrecht University (NL)
- Maciej Piasecki, Wrocław University of Science and Technology (PL)
- Stelios Piperidis, Institute for Language and Speech Processing (ILSP), Athena Research Center (EL)
- Kiril Simov, IICT, Bulgarian Academy of Sciences(BG)
- Marko Tadič , University of Zagreb (HR)
- Jurgita Vaičenonienė, Vytautas Magnus University (LT)
- Tamás Váradi, Research Institute for Linguistics, Hungarian Academy of Sciences (HU)
- Kadri Vider, University of Tartu (EE)
- Martin Wynne, University of Oxford (UK)

Reviewers:

- Ilze Auziņa, LV
- Bob Boelhouwer, NL
- Daan Broeder, NL
- Silvia Calamai, IT
- Roberts Darģis, LV
- Daniël de Kok, DE
- Riccardo Del Gratta, IT
- Christoph Draxler, DE
- Dimitrios Galanis, GR
- Maria Gavrilidou, GR
- Luís Gomes, PT
- Normunds Grūzītis, LV
- Jan Hajič, CZ
- Marie Hinrichs, DE
- Pavel Ircing, CZ
- Mateja Jemec Tomazin, SI
- Neeme Kahusk, EE
- Fahad Khan, IT
- Alexander Kŏnig, IT
- Jakub Mlynar, CZ
- Jiří Mírovský, CZ
- Marcin Oleksy, PL
- Petya Osenova, BG
- Haris Papageorgiou, GR
- Hannes Pirker, AT
- Marcin Pol, PL
- Valeria Quochi, IT
- João Rodrigues, PT
- Ewa Rudnicka, PL
- Irene Russo, IT
- João Silva, PT
- Egon W. Stemle, IT
- Pavel Stranak, CZ
- Thorsten Trippel, DE
- Vincent Vandeghinste, BE
- Jernej Vičič, SI
- Jan Wieczorek, PL
- Tanja Wissik, AT
- Daniel Zeman, CZ
- Claus Zinn, DE
- Jerneja Žganec Gros, SI

CLARIN 2018 submissions, review process and acceptance

- Call for abstracts: 17 January 2018, 28 February 2018
- Submission deadline: 30 April 2018
- 77 submissions in total were received and reviewed (three reviews per submission)
- Face-to-face PC meeting in Wroclaw: 21-22 June 2018
- Notifications to authors: 2 July 2018
- 44 accepted submissions: 21 oral presentations, 23 posters/demos

More details can be found at https://www.clarin.eu/event/2018/clarin-annual-conference-2018-pisa-italy.

Table of Contents

Parallel Session 3: Use of the CLARIN infrastructure

Parallel Session 4: Design and construction of the CLARIN infrastructure

Poster session

EXMARaLDA meets WebAnno

Steffen Remus* **Hanna Hedeland†** **Anne Ferger†** **Kristin Bührig†** **Chris Biemann***

***Language Technology, MIN**
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 an extension of the popular web-based annotation tool WebAnno, allowing for linguistic annotation of transcribed spoken data with time-aligned media files. Several new features have been implemented for our concomitant 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 annotation of spoken language data, apart from technical and data model related issues, the extension of WebAnno also offers a partitur view for the inspection of parallel utterances in order to analyze various aspects related to methodological questions in the analysis of spoken interaction.

1 Introduction

We present an extension of the popular web-based annotation tool WebAnno[1] (Yimam et al., 2013; Eckart de Castilho et al., 2014) which allows linguistic annotation of transcribed spoken data with time aligned media files.[2] 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 of teaching students to analyze and reflect on authentic classroom communication, and also on linguistic transcription as a part of that analysis. For this 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), which provides XML transcription and metadata formats. 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 (HTML) visualization of transcription data for qualitative analysis. However, within the relevant context of university teaching, apart from such requirements addressing the peculiarities of spoken data, several further requirements regarding collaborative annotation and management of users and data became an increasingly important part of the list of desired features: *a*) proper handling of spoken data (e.g. speaker and time information) *b*) playback and display of aligned audio and video files *c*) visualization of the transcript in the required layout *d*) complex manual annotation of linguistic data *e*) support for collaborative (i.e. pair-wise) annotation *f*) support for annotator agreement assessment *g*) reliable user management (for student grading). Furthermore, a web-based environment was preferred to avoid any issues with installation or differing versions of the software or the problems that come with distribution of transcription

[1] https://webanno.github.io

[2] https://github.com/webanno/webanno-mm

and video data. Another important feature was to use a freely available tool to allow others to use the teaching method developed within the project using the same technical set-up.

While WebAnno fulfills the requirements not met by the EXMARaLDA system or similar desktop applications, it was designed for the annotation of written data only and thus required various extensions to interpret and display transcription and video data. Since there are several widely used tools for the creation of spoken language corpora, we preferred to rely on an existing interoperable standardized format, the ISO/TEI Standard Transcription of spoken language[3], to enable interoperability between various existing tools with advanced complementary features and WebAnno.

In Section 2, we will further describe the involved components, in Section 3 we will outline the steps undertaken for the extension of WebAnno, and in Section 4, we will describe the novel teaching method and the use of the tool within the university teaching context. In Section 5, we present some ideas on how to develop this work further and make various additional usage scenarios related to annotation of spoken and multimodal data possible.

2 Related work

The EXMARaLDA system: The EXMARaLDA[4] transcription and annotation tool (Schmidt and Wörner, 2014) was originally developed to support researchers in the field of discourse analysis and research into multingualism, but has since then been used in various other contexts, e.g. for dialectology, language documentation and even with historical written data. The tool provides support for common transcription conventions (e.g. GAT, HIAT, CHAT) and can visualize transcription data in various formats and layouts for qualitative analysis. The score layout of the interface displays a stretch of speech corresponding to a couple of utterances or intonational phrases, which is well suited for transcription or annotations spanning at the most an entire utterance, but an overview of larger spans of discourse is only available in the visualizations generated from the transcription data. The underlying EXMARaLDA data model only allows simple span annotations of the transcribed text; more complex tier dependencies or structured annotations are not possible. 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 this task cumbersome.

WebAnno – a flexible, web-based annotation platform for CLARIN: WebAnno offers standard means for linguistic analysis, such as span annotations, which are configurable to be either locked to (or be independent of) token or sentence annotations, relational annotations between two spans, and chained relation annotations. Figure 1 (left) shows a screenshot of the annotation view in WebAnno. Various formats have been defined which can be used to feed data into WebAnno.

For analysis and management, WebAnno is also equipped with a set of assistive utensils such as *a*) web-based project management; *b*) curation of annotations made by multiple users; *c*) in-built inter-annotator agreement measures such as Krippendorff's α, Cohen's κ and Fleiss' κ; and *d*) flexible and configurable annotations, including extensible tagsets. All this is available without a complex installation process for users, which makes it particularly suitable for research organizations and a perfect fit for the targeted use case in this work.

The ISO/TEI Standard for Transcription of Spoken Language The ISO standard ISO 24624:2016 is based on Chapter 8, Transcriptions of Speech, of the highly flexible TEI Guidelines[5] as an effort to create a standardized solution for transcription data. As outlined in Schmidt et al. (2017), most common transcription tool formats, including ELAN (Sloetjes, 2014) and Transcriber (Barras et al., 2000), can be modeled and converted to ISO/TEI. The standard also allows for transcription convention specific units (e.g. utterances vs. phrases) and labels in addition to shared concepts such as speakers or time information, which are modeled in a uniform way.

[3] http://www.iso.org/iso/catalogue_detail.htm?csnumber=37338
[4] http://exmaralda.org
[5] http://www.tei-c.org/Guidelines/P5/

3 Adapting WebAnno to spoken data

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).

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 reserach 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. In addition to the written representation of spoken language, playback of the aligned original media file is another crucial requirement.

From EXMARaLDA to ISO/TEI The existing conversion from the EXMARaLDA format to the tool-independent ISO/TEI standard is specific to the conventions used for transcription, in this case, the HIAT transcription system as defined for EXMARaLDA in Rehbein et al. (2004). Though some common features can be represented in a generic way by the ISO/TEI standard, for reasons described above, several aspects of the representation must remain transcription convention specific, e.g. the kind of linguistic units defined below the level of speaker contributions.

Furthermore, metadata is handled in different ways for various transcription formats, e.g. the EXMARaLDA system stores metadata on sessions and speakers separated from the transcriptions 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.

Parsing ISO/TEI to UIMA CAS The UIMA[6] (Ferrucci and Lally, 2004) framework is the foundation of WebAnno’s backend. UIMA stores text information, i.e. the text itself and the annotations, in so-called CASs (Common Analysis Systems). 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 this, we parse the ISO/TEI[7] XML content and store utterances of individual speakers in different views (different CAS of the same document) and keep time alignments as metadata within a CAS.

We use the `annotationBlock` XML element as a non-disruptive unit since we can safely assume that ISO/TEI span annotations are within the time limits of the utterance. Note that annotations, such as incidents, which occur across utterances, are not converted into the WebAnno annotation view, but are present in the partitur view. Other elements, such as utterances, segments, incidents, and existing span annotations are converted to the main WebAnno annotation view.

New GUI features In order to show utterances and annotations in a well known and established parallel environment similar to EXMARaLDA’s score layout of the partitur editor, we adapt the existing online show case demos[8] and call this view the partitur view henceforth. Figure 1 (right) shows a screenshot of the adjustable partitur view. Both views, the annotation view and the partitur view are synchronized, i.e. by clicking on the correct marker in the particular window, the focus changes on the other.

Also, the partitur view offers multiple media formats for selection, viewing speaker or recording related details and a selectable width of the partitur rows. In the annotation view, we use zero width span annotations for adding time markers. Each segment starts with a marker showing the respective speaker. All markers are clickable and trigger the focus change in the partitur view and start or pause the media.

[6] Unstructured Information Management Architecture: `https://uima.apache.org/`

[7] Since ISO/TEI is too powerful in its general form, we restrict ourselves to the HIAT conventions.

[8] available at `http://hdl.handle.net/11022/0000-0000-4F70-A`

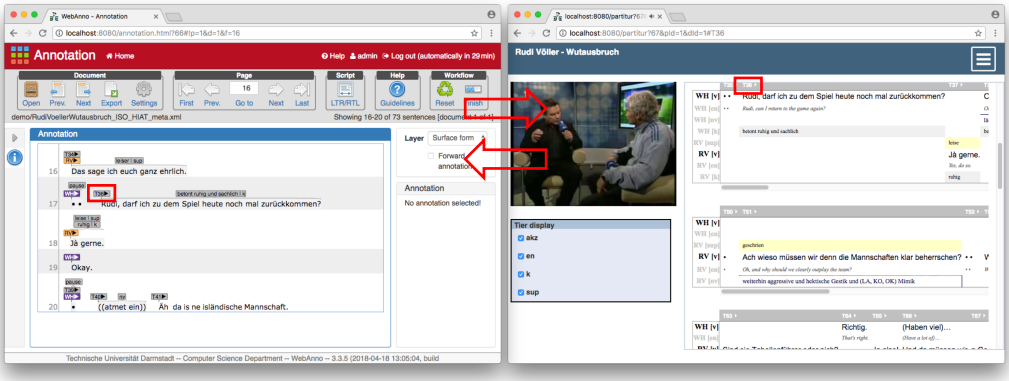


Figure 1: Screenshot of the WebAnno-EXMARaLDA plugin. Left: WebAnno's annotation view; Right: approximate EXMARaLDA partitur view. Both sides are synchronized by clicking the correct markers.

For media management, we added a media pane to the project settings, where we included support for uploading media files, which implies hosting them within the WebAnno environment, 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. Furthermore, multiple media files can be mapped to multiple documents, which allows proper reuse of different media formats for multiple document recordings.

4 WebAnno goes innovative teaching

As part of a so-called "teaching lab" the extended version of the WebAnno tool 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. Benefitting 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, in this case transcribed according to the HIAT conventions using the EXMARaLDA transcription and annotation tool. The ISO/TEI standard, which can model transcription data produced by various tools according to different transcription conventions, was used as an exchange format. Obvious next steps would therefore be to extend the interoperability to include full support and transcript visualization for further transcription systems, as well as a generic fallback option. 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.

References

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).

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.

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.

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. DE.

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 2016, Aix-en-Provence, 26–28 October 2016*, number 136, pages 113–130. Linköping University Electronic Press, Linköpings universitet.

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.

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.

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

L. Hunyadi
University of Debrecen, Hungary
hunyadi@undieb.hu

T. Váradi
Hungarian Academy of Science, Budapest, Hungary
varadi.tamas@nytud.mta.hu

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

Gy. Kovács
University of Szeged, Hungary
gykovacs@inf.u-szeged.hu

H. Kiss
University of Debrecen, Hungary
kiss.hermina@arts.unideb.hu

K. Takács
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.

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 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.

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 both for online browsing and download at the following addresses: https://clarin.nytud.hu/ds/imdi_browser/ under External Resources.

2.2 The annotation scheme

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) or 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 the separation of modalities in the annotation offers an interesting opportunity to study the interdependence of the two modalities in actual communicative events.

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 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.

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.

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 shape (descending, falling, rising etc.), the absolute and relative vertical position of every single prosodic event through their starting and ending points. 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.

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.

4.1 Modelling turn management: automatic detection of turn taking

The HuComTech corpus provides detailed data on turn management. For each discourse unit it contains annotation to indicate topic initiation, topic elaboration and topic change. 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.

The task is certainly challenging and the experiments so far represent tentative initial steps. Earlier studies on topic structure discovery relied mostly on text and/or prosody, the HuComTech corpus, on the other hand, allows a much wider sources of information to be used as cues, such as gaze, facial expression, hang gestures and head movement etc. Kovacs et al. (2016) built a topic unit classifier with the use of Deep Rectifier Neural Nets (Glorot et al, 2011) and the Unweighted Average Recall metric, applying the technique of probabilistic sampling. We demonstrate in several experiments that this method attains a convincingly better performance than a support vector machine or a deep neural net by itself. For further information see (Kovács et al. 2016)

4.2 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 present a new approach based 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.

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.

Hunyadi et al. 2016 contains a tentative first analysis and in the full paper we will update it with more recent analyses

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)

[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)

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

[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.

[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)

[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.

[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.

[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.

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

[Hunyadi et al 2016] Hunyadi, L., Kiss, H., and Szekrenyes, I. 2016. Incompleteness and Fragmentation: Possible Formal Cues to Cognitive Processes Behind Spoken Utterances. In: Tweedale J. W., Neves-Silva R., Jain L. C., Phillips-Wren G., Watada J., Howlett R. J. (eds.) Intelligent Decision Technology Support in Practice. Cham: Springer International Publishing (pp. 231–257)

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

Florentina Armaselu
Luxembourg Centre for Contemporary and Digital History
University of Luxembourg
florentina.armaselu@uni.lu

Elena Danescu
Luxembourg Centre for Contemporary and Digital History
University of Luxembourg
elena.danescu@uni.lu

François Klein
Luxembourg Centre for Contemporary and Digital History
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 European contemporary history research and teaching. Two experiments are 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 an "inventory" of strengths and weaknesses of language technology applied to the study of history.

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 an added-value of 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 methods? The paper will address these questions, within the CLARIN 2018 *Multimodal data (Oral History)* topic, starting from two experiments based on an oral history collection, XML-TEI[1] annotation and textometric analysis.

It is from 1910 that language scientists began to be interested in oral history (Deschamp, 2013: 109-110). Bridging oral history and linguistics in a digital context has made the object of event-oriented initiatives and research, inside and outside CLARIN's framework (CLARIN-PLUS OH, 2016; Oral History meets Linguistics, 2015; Georgetown University Round Table on Languages and Linguistics, 2001). Different tools and perspectives have been approached, 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 less studied so far. This proposal aims to contribute to this topic (in our opinion of potential interest for the CLARIN community, as related to building and evaluating interdisciplinary connections between history and linguistics) and consists in a workflow for: (1) transforming and processing historical spoken data intended to linguistic analysis; (2) evaluating the impact of the use of language technologies in historical research and teaching.

2 Methodology

The study is based on a selection from the oral history collection on European integration published on the CVCE by UniLu Website[2]. The whole collection comprises more than 160 hours of interviews, in French, English, German, Spanish and Portuguese, with some of the actors and observers of the

[1] http://www.tei-c.org/index.xml.
[2] https://www.cvce.eu/histoire-orale. CVCE is now part of the Luxembourg Centre for Contemporary and Digital History (C²DH) of the University of Luxembourg, https://www.c2dh.uni.lu/.

European integration process. The selection included 5-10 hours of audio-video recordings and transcriptions, in French. The selected transcriptions were converted to a structured format, XML-TEI, then imported into the TXM[3] textometry software (Heiden et al., 2010), for linguistic analysis. Two experiments were devised. The first (EUREKA_2017), functioned as a pilot using a shorter corpus and involved a small group of C²DH researchers. The second (MAHEC_2018) was part of a course in *Political and Institutional History* for the Master 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 enquire on the role of the language technology in answering the proposed questions (or in discovering and formulating other related questions).

2.1 Corpus selection and research questions

The number of interviewees varied from six (EUREKA) to eight (MAHEC), including personalities such as Jean-Claude Juncker, Viviane Reding, Jacques Delors and Étienne Davignon. The selection criterion focused on important milestones in the construction 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 course, e.g. identify the European institutions mentioned in the interviews, their role and interconnections, reconstruct the process of the Economic and Monetary Union (EMU) or determine which of the interviewees is speaking more of the role of Luxembourg in the European integration, which less, and why (MAHEC).

2.2 Corpus preprocessing

The transcriptions were available in Microsoft Word format and contained markers for identifying the interviewer/respondent and, occasionally, timecodes. The transcriptions were first converted from Microsoft Word to XML-TEI[4]. Then, a set of XSLT[5] stylesheets, created for this purpose, were applied to the converted output[6], in order to transform it into specific TEI encoding for the transcription of speech. The extract below shows how the identity and type of speaker were encoded using the <u> tag (*utterance*) and the @who and @corresp attributes. The time points (when present) were encoded using <timeline> and <anchor/> elements, in order to mark the text with respect to time.

```
<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
```

2.3 TXM analysis

The corpus in XML-TEI format was imported into TXM, a textometry software, allowing part of speech tagging and lemmatisation[7], frequency of occurrence counts and statistical analysis of textual corpora. The analysed samples contained a total of 38687 (EUREKA) and respectively 110563 (MAHEC) occurrences. Given the encoding, it was possible to build sub-corpora and partitions corresponding to the type of speaker (respondent/interviewer) and the name of the speakers.

The following TXM features were used by the participants to find answers to the proposed questions: specificities (Lafon, 1980), index, concordances and co-occurrences (TXM manual). Figure 1 illustrates the specificities, i.e. a comparative view on the vocabularies of the speakers (e.g. over-use

[3] http://textometrie.ens-lyon.fr/?lang=en.
[4] Via the OxGarage online service, http://www.tei-c.org/oxgarage/.
[5] https://www.w3.org/TR/xslt/.
[6] Using oXygen XML Editor, https://www.oxygenxml.com/.
[7] Via TreeTagger.

for *banque centrale*[8] in the discourse of Yves Mersch and Jean-Claude Juncker, and respectively deficit in the speech of Étienne Davignon), for the top five European institutions most frequently mentioned in the text. Other features allowed particular queries (index), by a single property or in combination (e.g. *noun + adjective*), 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 ...*[9]) or document visualisation. Our hypothesis was that this type of linguistic analysis may help the participants in their quest for answers to the proposed questions.

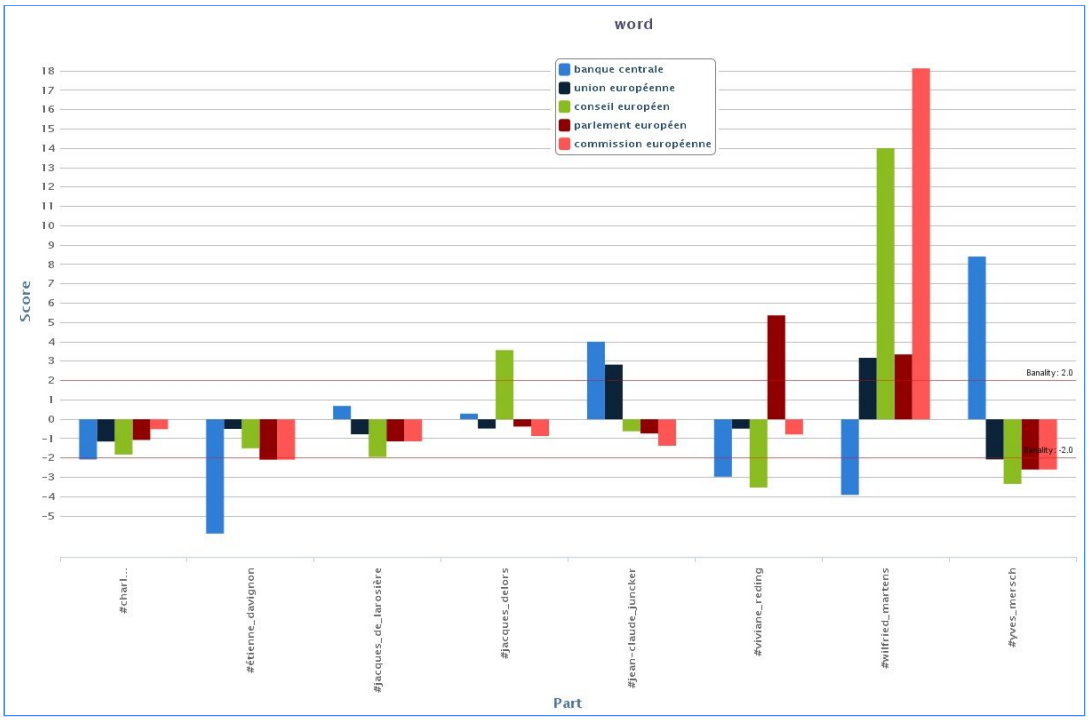


Figure 1. Specificities for European institutions within the respondents' partition (MAHEC_2018)

2.4 Evaluation

The evaluation was intended to confirm/disconfirm this hypothesis and to "measure" the impact of the linguistic technology, its innovative aspects and limitations, when applied to the study of history. The online (anonymised) questionnaires have included: *Yes/No* questions (e.g. *Have you found answers to the research questions?*), Likert-scale queries (e.g. *How do you appreciate the role played by the textometric analysis in the discovery of the answers?* with five possible answers from *Very weak* to *Essential*), open questions (e.g. *Can you shortly describe the added value of this approach, if any?*).

3 The experiments

The pilot experiment EUREKA_2017[10], took place from 11 to 15 and 18 to 22 September 2017 and implied the study of: (1) online audio-video interview sequences and transcriptions; (2) transcriptions using TXM analysis. Evaluation questionnaires were filled-in at the end of each phase. The participants were four C²DH researchers specialised in *European integration*, *Contemporary history*, *Historical and political studies*. Their knowledge varied on a five values scale from *Not at all* to

[8] Eng. *Central Bank*.
[9] Eng. *the European Central Bank is in charge of the monetary policy … .*
[10] Enquiring on the "Eureka effect" of the use of linguistic technology in historical research. The experiment was presented at *Les rendez-vous de l'histoire. Eurêka-inventer, découvrir, innover*, Blois, France, 4-8 October, 2017.

Expert in the fields of: *European integration history*, *Multimedia and oral history*, and *Textometric analysis*. While the data showed specialisation in European integration history with medium knowledge in multimedia and oral history, the self-evaluation of the textometry skills was placed at the lower end of the scale. The second experiment, MAHEC_2018, involved five Master students in *Contemporary European History*, and took place from 16 April to 14 May 2018. The assignment consisted of seven research questions and the evaluation of the added-value/limitations of the language technology in completing the task. The students' background varied from *History* and *Contemporary European history* to *Medieval 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 from *Not at all* to *Good*.

The results of the first experiment (Figure 2) indicate moderate valuation by the participants concerning the role of textometric analysis in finding the answers (left) and the response to the question whether there is a discovery, "Eureka" effect determined by the use of this technology (right), on a scale from -2 to +2, *Very weak* to *Essential* and *Not at all agree* to *Fully agree,* respectively.

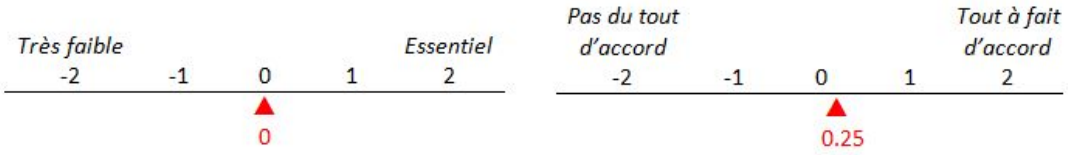


Figure 2. Average scores for textometric analysis (EUREKA_2017): role; "Eureka" effect

As an added value of the method, the participants mentioned: usefulness for analysing large corpora, allowing both local and global observation, rapid identification of the main themes, graphical representation of results. It was also observed that the textometric analysis alone is not sufficient in research. Less positive points were: the interface could have been more intuitive[11], the graphics more attractive and the selected sample larger, in order to fully exploit the potential of the method.

For the second experiment, the average value regarding the role played by the textometric analysis in finding the answers was a bit higher than above (0.4 instead of 0 on the -2 to +2 scale). The aspects evoked as added value were similar to those mentioned in the first case, e.g. allowing the analysis of a large corpus of documents instead of reading them one by one, "fast reading", speed and rigour. As strong points, it was noted the use of part of speech based queries and the suitability of textometric analysis for assisting interpretation. As weak points were mentioned the results window that should have been larger and the heterogeneity of the questions proposed to the interviewees instead of a common set that would have allowed a better basis for comparing their responses. Concerning the innovative side of the studied technology, it was pointed out that it often served just to prove the position or the role of a given personality within the European integration process, rather than providing new information. An aspect that ought to be further examined in future experiments.

4 Conclusion and future work

The project combined oral history and digital linguistic analysis, and evaluated the use of language technology in history research and teaching. Two experiments have been devised. Although rapidity in processing and visualising linguistic features in large amounts of texts were mainly valued, the results showed a certain reserve concerning the innovative added value of the analysis tool. Perhaps, since, as specialists or students in the field, the topic of European integration was, to a certain extent, already known to the participants. For comparison purposes, more evaluation results, from different groups of participants with different degrees of knowledge about the proposed topic, are needed. The longer term objective of this type of evaluations would be to draw an "inventory" of strengths and weaknesses of language technology applied to the study of history.

[11] For EUREKA, no initial TXM training was provided, just a tutorial and assistance with the tool during the experiment. For MAHEC, a short TXM training was provided, as well as a tutorial and assistance.

References

[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

[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.

[Descamps 2013] Florence Descamps. 2013. "Histoire orale et perspectives. Les évolutions de la pratique de l'histoire orale en France". In F. d'Almeida et D. Maréchal (dir.), *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.

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

[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/.

[Lafon 1980] Pierre Lafon. 1980. "Sur la variabilité de la fréquence des formes dans un corpus". *Mots*, N°1 , p 127-165. http://www.persee.fr/doc/mots_0243-6450_1980_num_1_1_1008.

[ENS de Lyon & Université de Franche-Comté 2017] ENS de Lyon & Université de Franche-Comté. 2017. *Manuel de TXM 0.7.8*, http://txm.sourceforge.net/doc/manual/manual.xhtml.

[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.

[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*, edited by Deborah Tannen and James E., pp. 84–113. Alatis, Georgetown University Press, Washington, D.C. http://faculty.georgetown.edu/schiffrd/index_files/Linguistics_and_oral_history.pdf.

[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.

The Acorformed Corpus: Investigating Multimodality in Human-Human and Human-Virtual Patient Interactions

M. Ochs[1], P. Blache[2], G. Montcheuil[2,3,4], J.M. Pergandi[3], R. Bertrand[2], J. Saubesty[2], D. Francon[5], and D. Mestre[3]

Aix Marseille Université, Université de Toulon, CNRS,
[1]LIS UMR 7020, [2]LPL UMR 7309, [3]ISM UMR 7287 ; [4]Boréal Innovation,
[5]Institut Paoli-Calmettes (IPC), Marseille, France

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 model based on the analysis of 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 corpus 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.

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 experts annotated one third of the corpus each. 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).

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 patient's feedbacks, 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., 2018)). 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).

[2]The corpus is on Ortolang part of the CLARIN infrastructure

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.

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 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.

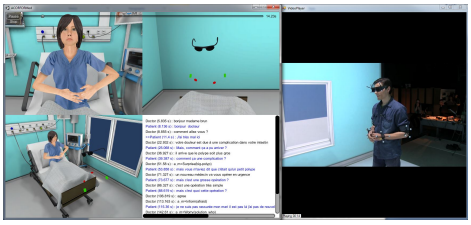


Figure 2: 3D video playback player

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 to the experimentation. Ten of them are real doctors that already have an 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[3].

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 http://crvm.ism.univ-amu.fr/en/acorformed.html). Note that thanks to the tools described in the previous section, some of the non-verbal participant behavior has be automatically annotated.

In order to evaluate the global exerience 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))[4]. These subjective

[3]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).

[4]The analyze of the subjective experience of the participants is out of scope of this paper and is described in other articles

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 behavior of the participants) to subjective measures (e.g. feeling of presence and perception of the virtual patient's believability).

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).

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.

R. Beale and C. Creed. 2009. Affective interaction: How emotional agents affect users. *International journal of human-computer studies*, 67(9):755–776.

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*.

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.

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.

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)*.

currently under review.

M. Ochs, P. Blache, G. Montcheuil, J.-M. Pergandi, J. Saubesty, D. Francon, and D. Mestre. 2018. 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*.

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*.

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.

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.

Media Suite: Unlocking 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

Carlos Martinez-Ortiz
Netherlands eScience Center
Amsterdam
The Netherlands
c.martinez@esciencecenter.nl

Julia Noordegraaf
Department of Media Studies
University of Amsterdam
The Netherlands
J.J.noordegraaf@uva.nl

Abstract

This paper discusses the rationale behind the development of a research environment –the Media Suite– in a sustainable, dynamic, multi-institutional infrastructure that supports mixed media scholarly research with large multimedia data 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 astronomy for instance, new collections of astronomical observations are made publicly available on a regular basis. In other domains such as Media Studies research focuses on data collections maintained at cultural heritage institutions, archives, libraries, and 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 (play-out, viewing). Therefore, scholars either fall back on collections that are openly available or spend considerable amounts of time in *onsite* visits to archives for consulting data collections. 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 *mixed-media* sources, such as radio and television broadcasts, film, oral history interviews but also (news)paper archives, film posters and eyewitness reports. The *Media Suite* serves as the online portal to the infrastructure where first of all, content and metadata can be explored, browsed, compared, and stored in personal collections. In addition, the Media Suite provides a workspace for working with mixed media collections, providing tools for manual and automatic annotation, visualization, analysis, and sharing.

The ultimate goal is to (i) enable distant reading (Schulz, 2011), that is, identifying patterns or new research questions in all aggregated collections, (ii) facilitate close reading: the detailed examination of individual items (e.g., videos) in a collection or parts of these items (e.g., video segments) during search and scholarly interpretation, and (iii) make sure that the "scholarly primitives" (Unsworth, 2000;

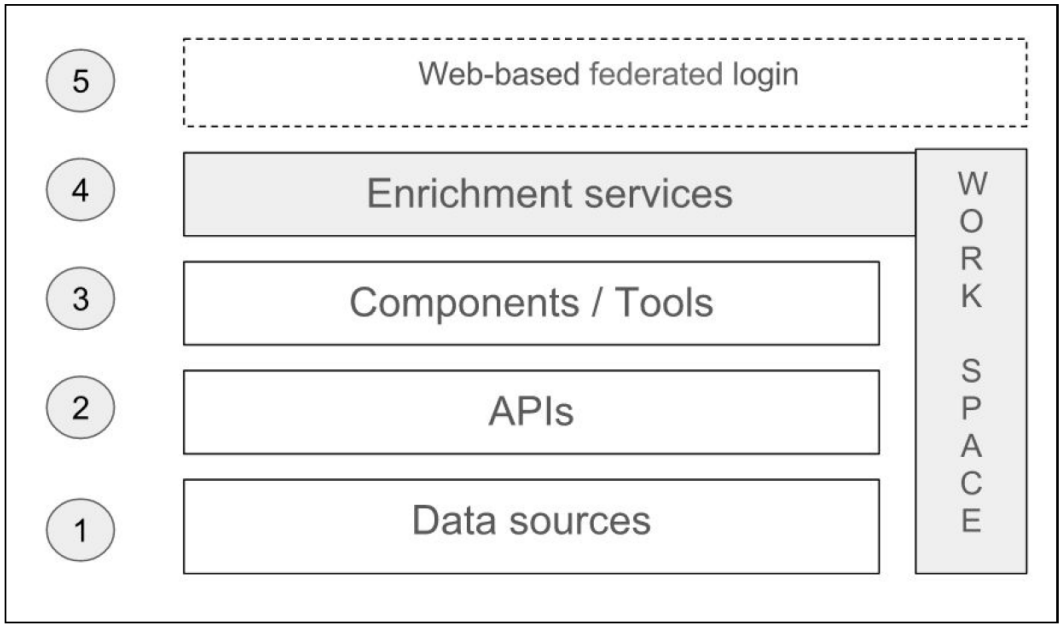


Figure 1: The building blocks of the CLARIAH Media Suite

Blanke and Hedges, 2013), basic activities common to research across humanities disciplines, are well supported.

1.1 Challenges

Questions however are: How to facilitate "close reading" when the media objects cannot be accessed because of copyright issues? How to enable "distant reading" when metadata is sparse, or diverse, and incomplete? How to cater to the needs of scholars with specific research questions and methods in the context of an infrastructure that has to be generic enough to be feasible? How to enable scholars to work with collections from different institutes using the same tools, when these collections are "locked"? How to enable scholars that are computer literate to work directly with the data or to deploy private content analysis tools such as computer vision or sentiment analysis?

The approach of the CLARIAH Media Suite to tackle these challenges is to provide mechanisms that enable researchers to work with tools and aggregated data *within* the closed environment of the infrastructure sealed with a federated authentication mechanism (SURFConext[1]) that currently only serves scholars with a university account in the Netherlands, but that soon will be expanded to the CLARIN federation. Also, the so called 'homeless users' that do not have an account with an academic institution, will eventually have the opportunity to request for a login. We refer to this approach as to "bringing the tools to the data", as opposed to "bringing the data to the tools".

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

Institutional collection maintainers have internal data governance processes to ensure that data assets are formally managed. One important aspect covered by governance processes is licensing: who has

[1] https://www.surf.nl/en/services-and-products/surfconext/index.html

permission to access the data. However, data governance with respect to external processes –loosely defined as being part of an 'infrastructure'– is typically not accounted for. 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)organized or (re)considered, formalized and supported by the Media Suite and the emerging infrastructure in which it is embedded.

3 APIs – Sustainable development

A digital infrastructure should use existing protocols, conventions, and standards. Besides obtaining data by harvesting using the OAI-PMH protocol, or using application programming interfaces (APIs), the functionalities have been organized in a modular approach, which includes (Martinez-Ortiz et al., 2017):

- Components that use API's to perform specific tasks.
- Tools that incorporate a number of components in a tool.

4 Components/Tools – User-friendly interaction design

Developing new tools "from scratch" for every research question would be a very inefficient (and costly!) endeavour. The digital infrastructure should provide tools that are suitable both for common scholarly tasks and for specific tasks required by each discipline. However, 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. We address this challenge by (i) focusing on the similarities in research methods from different disciplines (de Jong et al., 2011; Melgar Estrada and Koolen, 2018), (ii) analyzing tools that support qualitative methods (Melgar et al., 2017), and (ii) working with scholars as co-developers in the process . The resulting functionalities are built in a modular (lego) approach that supports both flexible software development of components and user-friendly interaction with assembled tools.

5 Work Space – Working with audio-visual content and private data

In addition to IPR 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 Eifel 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. We are implementing an automatic speech recognition service that resides within the CLARIAH infrastructure 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 workflows centred around a "Work Space" (see Figure 2) that has the following functionalities:

- Storing individual items from different "institutional" collections resulting in a private, virtual, multimedia, research collection.
- Storing private session data such as queries and filtering options.

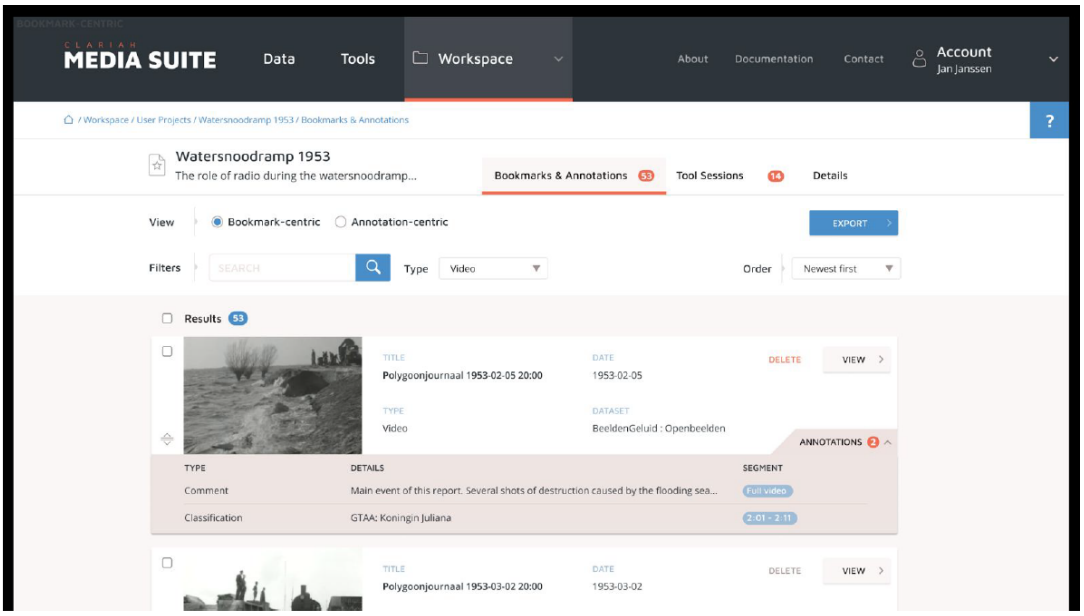


Figure 2: The CLARIAH Media Suite's Workspace

- Uploading private data and perform enrichment services to these data (e.g., speech recognition)
- Running private code on data collections in the infrastructure for creating data visualization (e.g., Jupyter Notebooks).

6 Conclusion and future work

We described the challenges found in building an infrastructure that satisfies the needs of humanities scholars working with audio-visual media and contextual collections. We choose the approach of building a research environment that adheres to infrastructural requirements while at the same time being flexible 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 incorporates flexible functionality capable of addressing very specialistic research questions. The Media Suite is currently functional and used by scholars doing actual research projects and will be developed further, e.g., by incorporating additional data sources (e.g., social media data), increasing metadata granularity (e.g., adding computer vision or emotion recognition), adding advanced annotation tools, and supporting missing data visualization (data critique) for heterogeneous datasets.

References

Tobias Blanke and Mark Hedges. 2013. Scholarly primitives: Building institutional infrastructure for humanities e-science. *Future Generation Computer Systems*, 29(2):654–661.

Franciska M.G. de Jong, Roeland J.F. Ordelman, and Stef Scagliola, 2011. *Audio-visual Collections and the User Needs of Scholars in the Humanities: a Case for Co-Development*, pages –. Centre for Language Technology, Copenhagen, 11. eemcs-eprint-20868.

Carlos Martinez-Ortiz, Roeland Ordelman, Marijn Koolen, Julia Noordegraaf, Liliana Melgar, Lora Aroyo, Jaap Blom, Victor de Boer, Willem Melder, Jasmijn van Gorp, Eva Baaren, Kaspar Beelen, Norah Karrouche, Oana Inel, Rosita Kiewik, Themis Karavellas, and Thomas Poell. 2017. From tools to "recipes": Building a media suite within the dutch digital humanities infrastructure clariah. DHBenelux.

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 Estrada and Marijn Koolen. 2018. Audiovisual media annotation using qualitative data analysis software: A comparative analysis. *The Qualitative Report*, 23(13):40–60.

Kathryn Schulz. 2011. What is distant reading. *The New York Times*, 24.

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.

Using Linked Data Techniques for Creating an IsiXhosa Lexical Resource - a Collaborative Approach

Thomas Eckart, Bettina Klimek, Dirk Goldhahn
Institute of Computer Science
University of Leipzig, Germany
{teckart,klimek,goldhahn}
@informatik.uni-leipzig.de

Sonja Bosch
Department of African Languages
University of South Africa
Pretoria, South Africa
boschse@unisa.ac.za

Abstract

The CLARIN infrastructure already provides a variety of lexical resources for many languages. However, the published inventory is unevenly distributed favouring languages with large groups of native speakers and languages spoken in highly developed countries. Improving the situation for so called "under-resourced languages" is possible by close collaboration with the language-specific communities and expertise that - naturally - reside in the countries where those languages are spoken. This submission presents an example for such a collaboration where a representative sample of an existing lexical resource for the isiXhosa language, which is spoken in South Africa, was processed, enriched, and published. The resource under discussion is intended to be a prototype for more resources to come.

1 Introduction

Many resources that are provided that use standards and interfaces of the CLARIN project, or adjoining initiatives, are of a high quality and form the cornerstone of research applications and endeavours. Unsurprisingly, a short evaluation of provided resources in applications like the VLO shows an emphasis on European languages or languages with large speaker groups. However, this is only a limited view on the world's languages and the availability of required language resources.

There are many projects established in the European research context that work on creating documentation about and data stocks for languages where those are missing. However, in many cases valuable resources of high quality which are unique in their extent and focus do already exist, but are not yet available for a broader audience via standard applications and Web portals.

In cases where data is still preserved on paper, standard digitisation procedures have to be carried out. From the perspective of modern research environments, this is only the beginning of the life-cycle of research data. Questions about data formats used, supported query interfaces and means to announce the existence of the resource to the public have to be considered as well, and are outstanding issues in projects like CLARIN, comparable projects like the South African SADiLaR (Roux, 2016), or the growing Linguistic Linked Open Data (LLOD) Cloud community (Chiarcos et al., 2012). This submission presents the collaborative work of three partners from different scientific domains with the goal of preparing and enhancing a valuable lexical and morphological resource of the isiXhosa language (ISO 639-3: xho) in the described context. The aim of this is to provide a high-quality resource that is publicly available and ready for use in future research endeavours.

2 Language Diversity in South Africa

South Africa is in the unique position of having 11 official languages entrenched in its constitution[1]. Nine of the official languages belong to the Bantu language family, and more specifically to

[1] https://www.gov.za/DOCUMENTS/CONSTITUTION/constitution-republic-south-africa-1996-1

Guthrie's Zone S in which the following language clusters are identified (Nurse and Philippson, 2003, 609-611):

- S20 Venda Group - Tshivenda [ven]
- S30 Sotho-Tswana Group - Setswana [tsn], Sesotho sa Leboa [nso], Sesotho [sot]
- S40 Nguni Group - IsiZulu [zul], isiXhosa [xho], Siswati [ssw], isiNdebele [nbl]
- S50 Tshwa-Ronga Group - Xitsonga [tso]

The typical word structure of the Bantu languages is of an agglutinating nature, where formatives may be affixed to root morphemes, thereby altering the meaning of the word. Each formative morpheme maintains a distinct and fixed meaning; even when used with different root morphemes, it still maintains its original meaning.

IsiXhosa, the language under discussion in this paper, is spoken predominantly in the Eastern Cape and Western Cape regions and has approximately 8.1 million first-language speakers, which is about 16% of the South African population. IsiXhosa shares a so-called conjunctive orthography with the other languages of the Nguni group which means that bound morphemes are attached to the words (unlike other South African Bantu languages) and thus cannot occur independently as separate words.

Compared to European languages, the South African Bantu languages are considered resource scarce in the sense that lexical resources are very limited. Although they all have at least one or two paper dictionaries available, ranging from monolingual to bilingual general purpose or learners' dictionaries, the only online dictionary is isiZulu.net[2], an online isiZulu-English dictionary containing bidirectional lookups as well as basic morphological decomposition. There are no machine-readable lexicons freely available for any of the languages. It should also be noted that the South African Bantu languages are not yet completely standardised with regard to orthography, terminology and spelling rules (Taljard and Bosch, 2006, 429).

3 IsiXhosa Data Set

The described isiXhosa data set is currently a representative sample of raw data for a Xhosa-English dictionary, containing approximately 6,800 lexical entries (Bosch et al., 2018). In its final state, the data set will contain approximately 10,000 lexical entries. The data set is accompanied by English translations and was compiled and made available for purposes of further developing Xhosa language resources. Bilingual (Xhosa-English) word lists were compiled with the intention of documenting Xhosa words and expanding existing bilingual Xhosa dictionaries.

The preparation process of the data involved digitisation into CSV tables and various iterations of quality control in order to make the data reusable and shareable. The resulting records were encoded using a RDF-based ontology supporting characteristics of Bantu languages, which is explained in more detail in the following section. Table 1 gives a short overview of the current inventory.

Dataset Feature	**Value**
Number of noun lexemes	4020
Number of verb lexemes	2763
Number of noun classes	15
Number of English translations	7807

Table 1: Characteristics of the isiXhosa data set.

The data set, CMDI-based metadata, and all required interfaces (including a SPARQL endpoint) are currently hosted at a German CLARIN centre (*hdl:11022/0000-0007-C655-A*) and at

[2] `https://isizulu.net`

a dedicated GitHub project[3]. The final result will be hosted in the context of the South African Centre for Digital Language Resources[4] (SADiLaR) and is to be expected until mid-2019.

4 The Bantu Language Model

For the creation of the isiXhosa data set the Linked Data format RDF (resource description framework) was chosen. This choice is motivated by the aim to convert the source data into a highly interoperable and machine-processable format, especially with regard to future language tools that will be based on this data and applications that concern interconnection and enrichment with other language data. The formal basis of such a Linked Data set builds its underlying ontology, also called vocabulary. For the isiXhosa data set a dedicated ontology, the Bantu Language Model[5] (in short BLM), has been created that is not only suitable for representing isiXhosa lexical and morphological language data but also for other Bantu languages. The structure of the ontology is illustrated in figure 1. As it is established practice in the Semantic Web environment to reuse already existing vocabularies, the BLM is based on the OntoLex-Lemon model[6] for representing lexical entries and translational data and the Multilingual Morpheme Ontology (MMoOn)[7](Klimek et al., 2016) for representing morphological language data. Lexemes, wordforms and morphs are all interconnected via object properties and linguistic categories like nominal classifiers which are already included and specified for Bantu languages. In order to facilitate the reuse of the vocabulary, already existing classes and properties from the two other vocabularies have been essentially duplicated and interrelated with an equivalence relation to its origin. As a result, the same name space can be used for the BLM as a whole, which spares the effort to consult the external vocabularies in order to understand the semantics of the ontology. Overall, the BLM fulfils the following defined goals which are exemplary proven with the isiXhosa data set: 1) create a consistent semantic core model that can serve as a shared basis for a multilingual Bantu language data environment, 2) model lexical and morphological data that takes Bantu language-specific characteristics into account and 3) use a machine-readable and interoperable format that allows future extensions and interconnectivity with other data sets. The last goal has been implemented by adding the isiXhosa data set to the LLOD Cloud[8] and interlinking its English translations with English lexical entries of the RDF WordNet data set.

5 Summary and further work

This contribution described the preparation and publication of a lexical resource for a so-called "less resourced language" while simultaneously reaching out to other linguistic communities using linked open data standards. While the landscape of the Linguistic Linked Open Data Cloud is growing, the amount of the data available via CLARIN compatible interfaces and their support in CLARIN services is still marginal and should be extended in the future. This might especially be a reasonable decision, as many aspects of the LLOD infrastructure are compatible with key ideas promoted by the CLARIN project. This contains decentralised access and query endpoints for locally stored data, describing and accessing resources via open standards, or the re-use of vocabularies and schemata that are also used in "traditional" fields of the linguistic community. Therefore, infrastructure projects like CLARIN and SaDiLaR have the potential to pave the way in this area, and to provide and promote both their long-time experience with the management of digital language resources and user-friendly applications that allow an easy integration of new data stocks.

It should also be seen as an example for the collaboration of CLARIN centres with language experts in non-European countries to promote key ideas of the CLARIN project, to increase the

[3] https://github.com/MMoOn-Project/OpenBantu
[4] http://www.sadilar.org
[5] http://mmoon.org/bnt/schema/bantulm/
[6] https://www.w3.org/community/ontolex/wiki/Final_Model_Specification
[7] http://mmoon.org/
[8] https://datahub.ckan.io/dataset/open-bantu-isixhosa-lexicon

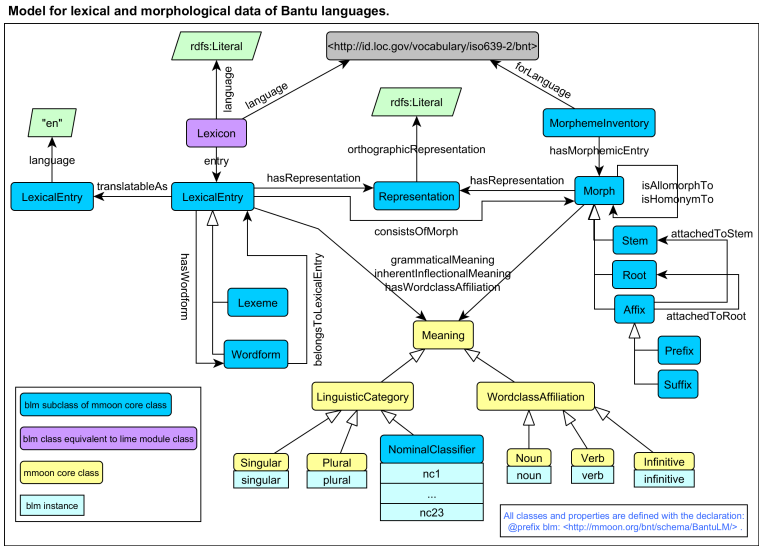


Figure 1: Ontology for the Bantu Language Model.

coverage of language resources for under-resourced languages and to ease the "bias" in favour of European languages in the current global resources landscape.

As the next working step, the described dataset for isiXhosa will be extended and improved, and the final result will be published by mid-2019. The collaboration is intended to be a working prototype for further available resources that wait to be digitised or transformed using modern standards and practices. Among those resources are dictionaries provided by the Comparative Bantu OnLine Dictionary project (CBOLD[9]) which often contain translations into languages like English or French. The explicit encoding of translation information in the Bantu Language Model is seen as a valuable foundation for cross-lingual alignment of those dictionaries in the context of a distributed environment for lexical resources.

References

[Bosch et al.2018] Sonja Bosch, Thomas Eckart, Bettina Klimek, Dirk Goldhahn, and Uwe Quasthoff. 2018. Preparation and usage of xhosa lexicographical data for a multilingual, federated environment. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki (Japan).*

[Chiarcos et al.2012] Christian Chiarcos, Sebastian Hellmann, and Sebastian Nordhoff, 2012. *Linking Linguistic Resources: Examples from the Open Linguistics Working Group*, pages 201–216. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Klimek et al.2016] B. Klimek, N. Arndt, S. Krause, and T. Arndt. 2016. Creating Linked Data Morphological Language Resources with MMoOn – The Hebrew Morpheme Inventory. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC-2016), 23-28 May 2016, Slovenia, Portoroz.*

[Nurse and Philippson2003] D. Nurse and G. Philippson. 2003. *The Bantu languages.* London: Routledge.

[Roux2016] Justus C. Roux. 2016. South African Centre for Digital Language Resources. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC-2016), 23-28 May 2016, Slovenia, Portoroz.*

[Taljard and Bosch2006] Elsabé Taljard and Sonja E. Bosch. 2006. A Comparison of Approaches to Word Class Tagging: Disjunctively vs. Conjunctively Written Bantu Languages. In *Nordic Journal of African Studies 15(4)*, pages 428–442. ISSN 1459-9465.

[9] `http://www.cbold.ish-lyon.cnrs.fr/`

A Platform for Language Teaching and Research

Maria Stambolieva
Laboratory for Language Technologies
New Bulgarian University, Sofia, Bulgaria
mstambolieva@nbu.bg

Valentina Ivanova
Digital Innovation Laboratory
New Bulgarian University, Sofia, Bulgaria
v.ivanova@nbu.bg

Mariyana Raykova
Digital Innovation Laboratory
New Bulgarian University, Sofia, Bulgaria
mraykova@nbu.bg

Abstract

The Platform for Language Teaching and Research was designed and developed at New Bulgarian University in answer to important educational needs, some of them specific to Bulgaria. The aim of the paper is to present the tool developed to match those needs, its functionalities, architecture and applications – actual and envisaged. The Platform can provide 1/ course development support for native and foreign language (and literature) teachers and lecturers, 2/ data and tools for corpus-driven and corpus-based lexicography, corpus and contrastive linguistics, 3/ an environment for research, experimentation and comparison of new methods of language data preprocessing. The educational content organised and generated by the Platform is to be integrated in the CLARIN part of the CLaDA-BG infrastructure, of which New Bulgarian University is a partner.

1 Introduction

Language teaching in Bulgaria presents a number of challenges at all educational levels. On the one hand, Bulgarian is not an international language, which imposes the need for foreign language tuition from a very early age, for different age groups and at all levels; on the other hand, Bulgarian students increasingly come from different ethnic backgrounds, with ensuing mixed groups of learners having a different level of proficiency in the Bulgarian language (Stambolieva et al. 2017b). At secondary school and university level, there is also a growing need for a specialisation of both native and foreign language learning (Cf. Stambolieva et al. 2017a). For all educational age levels and groups, the various types of non-homogeneity demand new levels of individualisation of the learning process. These factors combine to define the increasing demand for a wide range of e-tuition tools supporting native and foreign language teaching. It is initially with the aim of answering this demand only that the Platform was developed.

2 Aims of the Platform

The Platform for is a modular, versatile tool aiming to provide 1/ course development support for native and foreign language (and literature) teachers and lecturers, 2/ data and tools for

corpus-driven and corpus-based lexicography, corpus and contrastive linguistics, 3/ an environment for research, experimentation and comparison of new methods of language data preprocessing.

3 Modules and Functions

The Platform integrates: 1/ an environment for creating, organising and maintaining electronic text archives and extracting text corpora: a repository of general or domain-specific texts, further classified in accordance with the Common European Framework of Reference for Languages (CEFRL); 2/ modules for linguistic analysis, including a lemmatiser, an in-depth POS analyser; a term analyser; a syntactic analyser; an analyser of multiple word units (MWU – including complex terms, analytical forms, phraseological units); a parallel text aligner; a concordancer (Cf. McEnery and Hardie 2012, 37-48, Anthony 2013); 3/ a linguistic database allowing linguistic analysis to be performed on either a single text or a corpus of texts: corpora can be modified (reduced in size or expanded with additional texts from the archive) without loss of information; 4/ modules for the generation and editing of vocabulary or grammar drills (Cf. Sinclair 2004); 5/ modules for the extraction of linguistic information directly from texts/corpora or from the data base (Cf. on corpus-based and corpus-driven approaches McEnery and Hardie 2012, 5-14, 147-64). The implemented class hierarchy and the class specialisation are designed to maximise code reuse and to provide encapsulation of the data. New modules for linguistic analysis (as e.g. semantic analysers) can easily be integrated into the existing Platform (Cf. Section 4 below).

The environment for the maintenance of the electronic text archive organises a variety of metadata which can, individually or in combinations, form the basis for the extraction of text corpora. Following linguistic analysis, secondary ("virtual") corpora can be extracted, formed by sentences containing a particular unit – a lemma (e.g. *it*, *dislike*), a word form (e.g. *begins*), a MWU (e.g. *has been writing*, *put off*), an attribute (e.g. <intransitive verb>, <comparative degree>, <present perfect progressive tense>) or a combination of attributes. The sets of categories and attributes generated by the separate modules can be converted into tagsets compatible with other annotation formats.

The platform can thus be used to support a variety of linguistic activities – from the generation of drills to the compilation of corpus-driven and corpus-based reference materials – glossaries, thesauri, dictionaries or grammars.

4 Architecture

The architecture of the system is modular. It consists of input modules, modules for preprocessing, processing, analysis and data storage, and output modules. The input modules provide user interface for different linguistic tasks. They work independently. Further implementation of new input modules will extend the features of the system. This architecture allows for the addition of various modules for automatic (pre)processing. Each of the preprocessing modules can be implemented independently and added to the system at an appropriate stage. The architecture allows the implementation of different options – such as the use of statistical methods, of systems of rules or neural networks. It also allows the parallel use of several systems of preprocessing and the comparison of their results for the purpose of making an intelligent choice. In all these cases, the results are confirmed by a team of expert linguists – which, on the one hand, supports the self-learning aspect of preprocessing and, on the other, helps assess the relative reliability of each type of system. With these features, the Platform adds to its more standard functions those turning into an environment for experimentation and research, where new methods of preprocessing will be tested and compared.

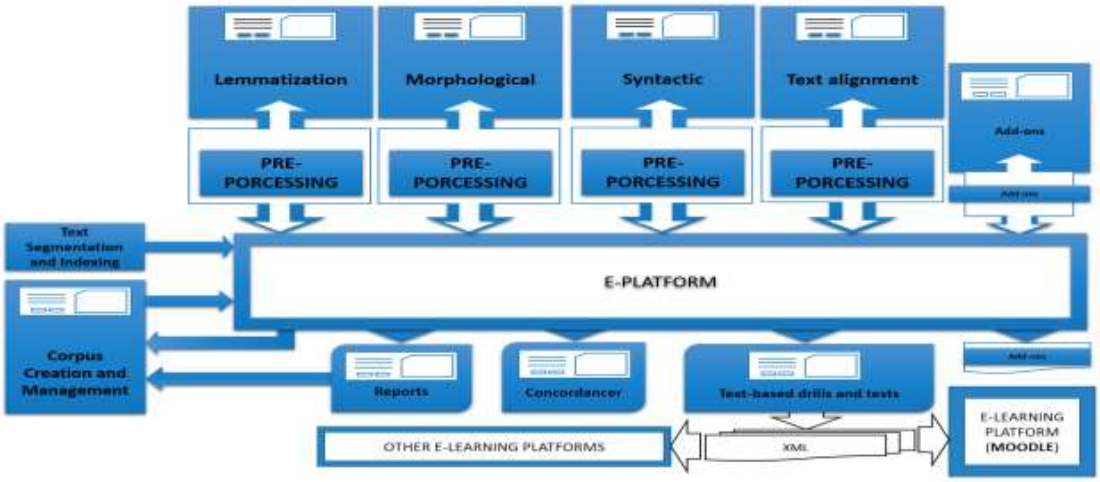


Figure 1. Architecture of the Platform

The interfaces between the input modules, the preprocessing modules and the Platform are custom, but the decoupling between the input modules and the storage allows for cascading preprocessing that may include in-time conversion to/from a standardised query interface.

The text-based drills generator can export the teaching content to the educational platform of the user. In the case of New Bulgarian University, the export format is Moodle XML format for Moodle Quiz Module. The Platform feeds the question banks of Moodle with three types of question items: Fill in the Blanks, Matching, and Reordering, each with a number of subtypes.

The Platform is web based and implements the MVC (Model-View-Controller) architectural pattern. The internal data representation follows the object-oriented paradigm. The fundaments of the system lie on the open source framework CodeIgniter. On one hand, it provides a skeleton for MVC functionality for user defined solutions. On the other, it is implemented on the object-oriented programming language PHP. The framework is light-weighted and fast; it offers ready to use plugins and a full-listed documentation.

The database of the Platform consists of 46 tables in third normal form (3NF), created using the MySQL database. To provide consistent user experience, the user interface design reuses a predefined HTML structure and utilizes a unified template of CSS-coded layout. The architecture of the system is maintainable and extendable.

5 Functions and spheres of application.

Unlike a large number of existing foreign language learning networks and platforms (some of which excellent – as e.g. the social network Busuu, the Duolingo paltform, the Rosetta Stone Dynamic Immersion system or the Babbel platform, and many others), the Platform for Language Teaching and Research developed at New Bulgarian University, Sofia, is not directly oriented towards the student as a final user and does not offer lessons; it offers support to the course developer/applied linguist (lexicographer, grammar book or course book author), corpus or computational linguist. Although it is language independent, it has so far only been developed for Bulgarian and English (with French in development). For these, it offers raw or analysed data, educational content, resources for the applied linguist. The Platform is thus neither simply an organiser / manager of teaching aids and tools (as e.g. the Platform for Technical French and Language Technologies[1]), nor a workbench integrating corpus analysis tools (Cf. Stambolieva and Dragostinov 2014). It is a system combining these functions – a powerful tool for corpus creation, manipulation and analysis and for the generation of varied corpus-based and/or corpus-driven educational and other content. A prominent feature of the system is its modularity and versatility, allowing the user to carry out a large number of linguistic tasks within the same platform, to combine

[1] http://www.beco.pub.ro

results in many different ways and for a variety of tasks, and to reuse all existing resources. The Platform can thus be defined, following L. Anthony's classification, as a fifth generation corpus tool (Anthony op. cit., 155-8).

The NBU Platform has been in use for the last two years as a support for test generation and online tuition and testing of general and specialised English. At present, it supports the development of supplementary teaching material in the Bulgarian language and literature. The modules of the platform can already offer support for the extraction of corpus-based grammars; the modules for lexicographic work are in development. We trust that the platform and the educational and linguistic content that it can generate will be among the important assets of the Bulgarian CLARIN (CLaDA-BG) infrastructure and will attract to it a considerable number of users.

References

[Anthony 2013] Laurence Anthony. 2013. A Critical Look at Software Tools in Corpus Linguistics. Linguistic Research 30 (2) 141-61.

[McEnery and Hardie 2012] Tony McEnery, Andrew Hardie. 2012. Corpus Linguistics. Cambridge University Press.

[Sinclair 2004] John McH Sinclair (ed.). 2004. How to Use Corpora in Language Teaching. Studies in Corpus Linguistics vol. 12. John Benjamins: Amsterdam/Philadelphia.

[Stambolieva et al. 2017a] M. Stambolieva, M. Hadjikoteva, M. Neykova, V. Ivanova, M. Raykova. 2017. The NBU Platform in Teaching Foreign Languages for Specific Purposes. Proceedings of the 13th Annual International Conference on Computer Science and Education in Computer Science, Albena.

[Stambolieva 2017b] M. Stambolieva, M. Hadjikoteva, M. Neikova, V. Ivanova, M. Raikova. 2017. Language Technologies in Teaching Bulgarian at Primary and Secondary School Level: the NBU Platform for Language teaching. Proceedings of RANLP 2017 Workshop" LTDHCSEE: Language Technology for Digital Humanities in Central and (South-) Eastern Europe, 32-8.ISBN 978-954-452-046-5

[Stambolieva and Dragostinov 2014] M. Stambolieva, D. Dragostinov.2014. The NBU Linguist's Workbench. Proceedings of the 10th Annual International Conference on Computer Science and Education in Computer Science. Petya Assenova, Vijaj Kanabar (eds.) New Bulgarian University, Hochschule Fulda University of Applied Science, Boston University, 259-266.

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. It specifically 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 [2, 3, 9, 12, 14].

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 and university courses are using the collection [4, 8, 9]; large educational programs were developed and implemented in German schools [15]. Whereas the Shoah Foundation initially had not transcribed its 53,000 interviews, CeDiS created 908 German-language (plus 50 foreign-language) transcriptions following specific guidelines. These transcripts are time-coded every minute enabling full text search over all 958 interviews [1]. 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 Archive*'s new commercial provider ProQuest [17]. In 2017, the Shoah Foundation provided another 984 transcripts in English language in their online archive [21].

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 [16]. Most of them were transcribed, translated into German, indexed and made available in an online archive together with accompanying photos and documents. 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, over 8,000 archive users – students, researchers, teachers, and other interested persons – have been granted access to the collection.

Faceted search options allow 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 [3]. 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.

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 the satellite imagery provided by Google Maps, 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 data visualization, digital mapping contextualizes the survivors' testimonies within current local cultures or memory – or forgetting.

In 2018, the archive will get 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 like *Archiv Deutsches Gedächtnis* and *Memories of the Occupation in Greece* use the same technology as the *Forced Labor 1939–1945* project, adding project-specific functionalities.

3 Digital Perspectives

The digital curation of interview collections faces a number of problems. Digital preservation strategies have to 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 has to 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 corresponding 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 or the oralhistory.eu group. CLARIN workshops in Oxford 2016 [5] and Munich 2018 [6] 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 language oral history interviews, developed by researchers from Pilsen, is accessible at the

Malach Center for Visual History in Prague [18]. Dutch scholars are currently combining a sequence different tools for various curation steps into a "transcription chain" [7].

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. In conversation analysis, the focus is on the dialogue with the interviewer, a dialogue which is somewhat hidden, because the camera focuses on the narrator, and which often gets overlooked by historians who are more interested in fact-finding than in the co-construction of the testimony. Even oral historians are being told "An interview is not a dialogue" in introductory texts [10].

Some preliminary studies have proved that the interview archives can be very useful for comparative studies – even without applying quantitative methods [8, 9, 11, 12]. For a more systematic approach, however, interview transcripts have to be standardized in a really machine-readable form. Therefore, CeDiS is working on a TEI schema for oral history interviews, building on the TEI guidelines for transcribed speech [20].

New research perspectives can open up, when "oral history meets linguistics" – the title of a 2015 workshop in Freiburg [12]. 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 listening – will have to be interpreted through a careful listening to individual testimonies – or close listening.

4 Cross-collection Search

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 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 or Holocaust research – are working with diverse metadata standards. Some library-based oral history collections in the US have created MARC21 records for their interviews; some European collections try to adapt 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.

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 "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.

5 Perspectives

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.

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.

In general, the digitized perception of historical sources 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 importance and meaning of "context" differs between disciplines: While linguists are used to work 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.

The digital de-contextualization gets more profound, when the researchers use a digital environment to search interview segments about a specific topic instead of listening to complete testimonies. They can find and copy some nice quotations but often will not understand their meaning correctly without knowing their context within the whole testimony.

While de-contextualization is inherent to digital research, digital environments for oral history allow to work 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 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.

References

[1] Abenhausen, Sigrid, Apostolopoulos, Nicolas, Körte-Braun, Bernd, Nägel, Verena (Eds.), Zeugen der Shoah: Die didaktische und wissenschaftliche Arbeit mit Video-Interviews des USC Shoah Foundation Institute, Berlin: Freie Universität 2012

[2] Apostolopoulos, Nicolas, Barricelli, Michele, Koch, Gertrud (Eds.), Preserving Survivors' Memories. Digital Testimony Collections about Nazi Persecution: History, Education and Media (Education with Testimonies, Vol. 3), Berlin: Stiftung EVZ, 2016, http://www.stiftung-evz.de/index.php?id=1655, 8 Sep 2018

[3] Apostolopoulos, Nicolas, Pagenstecher, Cord (Eds.), Erinnern an Zwangsarbeit: Zeitzeugen-Interviews in der digitalen Welt, Berlin: Metropol 2013

[4] Bothe, Alina, Brüning, Christina Isabel (Eds.), Geschlecht und Erinnern im digitalen Zeitalter. Neue Perspektiven auf ZeitzeugInnenarchive, Berlin: Lit 2015

[5] 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

[6] CLARIN workshop: "Oral History: Users and their scholarly practices in a multidisciplinary world, Munich, 19.-21.9.2018, http://oralhistory.eu/workshops/munich, 8 Sep 2018

[7] Hessen, Arjan van et al., Oral History & Technology: Transcription Chain, http://oralhistory.eu/workshops/transcription-chain, 8 Sep 2018

[8] Michaelis, Andree, Erzählräume nach Auschwitz: Literarische und videographierte Zeugnisse von Überlebenden der Shoah, Berlin: Akademie 2013

[9] Nägel, Verena, 'Zeugnis – Artefakt – Digitalisat. Zur Bedeutung der Entstehungs- und Aufbereitungsprozesse von Oral History-Interviews', in: Eusterschulte, Anne/Knopp, Sonja/Schulze, Sebastian (Eds.): Videographierte Zeugenschaft. Ein interdisziplinärer Dialog, Weilerswist: Velbrück Wissenschaft 2016, 347-368

[10] 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

[11] Pagenstecher, Cord, '"We were treated like slaves." Remembering forced labor for Nazi Germany', in: Mackenthun, Gesa, Hörmann, Raphael (Eds.), Human Bondage in the Cultural Contact Zone. Transdisciplinary Perspectives on Slavery and Its Discourses, Münster: Waxmann 2010, 275-291

[12] Pagenstecher, Cord, Pfänder, Stefan, 'Hidden Dialogues: Towards an Interactional Understanding of Oral History Interviews', in: Kasten, Erich, Roller, Katja, Wilbur, Joshua (Eds.), Oral History Meets Linguistics, Fürstenberg/Havel: SEC Publications 2017, 185-207

[13] Pagenstecher, Cord, 'Testimonies in Digital Environments. Comparing and (De-)Contextualizing Interviews with Holocaust Survivor Anita Lasker-Wallfisch', in: Oral History, Autumn 2018 (in preparation)

[14] Pagenstecher, Cord, Tausendfreund, Doris, 'Interviews als Quellen der Geschlechtergeschichte. Das Online-Archiv "Zwangsarbeit 1939-1945" und das „Visual History Archive" der „USC Shoah Foundation"', in: Bothe, Alina, Brüning, Christina Isabel (Eds.), Geschlecht und Erinnerung im digitalen Zeitalter. Neue Perspektiven auf ZeitzeugInnenarchive, Berlin/Münster: Lit 2015, 41-67.

[15] Pagenstecher, Cord, Wein, Dorothee, 'Learning with Digital Testimonies in Germany. Educational Material on Nazi Forced Labor and the Holocaust', in: Llewellyn, Kristina R., Ng-A-Fook, Nicholas (Eds.), Oral History and Education. Theories, Dilemmas, and Practices, New York 2017, 361-378

[16] Plato, Alexander von, Leh, Almut, Thonfeld, Christoph (Eds.), Hitler's Slaves: Life Stories of Forced Labourers in Nazi-Occupied Europe, New York/Oxford: Berghahn 2010

[17] ProQuest, 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

[18] Stanislav, Petr, Svec, Jan, Ircing, Pavel, 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

[19] Thonfeld, Christoph, Rehabilitierte Erinnerungen? Individuelle Erfahrungsverarbeitungen und kollektive Repräsentationen von NS-Zwangsarbeit im internationalen Vergleich, Essen: Klartext 2014

[20] Text Encoding Initiative, P5: 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

[21] USC Shoah Foundation, 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

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

Pawel Kamocki
IDS Mannheim / Université Paris Descartes
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
Universität zu Köln / IDS Mannheim / Universität Heidelberg
andreas.witt@uni-koeln.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 infrastructure 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. In principle, every reproduction or communication to the public of copyright-protected material requires authorisation from the rightholder. 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 related rights in the information society (hereinafter: InfoSoc Directive) contains (in its art. 5) a limitative 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), 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.

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, at least for research purposes, has been discussed at least since 2011, i.e. the publication of the Hargreaves review (Hargreaves, 2011). In 2013, a group on Text and Data Mining was created within the Stakeholder's Dialogue "Licenses for Europe" (European Commission, 2013). The academic community, unhappy with the adopted approach (focused on licensing rather than on statutory exceptions), largely withdrew from the process (LIBER, 2013). 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.

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), 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 (art. L. 122-5, 10° of the French Intellectual Property Code), 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 decree was supposed to clarify; unfortunately, a proposal for such a decree was rejected in 2017 (Langlais, 2017) 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 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 (cf. §23 UrhG). Interestingly, the new law expressly uses the word "corpus". Such a "corpus" can be shared with a "specifically limited circle of persons" (presumably a research team, perhaps also multi-institutional). However, once the research is over, the corpus has to be deleted or transferred to a library or an archive for permanent storage. The new German exception is expressly non-overridable by contractual clauses (cf. §60g), 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 (VG Wort; cf. §60h UrhG).

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, cf. art. 27 of the Polish Copyright Act). 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 of on copyright in the Digital Single Market (European Commission, 2016). 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) advocates to a solution similar to the one adopted in Germany, requiring payment of equitable remuneration and deletion of the compiled corpus upon the completion of the project. Its draft also stipulates that "lawful access" to mined works has to "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 (CULT, 2017). 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 that mining should be allowed also for other purposes than scientific research (IMCO, 2017). The Industry, Research and Energy Committee (ITRE) took a similar position (ITRE, 2017). 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 (JURI, 2017).

On 25 May 2018, the European Council (under the Bulgarian presidency) published its version of the proposal (European Council, 2018), which 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). 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 changes the paradigm from "TDM only with permission" to "open for TDM by default", but does not really provide the 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 (JURI, 2018), adopted on 29 June 2018 was partly inspired by the Council's proposal. JURI postulated 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.

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). This means that the adoption process has been slowed down, but not completely interrupted. However, it is still impossible to predict the content of the soon-to-be-adopted TDM exception. The Directive will probably have to be implemented twelve months after its entry into force (as per — seemingly undisputed — art. 21 of the Commission's draft).

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

While language researchers will receive substantial benefits and some legal certainty from the new TDM exceptions, 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 (because everything can be freely mined, at least for non-commercial research purposes), researchers have fewer incentives to care about proper licensing and sharing of their datasets and results (e.g. within research infrastructures) (Suber, 2012). This may in turn considerably reduce the "knowl-

edge commons" (i.e. immaterial resources that — due to proper licensing — can be freely accessed and re-used by anyone and for any purpose (Hess, Ostrom, 2006) 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 (von Hippel, 2017) may paradoxically be more difficult. In order to 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.

Reference

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

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

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

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

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 it allows 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. To this goal, we first present some fundamental concepts of data protection relevant for language research and then focus on the models that certain EU member states use to regulate data processing for research purposes.

1 Introduction

Language resources contain material subject to various legal regimes. For instance, language resources can 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 consent of the data subject 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 person.[1] In the CLARIN Virtual Language Observatory (VLO), 95,502 language resources[2] could contain personal data.[3]

[1] 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).
[2] Resource type: Audio, Radio, Sound, Speech, Spontaneous, Television or Video.

Although the General Data Protection Regulation[4] (GDPR) provides a general framework for personal data protection, it leaves a certain degree of freedom for the EU Member States to regulate data processing for research purposes. It means that they can adopt different regulatory models. This article preliminarily maps and provides insights into different models.[5] 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 Article 29 Working Party[6] (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). A natural person can be identified by reference to the identifier (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).

Voice can be considered biometric data (see González-Rodríguez et al. 2008; Jain et al. 2004).[7] Biometric data for uniquely identifying a natural person belongs to a special category of personal data[8] the processing of which is even more restricted than for other personal data. The similar case is with 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 as well. Therefore, the requirements concerning the processing of special categories of personal data do not need to be followed until the oral speech contained in language resources is not 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.

[3] 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).
[4] The GDPR is applicable in all EU Member States from 25 May 2018. It replaces the Data Protection Directive.
[5] For lack of space not all the EU countries are addressed in the present paper.
[6] 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.
[7] 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).
[8] 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".

3 Processing personal data for research purposes

3.1 GDPR requirements

The GDPR provides six legal grounds for processing personal data: consent, performance of a contract, compliance with a legal obligation, protection of the vital interests, public interest and legitimate interests (Art. 6). If possible, the processing for research purposes should rely on consent (for further discussion on consent see WP29 2017). This paper concentrates on cases where there is no consent, and another legal ground is needed.[9] According to WP29, the legitimate interest can serve as a legal ground for processing personal data in the research context (2014b: 24-25). The concept of legitimate interests is rather complicated and requires weighing different interests.[10]

The GDPR establishes the following requirements for processing[11] data in the research[12] context (Art. 89):

1) processing is subject to appropriate safeguards (technical and organisational measures (e.g., pseudonymisation) to ensure data minimisation);

2) the EU Member States are allowed to limit the following data subject's rights[13]: the right of access, right to rectification, right to restriction of processing and right to object.

Since the GDPR offers the Member States some flexibility to specify requirements concerning processing for research purposes[14], it is necessary to analyse national laws.

3.2 National models

The first example to be considered is **Estonia**. The Estonian draft Act on Personal Data Protection (Draft PDPA 2018a) sets the following requirements for 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 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.

[9] As a matter of fact, one option to avoid problems with personal data protection is the anonymisation of data used for language research (for further discussion on anonymization of data see WP29 2014a).

[10] 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. 6).

[11] Including further processing of previously collected data (Art. 5).

[12] The GDPR defines research in a broad manner. It covers "technological development and demonstration, fundamental research, applied research and privately funded research" (Recital 159).

[13] The GDPR itself limits the right to erasure ('right to be forgotten') in research context (Art. 17).

[14] MS have flexibility to regulate data processing in other areas as well (e.g., processing for the purpose of academic artistic or literary expression (Art. 85)).

According to the **Finnish** model, the draft Act on Personal Data Protection (Draft PDPA 2018b) and its preamble outline the following for processing personal data for scientific research[15]:

1) Personal data may be processed by university researchers according to § 6.1e in the GDPR, i.e. *performance of a task carried out in the public interest* based on the university's legal mandate to do research. Universities also have the right to archive data for scientific and historical research based on §9.2j in GDPR, i.e. *processing is necessary for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes*.

2) For a researcher to use datasets containing personal data for secondary purposes, a research plan is required, but its primary purpose from the GDPR perspective is to document that the research is of a scientific or historical nature. If the personal data is sensitive, a set of protective measures specified in the law also need to be carried out and documented.

3) Also, 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. The motivation for why the processing is necessary should be included in the research plan (mentioning the Principal Investigator). It is possible even for sensitive personal data, but then the research plan also needs to describe how it fulfils the ethical standards in the field of research and needs to be delivered to the Data Protection Authority at least 30 days ahead of starting the processing.

The next example is **Lithuania**. In contrast with the current regulation, the newly enacted Lithuanian Law Amending the Law on Legal Protection of Personal Data (LLPPD 2018), which implements the provisions of GDPR, contains no special provisions dealing with the research exemption and no prior checking procedure is required. 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 has to rely directly on and comply with the general provisions of GDPR.

Italy has not yet released detailed legislation to link the Italian Personal Data Protection Code (IPDPC) to the GDPR. However, the subject is addressed at the policy level. The Law of European Delegation (LED) establishes that the Italian Government enacts detailed legislative decrees with the aim of adapting the national framework specifying the GDPR (art. 13). In the Delegation Law, a series of objectives are defined: (i) the abrogation of the provisions of the Privacy Code which are not compatible with the provisions of the GDPR; (ii) the coordination and integration of the Code on personal data protection, in order to implement the non-directly applicable GDPR provisions; (iii) the adoption of specific implementing Acts by the Italian data protection Authority (i.e. "Garante Privacy"), for the purposes envisaged by the GDPR.

It is likely that clause (iii) also deals with the use of data for research purposes. The data protection authority assesses whether the provisions contained in Attachment A.2 of the Privacy Code are fully compliant to the GDPR rules, or whether the additional regulatory action is necessary.

France and **Germany** have adopted very different views on the processing for research purposes. The new German law[16] seems to be as favourable to researchers as possible, while the new French law is much more conservative. For example, the French law requires authorisation from the national data protection authority for processing of special categories of personal data for public research purposes, whereas the German law generally allows for such data to be processed for such purposes simply if the processing is subject to appropriate safeguards and if it passes the "balance of interests" test. Likewise, any derogations from rights of data subjects in the French law seem to be limited to specific cases of medical research. Time shall tell which approach proves better.

[15] As far as possible, Finland plans to maintain its existing practice for collecting and using research data.

[16] It shall be kept in mind that the German Federal Data Protection Act (Bundesdatenschutzgesetz, BDSG) only applies to 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 (Landesdateschutzgesetze, LDSG). To the best of our knowledge, no LDSGs has yet been updated to conform to the GDPR. Therefore, for now the situation regarding processing of personal data for research purposes in German universities is not entirely clear.

In **Greece**, a Draft Bill for Personal Data (PDPA 2018c) implementing the GDPR has been recently released for public consultation (completed on March 5, 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.

4 Conclusion

As argued in the paper, the development and use of language resources often involve processing of personal data. Although the GDPR is applicable in the whole EU, it allows the Member States to specify processing for research purposes. This means that in addition to the GDPR, researchers that wish to construct and use LRs for language research must further follow national requirements.

Reference

[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)

[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)

[Draft PDPA 2018a] Estonian Draft Act on Personal Data Protection (Isikuandmete kaitse seaduse eelnõu) (22.08.2018). Available at https://www.riigikogu.ee/tegevus/eelnoud/eelnou/5c9f8086-b465-4067-841e-41e7df3b95af/Isikuandmete%20kaitse%20seadus (3.4.2018)

[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)

[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

[Personal Data Protection Act]. Personal Data Protection Act. Entry into force 01.01.2008. English translation available at https://www.riigiteataja.ee/en/eli/507032016001/consolide (3.4.2018)

[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)

Toward a CLARIN Data Protection Code of Conduct

Pawel Kamocki
IDS Mannheim / Université Paris Descartes
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
Universität zu Köln / IDS Mannheim / Universität Heidelberg
andreas.witt@uni-koeln.de

Abstract

This abstract discusses the possibility to adopt a CLARIN Data Protection Code of Conduct pursuant art. 40 of the General Data Protection Regulation. Such a code of conduct would have important benefits for the entire language research community. The final section of this abstract proposes a roadmap to the CLARIN Data Protection Code of Conduct, listing various stages of its drafting and approval procedures.

1. Overview of the General Data Protection Regulation (GDPR)

The General Data Protection Regulation (hereinafter: GDPR) is the EU Regulation 2016/69 of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and the free movement of such data. On 25 May 2018 it became directly applicable in all the EU Member States and replaced the Personal Data Directive of 1995. Unlike a directive, which requires implementation in the national law of each Member State, a regulation applies directly and supersedes national rules.

The GDPR is built around the same concepts as the Personal Data Directive. The essential notion of personal data remains the same: it is defined as "any information relating to an identified or identifiable natural person ('data subject')" (art. 4 of the GDPR). According to art. 5 of the GDPR, processing of personal data (i.e. any operation performed on such data) shall comply with the principles of lawfulness, fairness and transparency; purpose limitation; data minimisation; accuracy; storage limitation; integrity and confidentiality and accountability. In principle, processing is lawful if the data subject has given his informed consent; exceptionally, other grounds for lawfulness are also possible (as listed in art. 6 of the GDPR). Special categories of personal data (i.e. revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership, health, sex life or sexual orientation) benefit from stricter protection (art. 9 of the GDPR).

The GDPR also reinforces the rights of the data subjects (such as right to information, access, rectification, erasure, restriction, right to data portability or right to object) and — correlatively — obligations of data controllers (e.g. implementing "privacy by design and by default", keeping a record of processing activities or carrying out a data protection impact assessment) (Voigt, von dem Bussche, 2017).

2. Research Under the GDPR

The rules of the GDPR are meant to protect individual rights and freedoms, but their strict application to certain activities might have a chilling effect e.g. on freedom of research or artistic expression. This

is why the European legislator provided for some exceptions concerning processing of personal data in certain special situations, including research or archiving in the public interest. Art. 89 of the GDPR allows for significant derogations from the general framework in such context; however, for fear of an excessive intervention (and violating the subsidiarity principle), the European legislator left some leeway to the Member States concerning this matter. Therefore, the exact rules and procedures concerning processing of personal data for research purposes may vary slightly between Member States (Kelli et al., 2018). Nevertheless, for any derogations to apply, the processing must be "subject to appropriate safeguards for the rights and freedoms of the data subject". The GDPR remains vague as to what can be construed as such an 'appropriate safeguard', providing only one example: pseudonymisation. It is obvious, however, that other 'appropriate safeguards' are also envisageable.

3. Bottom-up Standardisation under the GDPR

Given its huge scope, the multitude of interests at stake, and the rate at which technology is advancing, the GDPR necessarily remains vague and leaves room for interpretation in order not to quickly become obsolete. This is also why the European legislator created some instruments for bottom-up standardisation which will hopefully help 'fill in the blanks' and contribute to the creation of sector-specific best practices while at the same time guaranteeing a degree of flexibility. These instruments include certification (e.g. data protection marks and seals) and codes of conduct. In this paper, we would like to discuss how the latter could be implemented within CLARIN and the benefits this would achieve.

4. The Scope of a Code of Conduct

Art. 40 of the GDPR invites "associations and other bodies representing categories of data controllers or processors" to prepare codes of conduct "intended to contribute to the proper application of [the GDPR], taking into account the specific features of various processing sectors". In particular, such codes of conduct could "calibrate the obligations of controllers and processors, taking into account the risk likely to result from the processing" (recital 98).

CLARIN is undoubtedly a body that is competent to draft a code of conduct concerning processing of personal data for the purposes of research in the field of linguistics and language technology.
Such a code of conduct should address such questions as (cf. art. 40(2) a)):

- fairness and transparency of processing (when shall the processing be considered "not fair"? To what extent the data subject should be informed about the processing and/or involved at the various stages of the processing?);
- the legitimate interests of the controllers, i.e. in which types of situations researchers can be exempted from the obligation to obtain the data subject's consent?
- the pseudonymisation and anonymisation techniques: which are appropriate for various types of research projects?
- the collection of personal data: in what conditions shall personal data be collected from data subjects, particularly if they belong to a group that merits specific protection (e.g. children, the elderly, immigrants);
- the exercise of rights of data subjects (such as the right to be forgotten or the right of access and rectification);
- specific technical and organisational measures taken to guarantee the security of personal data processing, and to achieve "privacy by design and by default" (cf. art. 25 of the GDPR);
- out-of-court resolution of disputes related to the processing of personal data (e.g. establishing an independent expert arbitration board);
- transfer of personal data to non-EU countries and organisations.

These questions should be answered not only from the legal, but also — and perhaps more importantly — from the technical and the ethical perspectives. This is why in our opinion a multidisciplinary ad-hoc Working Group should be created to draft a CLARIN Data Protection Code of Conduct. This group could consist of delegated members of the Legal Issues Committee, the Committee on Technical Centres, the Standards Committee and possibly also external experts or representatives of data protection authorities.

5. Potential Advantages of a Code of Conduct

A CLARIN Data Protection Code of Conduct would have some obvious benefits for the entire community. First and foremost, it would allay doubts related to the application of the GDPR to the processing of personal data for language research purposes, especially with regards to the special framework applicable to research (art. 89 of the GDPR, cf. supra). This would provide for a greater degree of legal security among researchers and, in the long run, lower the costs of carrying out new projects. In addition, a CLARIN Data Protection Code of Conduct could provide for more consistency within and across national CLARIN consortia and thereby help achieve the GDPR's main purpose, while at the same time spreading and perpetuating good practices in the community, especially among young researchers. Furthermore, the Code would solve another problem related to cross-border sharing of research data, enabling transfer of personal data to partners in non-EU countries who adhere to the CLARIN Data Protection Code of Conduct (cf. art. 46(2)(e) of the GDPR). Last but not least, such a Code could be adhered to also outside of the CLARIN community, thereby increasing the visibility of the CLARIN infrastructure.

The benefits of a CLARIN Data Protection Code of Conduct, both immediate and long-term, are well worth the necessary effort invested in the drafting and approval of the Code.

6. Toward a CLARIN Code of Conduct: a Roadmap

In this section we outline various stages involved in the drafting and approval of the CLARIN Data Protection Code of Conduct. The procedure is outlined in art. 40 of the GDPR.

As mentioned above, we believe that a multidisciplinary ad-hoc Working Group (consisting of internal and external experts on legal, technical and ethical aspects of personal data processing) shall be established and assigned the task of drafting the Code of Conduct.

The first step shall consist of identifying internal and external stakeholders. Possibly, a questionnaire can be created and distributed among the stakeholders, covering all the issues that the Code of Conduct should address (cf. above). The Working Group should collect feedback from various CLARIN projects and institutions with experience in processing personal data both in the field of linguistics and social science. Whenever feasible, representatives of various groups of data subjects should also be included in the process.

A first draft of the Code should be based on the responses collected from the stakeholders. This draft could then be subject to a round of consultation among these stakeholders and other CLARIN bodies.

Once a final version of the draft Code is developed, it shall be submitted to a competent supervisory authority. In our opinion, the draft shall be simultaneously submitted to several national supervisory authorities in various CLARIN countries. This would increase the visibility of the Code of Conduct, as well as the chances for the draft to pass to the next step.

A supervisory authority can approve the draft Code if it finds that it provides for sufficient appropriate safeguards. An approved code of conduct can be adhered to by data controllers and processors. However, such an approval is only valid on the national level and has no cross-border effect (i.e. a code approved in the Netherlands is not automatically regarded as providing appropriate safeguards in Germany and vice versa). Nevertheless, if a supervisory authority decides that the draft code is relevant for many EU Member States, it shall, before approving the Code, submit it to the European Data Protection Board (a body composed of the head of one supervisory authority of each Member State and of the European Data Protection Supervisor, cf. art. 68 of the GDPR), which shall provide its opinion on the draft code.

Moreover, if the Board decides that the draft code provides for appropriate safeguards, it shall notify the European Commission. The EC may then decide to grant the code conduct a general validity in the EU; in other words, it would become a complement to the GDPR and could be adhered to by all data controllers and processors in the EU (i.e. even in those countries where the draft were not submitted to or approved by a supervisory authority). Such a generally valid code of conduct could be also adhered to by data controllers in non-EU member states (e.g. universities and research institutions in the US), which would facilitate the transfer of relevant personal data to these data controllers.

A code of conduct should contain mechanisms for monitoring compliance. This monitoring is mandatory and carried out by a body which is accredited by a supervisory authority and "has an appropriate level of expertise in relation to the subject-matter of the code". It is advisable that once a CLARIN Data Protection Code of Conduct is adopted, the national CLARIN consortia or CLARIN centres apply for accreditation to monitor compliance with the Code.

The action plan described in this abstract will require time and effort by many CLARIN stakeholders; however, it is our belief that the approval of a CLARIN code of conduct in a single member country, or several slightly different codes in various member countries, would be a significant advancement for the entire CLARIN community.

Reference

Regulation 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) (articles 4, 5, 9, 25, 40, 41, 46, 68 and 89, recitals 98 and 99).

Voigt, P. and A. von dem Bussche (2017). "The EU General Data Protection Regulation (GDPR): A Practical Guide", Springer 2017.

Kelli, A. et al. (2018). "Processing personal data without the consent of the data subject for the development and use of language resource", to be presented at the CLARIN Annual Conference 2018.

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 central 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.

1 Introduction

Lärka,[1] is an Intelligent Computer-Assisted Language Learning (ICALL) platform developed at the CLARIN B Center Språkbanken (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 of Swedish – second/foreign language learners 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., 2013) for querying lexical resources and language technology tools. Due 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 was also 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, 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 the exercises which went into research on metalinguistic awareness. Lärka also became 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, a lot of users are unknown to us, but we can see from the logs that Lärka is being used beyond the reported schools and universities.

Nowadays, Lärka is being adapted into a central 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. This addresses an obvious need within CLARIN, as evidenced by the great interest in the recent CLARIN workshop on "Interoperability of Second Language Resources and Tools".[2]

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

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

2 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).

Exercises for linguists Students learning grammatical analysis are in great need of exercises and feedback on their analysis. Lärka offers exercises for linguistic analysis of 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 (beginner and intermediate), whereas the third (semantic roles) is only available as one level. Students can choose whichever mode they want to use: self-study, diagnostic test or test. In both self-study and test mode the actual categories practiced can also be chosen, whereas the diagnostic test automatically gives a set of exercises (3 for each of the main categories, i.e. 33 for eleven parts of speech). After the diagnostic test, a summary is provided which can be emailed to one's teacher for further comments or to oneself in order to study the examples further or to be able to track one's learning.

Exercises for language learners Lärka offers a number of exercises for learners of L2 Swedish. 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. For vocabulary, distractors are chosen to be of the same word class as the target word. For inflection exercises, we look up all morphological forms of the target word and use those as distractors. This morphological selection can be 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.

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

Another exercise is the listening exercise *Liwrix*. This exercise makes use of Text-to-Speech (TTS) technology by SitePal[3] to dynamically generate audio of either single words, phrases or sentences.

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) 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 level according to two approaches, namely *first-occurrence* (Gala et al., 2013; Gala et al., 2014) and *threshold* (Alfter et al., 2016). *First-occurrence* assigns the first level where a word occurs as target level while *threshold* assigns the level where a word occurs significantly more often than at a previous level as target level, with the threshold of significance set at 30%, with the target level corresponding to the CEFR level at which a word should be understood or produced by the learner for receptive and productive vocabulary respectively.

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

[3] `sitepal.com`

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. The students felt that this was of great use and definitely thought that the platform should be used in the future. 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. accompanied by an immediate automatic feedback.

3 Lärka as research infrastructure

Lärka is used as an infrastructure for (1) collection of data from learners through their interaction with the platform, i.e. exercise logs, for (2) text-level annotation of learner essays and course book texts, as well as (3) for experimentation and visualization of the ongoing research in support of language learning. Among other things, we integrate natural language processing tools and algorithms for corpus example selection, text assessment and automatic exercise generation. 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.

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 perform 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 co-occurrence measures, as well as the absence of anaphoric expressions and sensitive vocabulary (e.g. profanities), just to name a few.

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. 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. We use the aforementioned lists SVALex and SweLLex to markup 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.

Lexical complexity prediction Based on the word lists SVALex and SweLLex, 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). An interested user can test a specially devoted interface to get predictions about the complexity of a word and its target level (receptive versus productive).

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. 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.

4 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 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.

Lars Borin and Anju Saxena. 2004. Grammar, incorporated. *Copenhagen studies in language*, 30:125.

Lars Borin, Markus Forsberg, and Johan Roxendal. 2012. Korp-the corpus infrastructure of språkbanken. In *LREC*, pages 474–478.

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

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).

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

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.

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.

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.

In this paper we focus on the culture of giving for education. The collected corpus comprises texts with a time span of 80-100 years. The task is 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 economical context. Thus, the initial steps include: adaptation of the existing tools, the creation of a specialized corpus, the creation of a web-based concordance tool and presenting useful statistics and content analysis over the corpus.

Our work aims towards the ideas presented in Fokkens et al., 2018. The similarities are as follows: we also 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 are not working 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 task. Last, but not least, our NLP chain at the moment has not incorporated any Wordnet concepts or semantic roles. These modules will be added.

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 NE recognition and linking module, a lemmatizer and a parser. The main problems in the tokenization were related to the proper handling of the abbreviations. 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. Thus, the main difficulty was the assigning the word form of a rare word to its lemma. The parser also depends on the previous steps. Apart from that, the parser had problems with the syntactically different codifications in the contemporary Bulgarian and the texts in previous times.

3 Statistics and Content Analysis

Two of the most effective ways for observing the behaviour of various words, collocations and phrases are: the statistics over the keywords in some domain and 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[2] 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* (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 the all 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.

[2] http://webclark.org/

For extracting the key words from the corpus we used the program *AntConc* (see Anthony, 2014). The visualization (in form of the word cloud) has been done through the following web service: https://www.jasondavies.com/wordcloud/. In Table 1 below we give the first ten most frequent words from the lists with ranked keywords for the three periods:

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 1: The first 10 words with the highest rank, presented per period.

While in the first period the concept of *will* dominates, in the second and third one this is the concept of *fund*. As a whole, mainly the terminology has changed, not the content.

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).

4 Named Entity Annotation and Linked Open Data

We annotated all the Named Entities 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) will support the better understanding of the culture of giving, motivation for donation, etc.

For each document we explicated all the persons mentioned in it. Also as a metadata we recorded the persons that 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.

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. 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 specialized corpus CoDar would be useful for investigating sociological, historical, cultural, education policy making and language phenomena. Concerning the NLP processing it can be concluded that names identification is less of a challenge in comparison to abbreviations, variety of lexica and wordforms, syntactic structures over the big time span.

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]

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.

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.

Multilayer Corpus and Toolchain for Full-Stack NLU in Latvian

Normunds Grūzītis
AI Lab at IMCS
University of Latvia
normunds.gruzitis@ailab.lv

Artūrs Znotiņš
AI Lab at IMCS
University of Latvia
arturs.znotins@ailab.lv

Abstract

We present a work in progress to create a multilayer text corpus for Latvian. The broad application area we address is cross-lingual natural language understanding (NLU), and the aim of the corpus creation is to develop a data-driven toolchain for NLU in Latvian. Both the multilayered corpus and the downstream applications are anchored in cross-lingual state-of-the-art representations: Universal Dependencies (UD), FrameNet, PropBank and Abstract Meaning Representation (AMR). The corpus and the toolchain also include named entity and coreference annotation layers. We are planning to add the data sets and the tools to the CLARIN infrastructure.

1 Introduction

NLU systems rely, explicitly or implicitly, on syntactic and semantic parsing of text. State-of-the-art parsers, in turn, typically rely on supervised machine learning which requires substantial language resources – syntactically and semantically annotated text corpora.

In the industry-oriented research project "Full Stack of Language Resources for Natural Language Understanding and Generation in Latvian" (Gruzitis et al., 2018b), we are creating a balanced text corpus with multilayered annotations (Figure 1), adopting widely acknowledged and cross-lingually applicable representations: UD (Nivre et al., 2016), FrameNet (Fillmore et al., 2003), PropBank (Palmer et al., 2005) and AMR (Banarescu et al., 2013).

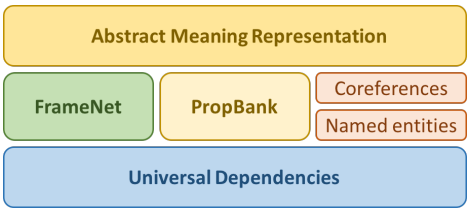


Figure 1: Annotation layers of the corpus.

The UD representation is automatically derived from a more elaborated manually annotated hybrid dependency-constituency representation (Pretkalnina et al., 2018). FrameNet annotations are added manually, based on the Berkeley FrameNet frame inventory (Nespore-Berzkalne et al., 2018). The FrameNet annotation process is guided by the underlying UD annotations. Consequently, frame elements are represented by the root nodes of the respective subtrees instead of text spans; the spans can be easily calculated from the subtrees. The PropBank layer is automatically derived from the FrameNet and UD annotations, provided a manual mapping from lexical units in FrameNet to PropBank frames, and a mapping from FrameNet frame elements to PropBank semantic roles for the given pair of FrameNet and PropBank frames. Draft AMR graphs are to be derived from the UD and PropBank layers, as well as auxiliary layers containing named entity and coreference annotation. The semantically richer FrameNet annotations (cf. PropBank) are also helpful in acquiring more accurate AMR graphs.

Although we focus on Latvian, we believe that our experience and findings can be useful for the systematic creation of similar multilayered corpora and toolchains for other less-resourced languages.

2 Treebank

In this 3-year project, we are aiming at a medium-sized[1] corpus: 10–15 thousand sentences annotated at all layers. Therefore it is important to ensure that the multilayer corpus is balanced not only in terms of text genres and writing styles but also in terms of lexical units (LU).

A fundamental design decision is that the text unit is an isolated paragraph. The treebank therefore consists of manually selected paragraphs from different texts of various types. Paragraphs are selected in proportions from a balanced 10-million-word text corpus: around 60% are newswire texts, around 20% is fiction, around 10% are legal texts, around 5% is spoken language (transcripts); the rest is miscellaneous.

As for LUs, our goal is to cover at least 1,000 most frequently occurring verbs, calculated from the 10-million-word corpus. Since the most frequent verbs tend to be the most polysemous, we expect that the number of LUs (verb senses w.r.t. semantic frames) will be considerably larger – around 2,000 units.

To capture the language-specific details and to accommodate the linguistic tradition on the one hand, and to meet the goal of the cross-lingual application on the other hand, the treebank annotation is provided in two complementary formalisms. First, the selected paragraphs are manually annotated according to a hybrid dependency-constituency grammar model developed with the linguistic tradition in mind (Barzdins et al., 2007; Pretkalnina et al., 2011). Second, the hybrid annotation is automatically converted to UD to achieve the cross-lingual compatibility (Pretkalnina et al., 2018), as well as to provide training data for efficient and robust parsers.

3 Named Entities and Coreferences

Named entities (NE) are essential for most NLU tasks, since they link the textual content to the real world, making the extracted facts (frames) meaningful for a knowledge base population task, for instance. From the multilayer corpus perspective, the AMR annotation heavily relies on NE recognition and linking, and on within-sentence coreference resolution, using the re-entrancy representation. This also allows for connecting individual AMR graphs and subgraphs into a wider context.

We annotate the following set of NE categories: *person*, *organization*, *geopolitical entity* (GPE), *location*, *product*, *time* (relative or absolute date, time, or duration), *event*, and *entity* (entities of other categories that occur rarely but could be re-considered in future). We mostly follow the MUC-7 annotation guidelines (Chinchor, 1998) which we have extended for compatibility with the top-level NE categories specified by the AMR guidelines. For serialization of the named entity layer, we use a version of the CoNLL-2003 data format on top of CoNLL-U.

In practice, many NEs contain references to other NEs. We annotate hierarchical entities, which enables us to exploit the annotations of either the outer or inner entities, or both, when annotating coreferences or deriving AMR graphs. This would not be possible with a flat annotation scheme. Furthermore, hierarchical annotations not only allow for the development of an automatic hierarchical named entity recognizer (NER) but also provide more training data for the development of a flat NER.

We approach coreference annotation in a pragmatic way, focusing on precision, and annotating coreferences only within the paragraph boundaries. This allows to annotate a lot of various text units, and it makes the annotation process easier and less error prone. We annotate pronominal and nominal noun phrases referring to real-word entities, and non-specific mentions if they are antecedents of pronouns. Bridging relations, discontinuous expressions, split antecedents and zero anaphora are ignored.

In addition to annotating the NE spans and categories, we also specify a corresponding Wikipedia identifier (URI) if one exists. For training a named entity linker (NEL), such corpus would be considered a very small one, but it will be helpful for evaluating a NEL. For this reason, we have specially included text units mentioning different persons with the same name, for instance. The manually verified Wikipedia identifiers are also useful when generating draft AMR graphs.

[1] Latvian UD Treebank is already categorized as a big one: `http://universaldependencies.org/conll18/results-las.html`

4 Semantic Frames

The annotation of PropBank frames is relatively more simple if compared to FrameNet, since PropBank frames are less abstract, and their semantic roles directly follow from the syntactic verb argument structure. However, we start with annotating FrameNet frames (Gruzitis et al., 2018a), and we derive PropBank annotations semi-automatically from the FrameNet and UD annotations.

While treebank, named entity and coreference annotation is done paragraph by paragraph, it is not a productive workflow for annotating abstract semantic frames. Instead, a concordance view is required, so that the linguist can focus on a target verb and its different senses (frames), without constantly switching among different sets of frames. This also improves the annotation consistency.

To provide such environment, we automatically extract UD-annotated sentences from the finalized paragraphs containing the requested target verb, and we store the result in a temporary CoNLL-U file. When more paragraphs are finalized at the UD layer, they are included in the follow-up concordance queries. The acquired concordance files are imported in the WebAnno platform (Eckart de Castilho et al., 2016) which we have specifically configured for the FrameNet annotation. When the annotation is done, the concordances are exported from WebAnno and are eventually merged into respective paragraphs.

The UD-based approach has a significant consequence: frame elements (FE) are not annotated as spans of text – only the head word of a UD subtree is annotated. This not only makes the annotation process more simple and the annotations more consistent, but it also facilitates the learning of automatic semantic role labeling, since it is easier to identify the syntactic head of a FE than a span of a string.

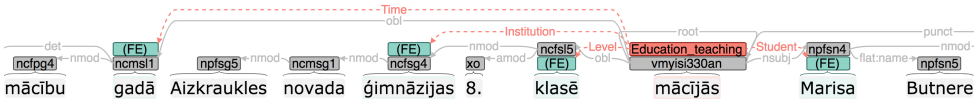


Figure 2: FrameNet annotation on top of a UD tree. The FE spans can be acquired automatically by traversing the respective subtrees: [.. *school year*]$_{\text{Time}}$ [*Aizkraukle county gymnasium*]$_{\text{Institution}}$ [*8th grade*]$_{\text{Level}}$ ***studied***$_{\text{EDUCATION\_TEACHING}}$ [*Marisa Butnere* ..]$_{\text{Student}}$.

To derive the PropBank layer, we are building on the previous work on SemLink (Palmer, 2009) and Predicate Matrix (Lopez de Lacalle et al., 2016). We use the mapping between English FrameNet and English PropBank as a draft configuration. The linguistically intensive manual task is to map the LUs from Latvian FrameNet to the semantic frames of English PropBank.

5 Toolchain

In parallel to the creation of the corpus, we have improved existing NLU tools and developed new NLU tools for Latvian. This has helped us to automate some of the most time consuming steps in the corpus annotation workflow, where post-editing is faster compared to annotating from scratch (incl. morphological tagging and dependency parsing). This also allows to gradually evaluate the impact of the corpus, and to notice potential issues sooner.

Task-specific NLU tools are often used together with other tools in a processing pipeline. This becomes a technical issue if tools are based on different frameworks and library versions. While integration of all tools in a monolithic framework definitely improves usability, it also imposes limitations on technology that can be used, and it adds extra overheads to integrate new components or replace existing ones.

To solve this problem, we have created NLP-PIPE: a simple, modular and scalable NLP/NLU toolchain for Latvian (Znotins and Cīrule, 2018). Modularity and containerization makes it easy to setup, and it offers a lot of flexibility to select and change pipeline components. Components of NLP-PIPE are distributed through Docker Hub.[2] NLP-PIPE supports both synchronous and asynchronous (batch) processing. A web-based API allows for easy integration of different NLP technologies.

[2] `https://hub.docker.com`

The closest existing solutions to NLP-PIPE are Taenga (Ziad et al., 2018) (source code not available), and OpeNER (Agerri et al., 2014) (difficult integration of new components; uses a complex XML-based data format; scalability is achieved through paid Amazon Web Services).

NLP-PIPE currently provides the following annotation services for Latvian: tokenization, morphological tagging, dependency parsing, NER, and coreference resolution. We hope that that this platform will facilitate further development of publicly available NLP components for Latvian, and will make Latvian NLP tools more accessible to a broader audience and for cross-lingual applications.

Segmentation Tokenization and sentence segmentation uses a fast and simple deterministic finite-state automaton which is a part of the Latvian morphological tagger (Paikens et al., 2013) implemented in Java.

Morphological Tagging A statistical morphological tagger achieving 97.9% accuracy for POS recognition and 93.6% for full morphological analysis (Paikens et al., 2013), implemented in Java.

Dependency Parsing A continuous transition-based dependency parser based on LSTM utilizing pre-trained word embeddings, learned character and morphological tag embeddings as features. It achieves 76.84% LAS (Labelled Attachment Score), 81.24% UAS (Unlabeled Attachment Score) on the Latvian UD test set. Derived from (Ballesteros et al., 2015) and (Znotins, 2016), implemented in Python and C++ using the DyNet library.

Named Entity Recognition The NER tagger is based on bidirectional LSTM neural network with an additional CRF layer. It utilizes pre-trained word embeddings, learned character and word shape embeddings as features. The model is currently trained using a flat annotation schema, achieving 74.01% F1-score. The tagger is derived from (Znotins, 2016) and (Lample et al., 2016), implemented in Python using the Keras framework.

Coreference Resolution A rule-based system achieving 58% F1-score (Znotins, 2014), implemented in Java.

6 Conclusion

The consecutive treebank and framebank annotation workflow has turned out to be very productive and mutually beneficial. The dependency layer facilitates the annotation of semantic frames, while the frame semantic analysis often unveils various inconsistencies in the dependency and morphological layers.

The corpus creation has significantly benefited from the CLARIN-supported customizable WebAnno platform. We plan to make the resulting data sets available for cross-linguistic research via CLARIN facilities. In fact, the Latvian UD treebank is already available through LINDAT for searching and browsing using the PML-TQ tool.[3] We also plan to integrate the NLU toolchain into the CLARIN infrastructure.

The multilayer corpus is gradually released[4] under the CC BY-NC-SA 4.0 license for non-commercial use, and under a commercial licence otherwise. An online demo of NLP-PIPE is available.[5] It is distributed under the GNU GPL 3.0 license, and under a commercial licence otherwise.

Acknowledgements

This work has received financial support from the European Regional Development Fund under the grant agreement No. 1.1.1.1/16/A/219. The integration of the toolchain in the CLARIN infrastructure is being supported by CLARIN Latvia.

References

[Agerri et al.2014] Rodrigo Agerri, Josu Bermudez, and German Rigau. 2014. IXA pipeline: Efficient and Ready to Use Multilingual NLP tools. In *Proceedings of the 9th LREC Conference*.

[3] https://lindat.mff.cuni.cz/services/pmltq/

[4] https://github.com/LUMII-AILab/FullStack

[5] http://nlp.ailab.lv

[Ballesteros et al.2015] Miguel Ballesteros, Chris Dyer, and Noah A. Smith. 2015. Improved Transition-based Parsing by Modeling Characters instead of Words with LSTMs. In *Proceedings of EMNLP*.

[Banarescu et al.2013] Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for Sembanking. In *Proceedings of the 7th LAW and Interoperability with Discourse*.

[Barzdins et al.2007] Guntis Barzdins, Normunds Gruzitis, Gunta Nespore, and Baiba Saulite. 2007. Dependency-based hybrid model of syntactic analysis for the languages with a rather free word order. In *Proceedings of the 16th NODALIDA Conference*.

[Chinchor1998] Nancy Chinchor. 1998. Appendix E: MUC-7 Named Entity Task Definition. In *Proceedings of th 7th MUC Conference*.

[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 LT4DH Workshop*.

[Fillmore et al.2003] Charles J. Fillmore, Christopher R. Johnson, and Miriam R.L. Petruck. 2003. Background to FrameNet. *International Journal of Lexicography*, 16(3):235–250.

[Gruzitis et al.2018a] Normunds Gruzitis, Gunta Nespore-Berzkalne, and Baiba Saulite. 2018a. Creation of Latvian FrameNet based on Universal Dependencies. In *Proceedings of the International FrameNet Workshop*.

[Gruzitis et al.2018b] Normunds Gruzitis, Lauma Pretkalnina, Baiba Saulite, Laura Rituma, Gunta Nespore-Berzkalne, Arturs Znotins, and Peteris Paikens. 2018b. Creation of a Balanced State-of-the-Art Multilayer Corpus for NLU. In *Proceedings of the 11th LREC Conference*.

[Lample et al.2016] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural Architectures for Named Entity Recognition. In *Proceedings of HLT-NAACL*.

[Lopez de Lacalle et al.2016] Maddalen Lopez de Lacalle, Egoitz Laparra, Itziar Aldabe, and German Rigau. 2016. A Multilingual Predicate Matrix. In *Proceedings of the 10th LREC Conference*.

[Nespore-Berzkalne et al.2018] Gunta Nespore-Berzkalne, Baiba Saulite, and Normunds Gruzitis. 2018. Latvian FrameNet: Cross-Lingual Issues. In *Human Language Technologies – The Baltic Perspective*, FAIA. IOS Press.

[Nivre et al.2016] Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of LREC*.

[Paikens et al.2013] P. Paikens, L. Rituma, and L. Pretkalnina. 2013. Morphological analysis with limited resources: Latvian example. In *Proceedings of the 19th NODALIDA Conference*.

[Palmer et al.2005] Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–106.

[Palmer2009] Martha Palmer. 2009. SemLink: Linking PropBank, VerbNet and FrameNet. In *Proceedings of the Generative Lexicon Conference*.

[Pretkalnina et al.2011] Lauma Pretkalnina, Gunta Nespore, Kristine Levane-Petrova, and Baiba Saulite. 2011. A Prague Markup Language profile for the SemTi-Kamols grammar model. In *Proceedings of NODALIDA*.

[Pretkalnina et al.2018] Lauma Pretkalnina, Laura Rituma, and Baiba Saulite. 2018. Deriving Enhanced Universal Dependencies from a Hybrid Dependency-Constituency Treebank. In *TSD*, volume 11107 of *LNCS*. Springer.

[Ziad et al.2018] Housam Ziad, John Philip McCrae, and Paul Buitelaar. 2018. Teanga: A Linked Data based platform for Natural Language Processing. In *Proceedings of the 11th LREC Conference*.

[Znotins and Cīrule2018] A. Znotins and E. Cīrule. 2018. NLP-PIPE: Latvian NLP Tool Pipeline. In *Human Language Technologies – The Baltic Perspective*. IOS Press.

[Znotins2014] A. Znotins. 2014. Coreference resolution in Latvian. In *Human Language Technologies – The Baltic Perspective*, volume 268. IOS Press.

[Znotins2016] A. Znotins. 2016. Word embeddings for Latvian natural language processing tools. In *Human Language Technologies – The Baltic Perspective*, volume 289. IOS Press.

(Re-)Constructing "public debates" with CLARIAH MediaSuite tools in print and audiovisual media

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 Institue
Utrecht University, The Netherlands
t.pieters@uu.nl

Abstract

This paper focuses on the proceedings of CLARIAH research pilot Debate Research Across Media (DReAM) by reflecting on the used conceptualization of public debates. In the pilot, heterogenous datasets (of digitized print and audiovisual media) are searched with the levelled research approach (combining distant and close reading techniques) to do historical public debate analysis with tools of the CLARIAH MediaSuite. The qualitative research interest in public debates is fundamentally 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 reflected upon. The natures of the different media and digitization processes, 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 taken into consideration before we argue that using this approach in the MediaSuite can help the researcher to gain an improved understanding of historical public debates.

1 Introduction

In research pilot Debate Research Across Media (DReAM)[1] we tested and contributed to the development of the Comparative Search tool and related tools in the CLARIAH media research environment MediaSuite[2][3]. We worked to accommodate the levelled research approach (Van der Molen and Pieters, 2017) in the MediaSuite. This explorative historical research approach works with the assumption that a combination of distant reading techniques (keyword search, word cloud analysis and timeline graph analysis) with historical analysis of a thematic subselection can help us to trace and understand public debates in historical material. By combining a number of relevant tools in the MediaSuite, 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 upon media studies principles, and one of the main aims of DReAM was to make the MediaSuite equipped for historical public debate research too. In this paper we critically reflect on a conceptualization of *public debate* implied in the research infrastructure and on how the levelled approach can be used to understand such public debates in the available heterogeneous digital sources.

[1] < www.clariah.nl/projecten/research-pilots/dream/dream>.
[2] The MediaSuite 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.
[3] CLARIAH.nl is the Dutch infrastructure related to CLARIN.eu and DARIAH.eu.

2 Public debate research in the MediaSuite

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 methods and tools in the MediaSuite. The most important of these tools, Comparative Search, is based on a previous CLARIAH cross-media analysis tool called AVResearcherXL (Bron et al, 2013; Huurnink et al, 2013). AVResearcherXL simultaneously searches the previously mentioned KB newspaper and NISV radio and television archive.

The development process was iterative: as end users, we first outlined our ideas and needs in a so called Demonstration Scenario, developers then worked on this, we then finally 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 levelled 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 sign-posting strategy to find relevant material. We safeguarded this historical approach with a number of important insights from media studies in the MediaSuite.

Public debate research in the MediaSuite can be done by using a combination of its digital tools: the infrastructure can in that sense generate thematically and historically connected subsets that refer to "public debates". First, each datasets is loaded in the tool Collection Inspector for metadata quality assessment. This allows the researcher to understand the composition of the different datasets. The researcher is then able to send a selected data set based on specified complete metadata to the next tool: Comparative Search. Historical interpretation of the data is only possible with a sufficiently complete date field for both datasets and with sufficient

a. Optical Character Recognition metadata for the newspaper dataset
b. Automatic Speech Recognition metadata for the radio and television datasets[4]

In Comparative Search these heterogeneous datasets that have been selected based on the metadata analysis can be queried next to each other by means of keyword search (macro level). The researcher can choose whether one query or different (medium specific) queries are used. The search results can be further explored by means of timeline graphs (macro level) and word clouds (meso level)[5]. Furthermore, individual results are listed and can be sorted in several ways. The researcher has an option to save queries and bookmark results to their Workspace to allow for structural analysis of the results (micro level).

When all of these functionalities are combined in a savvy manner, they allow for analysis of a cross-media dataset ("public debate") that is thematically and chronologically linked. Below we will outline how this is grounded in historical research and media studies.

3 (Re-)Constructing 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?"

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[6]. The careful contextualization of events that does justice to the actors

[4] At the time of writing (September 2018) the data integration process for the NISV data is ongoing but not completed yet.
[5] At the time of writing (September 2018) the word cloud functionality has not been integrated yet. In Summer 2018 this has been accommodated in a Jupyter notebook.
[6] <www.nwo.nl/onderzoek-en-resultaten/onderzoeksprojecten/i/46/13546.html>.

involved is the historian's primary concern. 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). Our claim is that drawing on insights from media studies can help historians to critically bridge the gap between distant and close reading of digital media sources to reconstruct public debates.

Doing 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 and in this case television, radio and print media function entirely different and, according to Marshal McLuhan (1964), even *are* the message (as opposed to what we would traditionally understand as the content): what these media convey is defined in the first place by each medium. In that sense, in order to describe a historical public debate, it is necessary to understand how different media can contribute to a meaningful public debate. 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 MediaSuite. How are these related to each other meaningfully as cross-media public debates? The textual data is searchable by means of the OCR data; the audiovisual data is searchable by ASR data. On this scale, this is unexplored methodological territory, and it forces reflection on how we can still do justice to the *visual* meanings of the television data.

Secondly, just as there is a historical difference between the source date of the objects of research and the time they are researched, there are different meaningful focus points in media studies. 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)? There are many ways to understand and account for the different levels of meaning on this continuum, for instance the encoding/decoding model (Hall, 1980). 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 MediaSuite 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 read in the digitally combined source material itself, which precludes a focus on production and reception.

The last related point of reflection is that this more or less artificial nature of public debate requires an explicit theoretical approach. Whereas, put crudely, historical research excels at contextualizing historical events, media studies is more targeted towards picking up on particular discursive relations: media objects are approached from a specific angle (i.e. gender, politics etc.). Since the MediaSuite does not isolate, group nor contextualize historical events, knowledge of the historical contextualization of the sources along with a sensibility towards discursive relations are necessary to signalize meaningful historical public debates within the search results. This brings us back to the importance of researcher expertise that is crucial in the levelled approach: close reading of the source material is required to trace and understand the possible connections 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 informed partly by media studies insights. By recognizing how a particular topic, for instance the psychoactive drug MDMA (searched for with query `xtc OR mdma OR ecsta*y`), 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.

4 Conclusion

In this paper we described a conceptual approach to "public debates". 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 MediaSuite infrastructure, grounded in historical research and safeguarded with reflections gained from media studies. By approaching the relevant datasets with the levelled approach in the MediaSuite it is possible to become aware of shifts in the discursive formation of particular topics. Although this is a fundamentally constructive exercise, our reliance on historical contextual expertise makes it possible to use this awareness to improve our understanding of historical relations and discursive dynamics of public debates across media. 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 interaction between regulation and public debates based on fine-grained reading of the digitized source material.

References

Bron, M., Gorp, J. van, Nack, F., Rijke, M. de, Vishneuski, Andrei & Leeuw, J.S. de (2012). A Subjunctive Exploratory Search Interface to Support Media Studies Researchers. SIGIR '12: 35th international ACM SIGIR conference on Research and development in information retrieval Portland, Oregon: ACM.

Derrida, Jacques. *Archive fever. A Freudian impression*. Chicago: University of Chicago Press, 1996.

Hall, Stuart. (1980). "Encoding/decoding." In: Stuart Hall, Dorothy Hobson, Andrew Love and Paul Willis (eds.) *Culture, Media Language*. London: Hutchinson.

Huurnink, B., Bronner, A., Bron, M., Gorp, J. van, Goede, B. de & Wees, J. van (2013). AVResearcher: Exploring Audiovisual Metadata. DIR 2013: Dutch-Belgian Information Retrieval Conference Delft: DIR.

McLuhan, Marshal. (1964). *Understanding Media*. London: Routledge.

Nicholson, Bob. (2013). "The digital turn. Exploring the methodological possibilities of digital newspaper archives" *Media History* 19.1

Van der Molen, Berrie, 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." 2017 arXiv:1701.00487.

Van der Molen, Berrie, 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, Gerben. (2013). "On Digital History." *BMGN - Low Countries Historical Review*, 128.4, 3–29.

Improving Access to Time-Based Media through Crowdsourcing and CL Tools: WGBH Educational Foundation and the American Archive of Public Broadcasting

Karen Cariani **WGBH Media Library and Archives** WGBH Educational Foundation, Boston, MA, USA Karen_Cariani@wgbh.org	**Casey Davis-Kaufman** **WGBH Media Library and Archives** WGBH Educational Foundation, Boston, MA, USA Casey_Davis-Kaufman@wgbh.org

Abstract

In this paper, we describe the challenges facing many libraries and archives trying to provide better access to their media collections through online discoverability. We present the initial results of a project that combines technological and social approaches for metadata creation by leveraging scalable computation and engaging the public, the end users, to improve access through crowdsourcing games and tools for time-based media. The larger need is for more accurate output and ease of use of computational tools for audiovisual archives to create descriptive metadata and annotations. As leaders in preservation, access, and analysis of culturally significant audiovisual material, WGBH is continually confronted with the need to enhance the descriptive data to improve discoverability for large-scale digital indexing and analysis of media collections.

1. The problem of scale for audiovisual collections

The creation of new audiovisual materials is increasing daily. In 2014, AVPreserve and the Northeast Document Conservation Center reported that over 537 million sound recordings already exist in collecting institutions across the US, and the number of video items is likely even higher. Over 57% of these items are unique. Audiovisual materials, often stored on obsolete analog tape-based formats, are deteriorating quickly and need to be digitized to be preserved and accessed. In addition, archives are increasingly collecting more "born digital" audiovisual content. Once digital, audiovisual recordings need to be adequately described, annotated, or "cataloged," in order to be made accessible and discoverable by researchers.

Creating metadata requires significant amounts of human labor, particularly for audiovisual materials. Compared to text-based and photographic archival collections, moving images can contain thousands of individual images, and audio can be more time-consuming because there are no visual cues and attentive listening is required. Standard audiovisual cataloging practice requires the item to be fully viewed or listened to in real time and descriptive information written on an item by item basis. Generally speaking, full descriptive cataloging takes a 1:1 ratio of time for audiovisual content – an hour video takes an hour to fully catalog. It is labor intensive for this content to be made accessible by standard cataloging practices. Yet with online search engines dependent on text exposed to the web, cataloging is critical for access and discoverability.

WGBH is the premier public radio and TV station in the US. It produces roughly 1/3 of the TV programs that are distributed across the US. It has been on the air since 1951 and has an archive of roughly 400,000 media items consisting of analog formats from 2" videotape to ¼" audio. WGBH has faced this challenge directly in connection with its work on the American Archive of Public Broadcasting (AAPB). The AAPB, a collaboration between WGBH and the Library of Congress (the Library), is coordinating a national effort to identify, preserve, and make accessible as much as possible a digital archive of public television and radio dating back to the late 1940s.

The collection has over 50,000 hours, or 90,000 files, of digital media selected by more than 100 public media stations and organizations with little consistent descriptive data. Often there is very little descriptive metadata beyond title and contributing station. Less than 50% of the records contain a date. The resources necessary to catalog large collections like the AAPB by standard cataloging practices are overwhelming and unattainable. We have calculated that to 'lightly' catalog the AAPB collection of 90,000 digital items, only spending 15-20 minutes per item, it would take one person 13 years. To fully catalog the collection, complete with subject headings, and name authorities, it would take one person approximately 43 years working full time.

The AAPB faces a massive growing collection of digital media files with incomplete metadata. We realize we will always have a significant backlog as we continue to grow the collection by up to 25,000 hours annually. At the same time, the AAPB needs metadata to know what we have, in order to determine its level of access, and to make the content discoverable by users. Our questions are, "How can we transform large amounts of audiovisual content into a searchable dataset for search engines and indexers? Is the public is a viable resource that could help with cataloging?"

Audiovisual media must be digital or digitized to be broadly accessible. However, just because an audiovisual collection is digital does not guarantee that it will be discoverable to scholars or the general public. A 2010 survey reported that 84% of searching on the web begins with search engines. No matter how many hours of digital content are available on the web, without robust descriptive metadata, it is not discoverable. To improve discoverability and access for scholars and researchers, the content needs better descriptive information that can be indexed by search engines. In addition, the rate at which born digital media is created increases every day, if cultural heritage institutions are to keep up with the pace of audiovisual content creation, they need practices that can radically scale to meet the pace of creation.

2. Potential solutions using audio analysis and computational linguistics tools

Creating transcripts of the audio is a potential solution to describe the content and expose the text to search engines. Using CL tools such as speech, or audio, to text tools can be adapted or "trained" for use with specific materials to achieve greater degrees of accuracy than ever before, but they do not generate perfect transcripts, and the data is not always clean enough for consistent indexing. There is no easy method to correct inaccuracies for large quantities of media transcripts. Similarly, audio analysis tools have progressed far enough (and have been implemented for collections ranging from bird calls to poetry) that they could be trained and applied at scale to oral history and archival audiovisual collections containing speech. However, training such tools and refining their output requires human oversight and quality control and expertise beyond most archivists.

Harnessing the interest of the public and spreading the work over many volunteers may be a solution to the lack of resources available to describe this content. Crowdsourcing spreads the amount of work across many people and gives them an opportunity to engage with the archive, beyond simply viewing or listening to it. It allows them to gain a sense of ownership as well learn something new. Public broadcasting fans, scholars, lifelong learners, students, and others get a chance to access historic programming while contributing back to the archive. Their contributions enable long-term access. Through this project we have found that crowdsourcing corrections, although it gets our users involved, is not practical with the current volume of transcripts. Corrections are not happening quickly enough.

Alternatively, collaborating with the computational tool-making experts – linguists, computer scientists – to increase accuracy of the tools, allow easier use, and the ability to enhance outputs,

would greatly benefit both communities. The challenge is to make the tools for capturing the data easy, intuitive and engaging – tools that anyone could use or implement, and use the tools at scale with large volumes of files. In addition, the data output needs to conform to archivists needs and metadata schemas.

3. Computation + crowdsourcing to describe audiovisual content

Prior to Pop Up Archive's acquisition by Apple, WGBH and Pop Up Archive received an Institute of Museum and Library Services (IMLS) Research Grant to test such methods through the AAPB. Through an iterative, design-based research approach we are exploring the following research questions:

- How can crowdsourced improvements to machine-generated transcripts and tags increase the quality of descriptive metadata and enhance search engine discoverability for audiovisual content? How can a range of web-based games create new points of access and engage the public engagement with time-based media through crowd source tools? What qualitative attributes of audiovisual public media content (such as speaker identities, emotion, and tone) can be successfully identified with spectral analysis tools, and how can feeding crowdsourced improvements back into audio analysis tools improve their future output and create publicly available training data that can help with cataloging other audiovisual collections at scale?

Using content from the AAPB as the sample data set to answer our questions, the project used open source speech-to-text and audio analysis tools (Kaldi) to create transcripts and qualitative waveform analysis for approximately 40,000 hours of AAPB digital files. We developed time-based media crowdsourcing games, open source web-based games, to improve transcripts and descriptive data by engaging the public in a crowd-sourced, participatory cataloging project.

4. Building on current computational techniques

While computational approaches for cataloging and analyzing text and image-based collections are increasingly common, audiovisual collections have received less attention to date. WGBH and Pop Up Archive tried to improve on the work of COMMA, a BBC R&D project, and High Performance Sound Technologies for Access and Scholarship (HiPSTAS), both of which have been developing new technologies for facilitating automated metadata extraction from time-based media collections. This emerging technology increasingly enables metadata to be created automatically for digital time-based media through programmatic analysis and speech-to-text software. Specifically, we are interested in building upon the BBC's and HiPSTAS's work to create extensive training data for speaker identification and other qualitative audio attributes, and to provide a baseline workflow leveraging multiple audio analysis tools, crowdsourcing, and training data that can be used by other audiovisual collections. Although we have a great data set for training the tools, archives need these tools to generate meaningful metadata to improve access to the collections, and archivists need toolkits and understandable documentation to be able to incorporate the improved tools into their processing and cataloguing workflows.

Automatically generated metadata from speech can provide a map of content and topics addressed in US public broadcast material from the past 70 years. We seek to augment approximately 50,000 hours of U.S. public broadcast content with automatically generated metadata improved through crowd-sourced contributions. This will serve as a national "jumping off point" for digital discovery and analyses of US public media and archival spoken word audio. Tools and methodologies will be shared with the academic and open source communities with the explicit goal of creating reusable workflows for similar endeavors.

5. Future Work

There is a great need to develop open workflows that employ emerging open source technology to create large quantities of metadata for digitized audiovisual content through highly accurate speech-to-text software, semantic tagging, and digital waveform data. This can serve as a great opportunity for both computational linguistics and archivists communities to work with each other to create enriched metadata to enhance discoverability as well as to create data sets that further attracts digital humanity research and social science research.

The AAPB collection includes local content from over 100 stations in various regions across the country. The audio in the collection represents spoken language in different accents of the English language and some non-English in various audio quality, often overlapping as well as non-linguistic sounds such as music, animal and object sounds, noise, and other sounds not translatable to text. All these factors do not lend it to a great speech-to-text testbed, or rather, it is a great challenge for computational linguistics tools. The programs vary from nightly news, to in-studio magazine shows, to cooking and fishing programs, to musical performances, to interviews and panel discussions. With a data set so diverse, it is a great test set to improve the tools.

We are currently working with James Pustejovsky at Brandeis University to utilize CL tools to improve the metadata. This effort includes work to develop tools to: time stamp videos; identify named entities; and episode videos in order to separate thematic content. We hope to also develop a "talking heads" classifier to further identify the content of the program and the type of program such as talking heads, an individual interview, or an event or scene.

One problem is that output of many of these computational linguistic tools are not easily ingested into archival metadata systems, and therefore less useful for archives. For WGBH, ingest data is best expressed as hierarchical xml or csv. WGBH utilizes PBCore (http://pbcore.org) as a metadata schema. PBCore was developed specifically as a derivative of Dublin Core to better describe audiovisual materials. (EBUCore, another PBCore family schema developed is utilized by the European Broadcasting Union.) An interchange data format is under development that easily maps output of the CL tools to archival ingest fields as well as enables interoperability between different tools. Our goal is to develop a flexible but syntactically and semantically consistent data format so that it can further be mapped to other metadata format for archives and language resources such as Component MetaData (CMDI) of Clarin project.

WGBH often has complete clean printed transcripts of interviews used in the editing process of a broadcast program. However they are typically not synced to the digital video file. Since the interviews are cut into the final program and themselves are not the final broadcast content, they are not captioned. Forced alignment of transcription to the digital video file with time stamps would be enormously helpful. Much of the collection was digitized from full broadcast tapes complete with "bars and tone" at the beginning along with a slate countdown before air. Being able to jump beyond the bars and tones directly to the beginning of the program will enhance the user experience. And finally, recognizing non-speech in an audio or video file so that speech to text tools don't try to translate non speech into words, would also help clean up transcripts. WGBH is looking forward to sharing our work with larger CL community in hopes of building more collaborations.

References

BBC. 2018. BBC world service archive. http://worldservice.prototyping.bbc.co.uk/. Accessed May 25, 2018.

British Library. 2018. British library georeferencer. http://www.bl.uk/maps/. Accessed May 25, 2018.

Tanya E. Clement, David Tcheng, Loretta Auvil, and Tony Borries. 2015. High performance sound technologies for access and scholarship (hipstas) in the digital humanities. Proceedings of the American Society for Information Science and Technology , 51(1):1–10.

COMMA. 2015. Comma: A cloud platform for metadata extraction. https://www.bbc.co.uk/rd/projects/comma. Accessed May 25, 2018.

Component Metadata. (n.d.). CLARIN. https://www.clarin.eu/content/component-metadata. Accessed September 11, 2018.

Family Tree. 2017. Family search. https://www.familytree.com/. Accessed May 25, 2018.

Nancy Ide, James Pustejovsky, Christopher Cieri, Eric Nyberg, Denise Dipersio, Chunqi Shi, Keith Suderman, Marc Verhagen, Di Wang, and Jonathan Wright. 2016. The language application grid. In Revised Selected Papers of the Second International Workshop on Worldwide Language Service Infrastructure - Volume 9442 , WLSI 2015, pages 51–70, New York, NY, USA. Springer-Verlag New York, Inc.

Peter B. Kaufman. January 2018. Towards a new audiovisual think tank for audiovisual archivists and cultural heritage professionals. Netherlands Institute for Sound and Vision. Hilversum, NL. http://dx.doi.org/10.18146/2018thinktank01. Accessed September 11, 2018.

Netherlands Institute for Sound and Vision. (n.d.). In the spotlight: Waisda? https://www.beeldengeluid.nl/en/knowledge/blog/spotlight-waisda. Accessed April 20, 2018.

New York Public Library. (n.d.). Together We Listen. http://togetherwelisten.nypl.org/. Accessed April 20, 2018.

Jack Nicas. February 27, 2017. YouTube tops 1 billion hours of video a day, on pace to eclipse tv. https://www.wsj.com/articles/youtube-tops-1-billion-hours-of-video-a-day-on-pace-to-eclipse-tv-1488220851. Accessed February 20, 2018.

OCLC. 2010. Perceptions of libraries. http://www.oclc.org/content/dam/oclc/reports/2010perceptions/2010perceptions_all.pdf, p. 32.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. 2011. The kaldi speech recognition toolkit. In IEEE 2011Workshop on Automatic Speech Recognition and Understanding. IEEE Signal Processing Society, December. IEEE Catalog No.: CFP11SRW-USB.

James Pustejovsky, Marc Verhagen, Keongmin Rim, Yu Ma, Liang Ran, Samitha Liyanage, Jaimie Murdock, Robert H McDonald, and Beth Plale. 2017. Enhancing access to digital media: The language application grid in the htrc data capsule. In Proceedings of the Practice and Experience in Advanced Research Computing 2017 on Sustainability, Success and Impact, page 60. ACM.

Smithsonian Institution. (n.d.). Smithsonian digital volunteers. https://transcription.si.edu/. Accessed April 20, 2018.

Tiltfactor Laboratory Dartmouth College. (n.d.). Metadata games. http://www.metadatagames.org/. Accessed April 20, 2018.

Trove. 2017. Trove: National Library of Australia. http://trove.nla.gov.au/. Accessed May 25, 2018.

Discovering software resources in CLARIN

Jan Odijk
UiL-OTS
Utrecht University, the Netherlands
j.odijk@uu.nl

Abstract

We 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. We have tested the profile by making metadata for over 70 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.

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. In order to address this issue we present (1) a CMDI profile for the description of software that enables discovery of the software and formal documentation of aspects of the software, and (2) a proposal for faceted search in metadata for software. We have tested the profile by making metadata for over 70 pieces of software. We describe how we ensured the quality of these metadata descriptions. We describe how we are testing the proposed faceted search. We propose to add this faceted search to the VLO, and 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 such faceted search.

2 Metadata Profile CLARINSoftwareDescription

The ClarinSoftwareDescription (CSD) profile[1] enables one to describe information about software in accordance with the CMDI metadata framework used in CLARIN (Broeder et al., 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.

We briefly describe the major components and elements of the profile. More details will be included in the actual presentation and the full paper. 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* is an extension of the component *cmdi-generalinfo*[2] with elements for specifying the *CLARIN Centre* hosting the resource and the *national project(s)* in which it has been made part of the CLARIN infrastructure.

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),

[1] clarin.eu:cr1:p_1342181139640.

[2] clarin.eu:cr1:c_1342181139620.

research domains and, for the linguistics domain, relevant *linguistic subdisciplines* for which it was originally developed.[3]

The *SoftwareImplementation* component enables one to describe information for users on the implementation and installation of the software. The most important components are for the description of the *interface*, the *input* and the *output* of the software.

The *Access* component enables one to describe information about the availability and accessibility of the resource.

The *ResourceDocumentation* component enables one to describe the documentation of the resource. The *SoftwareDevelopment* component is intended for information on the history and development of the software. The *TechnicalInfo* component enables one to describe technical information on a resource and is mainly aimed at developers.

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.[4]

The *LRS* component is intended for the description of the properties of a particular task for the CLARIN Language Resource SwitchBoard (CLRS, (Zinn, 2016)). Multiple LRS components can be present. It is our viewpoint that specifications for an application for inclusion in the CLRS registry[5] 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 we added a component to offer facilities for supplying the missing information. We 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).[6]

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)).[7] For the ones that were lacking we created definitions and provided other relevant information required for inclusion into the CCR. We submitted this file (2017-09-08), in the format required, to the maintainers of the CCR. However, the CCR coordinators'[8] mill runs slowly, and so far none of them have been incorporated in the CCR. After our submission to the CCR, we made some new modifications to the profile, so there are new elements and values for which the semantics does not exist yet.

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 are *ToolProfile*[9] (49 resources), *WeblichtWebService*[10] (287 resources), *resourceInfo*[11] with the value *toolService* for the element *resourceType* (68 resources), and *OLAC-DcmiTerms*[12] (189 software resources)..

In the presentation and in the full paper we will make a detailed comparison between these profiles. Here we mention the most important differences: (1) the CSD profile is fully dedicated to the description of software (v. the *OLAC-DcmiTerms* profile); (2) the CSD profile can be used to describe any type of software (v. *WebLichtWebservice*); (3) it offers more elements, and more formalised elements than the other profiles, not only elements useful for discovery but also for (formalised) documentation.

[3] which, of course, does not preclude its use in other research domains that were not foreseen during development.
[4] This component was created by Menzo Windhouwer, and adapted to the requirements of CMDI version 1.2.
[5] `https://github.com/clarin-eric/LRSwitchboard/blob/master/app/back-end/Registry.js`
[6] https://languagemachines.github.io/frog/
[7] `https://concepts.clarin.eu/ccr/browser/`.
[8] https://www.clarin.eu/content/concept-registry-coordinators
[9] `clarin.eu:cr1:p_1290431694581`.
[10] `clarin.eu:cr1:p_132065762964428`.
[11] `clarin.eu:cr1:p_1360931019836`.
[12] `clarin.eu:cr1:p_1288172614026`.

3 Metadata Descriptions using the CSD profile

We have described more than 70 software resources with the CSD profile, and describing these software resources resulted in various improvements of earlier versions of the profile.These software resources mainly concern resources from the Netherlands. Most descriptions started from the information contained in the CLARIN-NL Portal, Services part.[13] 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.

4 Metadata Quality

All metadata descriptions have been validated against the profile definition. We created several schematron[14] 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*. We 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.[15]

The quality checks offered by the CLARIN Curation Module[16] (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.

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, we 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, input modality, licence

Display Facets name, title, version, inputMimetype, outputMimetype, outputLanguage, Country, Description, ResourceProxy, AccessContact, ProjectContact, CreatorContact, Documentation, Publication, sourcecodeURI, Project, MDSelfLink, OriginalLocation

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*).

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 a little bit of metadata curation is needed in some cases and existing values must be mapped to a restricted vocabulary for use in the faceted search. This is the topic of the next section.

[13] `http://portal.clarin.nl/clarin-resource-list-fs`.

[14] `http://schematron.com/`

[15] On 2018-03-26, 750 URLs were correctly found, 22 were not found, and 73 exceptions were raised. Most exceptions raised are due to on-going changes on the INT website.

[16] https://clarin.oeaw.ac.at/curate/

6 Curation of existing metadata for software

The basic idea is as follows: we create a new standardised metadata record 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 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, e.g. to restrict them to a specific closed vocabulary. We report on our experiments with such a curation file for the *WebLichtWebService* profile, since it 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*, *Publication*, *modality* and *license*. We 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*, *Country*, and *modality*.

We still have to make curation files for the *ToolProfile*, *resourceInfo* and the *OLACDcmiTerms* profiles. We 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.

7 Concluding Remarks

In the full paper we will summarise our conclusions. We will also describe some problems we encountered, which we only briefly mention here: (1) definition of closed vocabularies (2) no possibility to reuse metadata elements ; (3) a lot of variety in the contents of the CMDI envelope element *MdSelfLink*, resulting in several unresolved or syntactically incorrect results; (4) lack of good CMDI metadata editors. Finally, we will identify some future work, in particular on deriving CLRS registry entries for CLAM-based applications and web services.[17]

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.

References

[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).

[Ostojic et al.2017] Davor Ostojic, Go Sugimoto, and Matej uro. 2017. The curation module and statistical analysis on VLO metadata quality. In *Selected papers from the CLARIN Annual Conference 2016, Aix-en-Provence, 2628 October 2016*, number 136 in Linkping Electronic Conference Proceedings, pages 90–101. Linkping University Electronic Press, Linkpings Universitet.

[17] `https://proycon.github.io/clam/`.

[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 1416, 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.

[Zinn2016] Claus Zinn. 2016. 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.

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 a 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. The case study is offered by the discovering of Anna Maria Bruzzone archive, containing the voices of people with mental disabilities recorded in 1977 in a psychiatric hospital.

1 Introduction

The 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. The case study is offered by the discovering of Anna Maria Bruzzone archive[1]. The Bruzzone's interviews were recorded before the Italian Data Protection Code (IDPC) was issued, 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 the article, the analysis is limited to personal data protection.

In the paper, some legal issues affecting the use and re-use of the archive is presented and discussed. The model envisaged aims to find 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).

The paper is conceived as follows: in § 2 the 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.

[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).

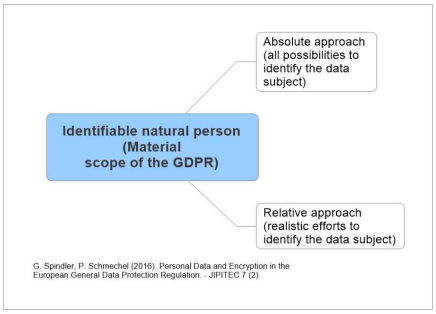
2 The speech archive of Anna Maria Bruzzone

Anna Maria Bruzzone's book *Ci chiamavano matti. Voci da un ospedale psichiatrico* (Einaudi, Torino 1979) contains the testimonies of thirty-seven patients of the Arezzo psychiatric hospital collected in 1977. The book testimonies 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 to 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, which were donated to the Department of Educational Sciences, Human Sciences and Intercultural Communication of the University of Siena – Arezzo. Such discovery is of high magnitude 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 of the UNISI Department, where also the Historical Archive of the Arezzo psychiatric hospital is hosted.
Reading a testimony and listening to it from the voice of the interviewee are apparently not the same thing and Bruzzone herself was well aware of it (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. 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 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 suited for publishing. This opens up the possibility to understand, document and examine the changes undergone by 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 name and not the pseudonym has been found in the box of every single tape.

3 Personal data and special categories of personal data

The curation and dissemination of vulnerable people archives are subject to the personal data regulation.

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 in the diagram on the right (from Spindler, Schmechel 2016). 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 is "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 has serious consequences also 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 shall be processed with special protection measures.

The next question is whether curation and dissemination of the archives is processing of 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 processing of personal data. The potential options for using the 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

[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.

[3] It was already in the 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.

(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 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 shortly 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 that "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. AIn fact, anonymisation may not correspond to the needs and scope of historical research, which may have interest, among others, in the analysis of the relational structures of individuals: in short, historians are interested in "names and faces". Therefore, the re-use of Bruzzone's interviews for historical purposes could be a leading case to set a "legal chain" for personal data processing, which can be summarized as follows.

Firstly, the research group shall provide for the identification of the real names of the patients and for the matching with the pseudonyms, as attested in the volume (Bruzzone 1979). Secondly, the research group should try to go back to the interviewees, also engaging the network of all the people – physicians, nurses, social workers, ordinary citizens – involved in the recent history of the Psychiatric hospital. This reconstruction could help to investigate the possibility and feasibility 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. If the consent form is obtained, the oral archive could be finally enjoyed by the research communities and the entire civil society.

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).

5 Conclusions

The analysed archive is subject to the GDPR since they contain special categories of personal data. The curation and dissemination of the archives is processing of personal data which requires legal grounds and other measures assuring the 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 in case of historical research. An additional option is to

[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).

acquire informed explicit 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).

Reference

[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);

[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 http://collections.internetmemory.org/haeu/20171122154227/http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2007/wp136_en.pdf (20.8.2018).

Versioning with Persistent Identifiers

Martin Matthiesen
CSC – IT Center for Science
Espoo, Finland
martin.matthiesen@csc.fi

Ute Dieckmann
Department of Digital Humanities
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 special landing pages.

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 also important for reproducibility of existing research. Publication principles like FAIR (Wilkinson et al., 2016) emphasise the importance of persistent identifiers (PIDs) and rich 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 rich metadata records means in practice when updating a large dataset.

In the final paper we will describe some of the tools and methods used at FIN-CLARIN's main service for researchers, the Language Bank of Finland[1]: For example, we will present our metadata and PID handling using META-SHARE[2], our system for managing versions[3] and our edition of Språkbanken's Korp concordance tool (Borin et al., 2012). We outline how using our tools and methods to perform the update was more challenging than anticipated. 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

[1] https://www.kielipankki.fi/language-bank/

[2] http://metashare.csc.fi/

[3] See The Language Bank's *Life cycle and metadata model of language resources*: http://urn.fi/urn:nbn:fi:lb-201710212

- 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

2 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. It was available for download licensed as CLARIN ACA +NC[4] (University of Helsinki, 2017a) and in Korp, licensed as CC BY (University of Helsinki, 2017c). The final paper will describe the conversion process of the raw data in more detail.

3 Initial assumptions

Even though the dataset consists of various individually identifiable newspapers and magazines, we had first 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.

4 The initial update process

Initially the dataset was updated as stated in the metadata: frequently and without changing the PID. Information on the updates of each variant was maintained on a separate wiki page, which was available as a link 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.

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 would be implicitly addressed, since new subcorpora would show 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. The updated dataset contains now 369 subcorpora in both variants, respectively. 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, they would be overwritten if more than one zip file is 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.

[4] `http://urn.fi/urn:nbn:fi:lb-2016050602`

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). We had to make decisions, which we will explain in more detail in the final paper:

- Abandon the idea of an accruing dataset variant behind a single PID.
- Begin versioning:
  - Create a version 1 of the current Korp variant of the dataset.
  - Create a version 1 of the current Download variant of the dataset.
  - Make explicit, that the variants of version 1 are not in sync.
- Keep the idea of having one PID per variant and version.
- Introduce "stop-over pages"[5] for PIDs pointing to corrected data in version 1.
- Mark non-significant changes in a new Change Log section in the metadata.

We considered 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).

We had to make at least minor changes to all subcorpora in Korp and Download. Version 2 of the downloadable corpus (University of Helsinki, 2017b) alone is a collection of 369 subcorpora consisting of 574 zip files and 88718 individual files.

Had we assigned 574 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, changing 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 old zip file would not have been provided automatically. PIDs to individual files would have required us to provide the content uncompressed and created a need for even more stop-over pages. Storage and bandwidth considerations also had to be taken into account.

Instead we used one PID for the Download variant and explained the changes in the metadata in a Change Log. We also used a Change Log in the metadata and two stop-over pages to explain the changes we made to the already published subcorpora in Korp. In the final paper we will describe the targets of our PIDs before and after the update in more detail using diagrams and examples from the dataset. We will also explain why automated PID assignment would work only to a limited extent in our scenario.

6 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 describe how we tried to find a balance between usability and transparency at every stage of the update process. In the final paper we will discuss PID granularity, machine readable PIDs and automated vs. manual PID handling in more detail.

[5] A "stop-over page" is a manually curated landing page accessed by a PID that pointed to data that has been corrected. The stop-over page is used to explain changes and direct the user further to the location of the corrected data. Information on how to access the previously available data is provided. In other words, a stop-over page is used in cases where it is not feasible to keep the orginal data online.

References

[Baker2016] Monya Baker. 2016. 1,500 scientists lift the lid on reproducibility. *Nature*. https://doi.org/10.1038/533452a.

[Borin et al.2012] Lars Borin, Markus Forsberg, and Johan Roxendal. 2012. Korp – the corpus infrastructure of Språkbanken. In *Proceedings of LREC 2012. Istanbul: ELRA*, page 474–478.

[Munafò et al.2017] 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.

[Rueda et al.2016] 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 Helsinki2017a] University of Helsinki. 2017a. 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 Helsinki2017b] University of Helsinki. 2017b. Corpus of Finnish Magazines and Newspapers from the 1990s and 2000s, Downloadable Version 2. The Language Bank of Finland. Retrieved from http://urn.fi/urn:nbn:fi:lb-2017091902.

[University of Helsinki2017c] University of Helsinki. 2017c. Corpus of Finnish Magazines and Newspapers from the 1990s and 2000s, Version 1. The Language Bank of Finland. Retrieved from http://urn.fi/urn:nbn:fi:lb-2016011101.

[Weigel et al.2015] 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.

[Wilkinson et al.2016] 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.

[Wittenburg et al.2018] Peter Wittenburg, Dieter Van Uytvanck, Thomas Zastrow, Pavel Straňák, Daan Broeder, Florian Schiel, Volker Boehlke, Uwe Reichel, and Lene Offersgaard. 2018. CLARIN B Centre Checklist. Technical Report CE-2013-0095, CLARIN ERIC. Accessible via http://hdl.handle.net/11372/DOC-78.

Interoperability of Second Language Resources and Tools

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

Silje Ragnhildstveit
Western Norway University of Applied Sciences
Norway
silje.karin.ragnhildstveit@hvl.no

Maarten Janssen
University of Coimbra
Portugal
maartenpt@gmail.com

Kari Tenfjord
University of Bergen
Norway
tenfjord@uib.no

Therese Lindström Tiedemann
University of Helsinki
Finland
therese.lindstromtiedemann@helsinki.fi

Koenraad de Smedt
University of Bergen
Norway
desmedt@uib.no

Nives Mikelić Preradović
University of Zagreb
Croatia
nives.mikelic@gmail.com

Abstract

Language learning based on learner corpora is an increasingly active area of research in CLARIN centres and beyond. In order to promote comparative research, the interoperability of data and tools in this area must be improved, and metadata and error annotation should be harmonized. A closer European collaboration in the field of learner corpus creation is desirable.

1 Introduction

In a changing Europe with increasing migration as well as internal mobility, there is an increased need for research on second language (L2) learning.[1] Among the sources of empirical data for such research are digital language learner corpora (L2 corpora), in which speech or text produced by L2 learners is compiled and annotated. In recent years several L2 corpora have been constructed around Europe, and many of those have been made available through CLARIN repositories and corpus management tools (Lindström Tiedemann et al., 2018). Given that L2 teaching and assessment is becoming a burning issue for a majority of European countries and languages, there is a need to re-assess existing resources and tools in relation to the ones that are under construction or planned, and to explore the possibility of comparative research.

While many CLARIN-related research groups have ongoing, planned or completed work on L2 resources and tools (Ivaska, 2014; Tenfjord et al., 2006; Volodina et al., 2016), these groups have until recently had little contact with each other. Comparative research across languages and national boundaries is non-trivial, not only due the large number of possible L1–L2 pairs but also due to the different approaches taken in the construction of L2 corpora. These include differences in linguistic annotation, differences in error taxonomies – which may be partly language specific – and differences in the metadata about the learners and the learning context. A step towards overcoming these differences would be an increase of the level of interoperability of L2 resources and tools in CLARIN.

[1] Second language or L2 in this article refers to any non-native language; L1 is native language or mother tongue.

On December 6 to 8, 2017, Swe-Clarin and Språkbanken at the University of Gothenburg hosted the *CLARIN Workshop on Interoperability of Second Language Resources and Tools*.[2] The goal of the workshop was to exchange experiences between people working on various approaches to L2 written corpora and tools, and to work towards a common understanding of useful standards, as well as suggestions for corpus composition and design in terms of age groups, proficiency levels, genres, etc.

During the workshop, it became obvious that all research groups face quite similar problems in the construction of L2 corpora, and that best practices are often elusive. Four core problem areas were identified, around which round-table discussions were organized: metadata, tools and formats, error annotation, and user interfaces. The first three of those have directly to do with interoperability, while the last one must also be seen against the background of various research goals, as well as ethical and legal concerns in the handling of learner language. The present paper elaborates on these issues, which highlight the need for interaction between research centres involved in building and using L2 corpora.

2 Interoperability

The interoperability of linguistic resources has been a focus of recent discussions due to their increasing number (Ide and Pustejovsky, 2010). Some state-of-the art approaches deal with both *structural interoperability* (aiming at the same formalism for annotations of different origin) and *conceptual interoperability* (aiming to link annotations of different origin to a common vocabulary) of general linguistic corpora (Chiarcos, 2012, p. 163). In the domain of L2 corpora, there are no formally established standards and there is no clear consensus on interoperability requirements, even though this field would have even more to gain from such a standardisation since learner corpora are mostly small, while researchers are often interested in comparing corpora with respect to learner characteristics, task specifications, etc.

A number of standardization initiatives within language learning have been promoted and developed by the IMS Global Learning Consortium.[3] However, these standards relate to assessment, and are not directly transferable to L2 corpora. Two formats are currently in use by most recent learner corpora. On the one hand, the PAULA/XML format is typically used in combination with ANNIS as a search engine. On the other hand, TEI/XML is typically used in combination with the CWB query language. For the construction of corpora in TEI/XML, a popular tool is TEITOK, which provides a graphical user interface to create, edit, and search often heavily annotated XML files. For querying corpora, different tools are available, examples within CLARIN centers being *Korp*[4] and *Corpuscle*[5].

If L2 corpora were fully interoperable, it would be possible, for example, to compare how learners with the same L1 perform on different L2 tasks with respect to grammatical, lexical, and other levels; and whether the same linguistic categories (e.g. plural versus the present tense) are learned in the same order despite different L1s. There are a lot of potential research scenarios in, for example, Second Language Acquisition (SLA), Foreign Language Teaching and Testing, Computer Assisted Language Learning (CALL) and Natural Language Processing (NLP). However, to achieve this interoperability, there is a need to make sure that L2 corpora have comparable error taxonomies (i.e. mark-up for deviations in orthography, tense, etc), associated metadata variables (e.g. age, gender, task, etc), file formats (e.g. json, xml), corpus design (e.g. L1 grouping), etc.

2.1 Metadata

Documentation of L2 resources is important, both as structured metadata which can be shared in repositories and catalogs like the VLO, and in human-readable form which makes it clear to L2 researchers how the data can be understood and used. Metadata provided by different corpora shows substantial variation. Some corpora have detailed metadata, such as ICLE (Granger et al., 2002) or CEDEL2 (Lozano,

[2] `https://sweclarin.se/eng/workshop-interoperability-l2-resources-and-tools`. Financial support for the workshop was provided by CLARIN ERIC, SWE-CLARIN, CLARINO and Riksbankens Jubileumsfond, the latter through two projects: SweLL – Electronic research infrastructure on Swedish learner language (IN16-0464:1) and Development of lexical and grammatical competences in immigrant Swedish (P17-0716:1).

[3] `https://www.imsglobal.org/specifications.html`

[4] `https://spraakbanken.gu.se/korp/`

[5] `http://clarino.uib.no/korpuskel/page`

2009), while others provide much less, either because there was insufficient information available, or because some information had to be left out for legal reasons, notably privacy considerations (Stemle et al., Submitted). Also, some corpora merge groups of languages (e.g. Serbian and Croatian) in metadata whereas other keep them apart. Such differences cause some problems for comparative research, for instance when age groups are delineated differently. There is increasing convergence, however, on the need of one relatively stable set of recommendable or obligatory metadata for learner characteristics, and on another set for corpus information (Granger and Paquot, 2018).

One of the most frequently used standards for representing the metadata is the TEI/XML header (teiHeader). The fact that TEITOK is used for a growing number of L2 corpora in TEI/XML has the side-effect that several specific extensions to the TEI metadata header initially proposed by the COPLE2 corpus have been adopted by the subsequent corpora using the same tool, leading to what could become a generally accepted extension to the TEI standard.

2.2 Tools

There is no clear lack of tools (see Stemle et al. (Submitted) for an overview of some, e.g. feat, TEITOK, Falko-tools, the SweLL-tool, etc.) but there is a general desire to have less computer savvy interfaces for existing tools. Moreover, there are currently three main obstacles with tools: firstly, it is difficult to know which tools are out there, which – especially for people just starting in the field – makes it hard to find tools that would suit them. The second issue is that too many of the tools are not available for those who would want to use them. And the third issue is that many of the available tools are not properly documented.

Some of the issues could be resolved through a better financing structure in which key tools are financed not by single projects, but through an infrastructure that provides them to the community, or through a commercial firm, as in the case of SketchEngine.[6]

2.3 Error taxonomy

L2 corpora are usually annotated with a description of 'errors', i.e. deviations from the language norm. It is, of course, rather utopian to expect that every L2 corpus project will have the same error taxonomy, e.g. compare taxonomies in Merlin (Boyd et al., 2014) with 64 error tags and ASK (Tenfjord et al., 2006) with 23 error tags. Languages are typologically different and the annotation therefore is at least partly language dependent. But various research projects also have different foci in error annotation which may be reflected in the taxonomy. Dobric (2015), e.g., lists six types of error classification approaches. It seems necessary to initiate steps to overcome differences and allow us to make some generalizations across different L2s. For instance, for languages that have morphosyntactic agreement, it would be useful to converge on a common tag or set of tags for agreement errors that would allow their comparison across languages to the extent possible. Also, annotators should generally avoid annotating the (speculated) cause of the error (such as the L1 influence), but should rather concentrate on a linguistic description of the error.

2.4 User interfaces

Different users may need different ways of using a corpus. Whereas a number of quite sophisticated, interfaces for L2 resources are readily available in web browsers, it is not easy for users to learn how to use the different modes of querying. The new interface for ASK[7] and that for Korp offer simple and advanced modes. On the one hand, a simple search makes it possible to start a query without knowing the details about the query language and the query can be extended through menus. On the other hand, the simple search is also translated to the query language (used) which can be further edited by more advanced users, if desired. A crucial point is that it is easy to switch between the different search modes.

It is often useful to access frequencies and to be able to make a subcorpus, e.g. if one is interested in a group of learners with a specific L1 or from a certain age group or level. Some researchers may want

[6] `https://www.sketchengine.eu/`
[7] `http://clarino.uib.no/ask/ask`

to download a corpus and use it for various purposes, so researchers should see a license which states in clear terms (and with CLARIN license category symbols) what is allowed and what not. Whenever new users start using the corpus, it is of utmost importance that the corpus is well documented in terms of an annotation code book, as well as documentation on the selection of material for the corpus etc., and that this information is easily available.

3 Investigating the CLARIN L2 landscape

Recently, CLARIN initiated an overview of L2 corpora in CLARIN countries[8] with contributions by Lenardič, Lindström Tiedemann and Fišer, partly based on the University of Louvain list of corpora around the world,[9] partly on items in the Virtual Language Observatory,[10] as well as on input from the Gothenburg workshop participants. The report made it clear that CLARIN is already making an important contribution to learner corpus research, with 33 corpora listed in the VLO (plus 3 on other CLARIN sites), but that there are clear needs for improvement of the VLO, inclusion of more documentation and standardisation of metadata (Lindström Tiedemann et al., 2018).

4 Conclusion and outlook

There is a lot of knowledge and there are a lot of data and tools on L2 corpora brought together in the CLARIN community, but researchers planning to construct new resources still face problems in finding out how to select the most appropriate tools and standards. Cooperation towards interoperability has so far not been the norm, but is becoming a priority. Therefore, the most important step is creating a network of the relevant researchers and research groups who are interested in discussing these issues. The first steps toward this goal have now been taken.

We have identified four areas with interoperability issues. These four areas have come forward from our experiences, but they are not at the same level: while error taxonomies and metadata are to some extent theory-dependent and less easily subject to converging practices, tools and user interfaces are more theory-independent but require different degrees of technical skills.[11]

A strength of the CLARIN workshop was that it brought together people of different profiles that work with L2 corpora: technical (such as software engineers, language engineers, NLP specialists) and non-technical (including general linguists, Second Language Acquisition researchers, language testing researchers, corpus linguists, and language teachers). This diversity has been highly productive and will also be beneficial to future initiatives in creating communication and collaboration between the various L2 corpus research centers in Europe.

We hope that the participants at the CLARIN workshop will continue as a working group for people interested in creating L2 corpora, hence facilitating and stimulating the compilation of new and more interoperable corpora. For the extended activities of such a working group, a larger project such as a COST action with strong participation from CLARIN and the Learner Corpus Association could be a promising next step in the right direction.

References

[Boyd et al.2014] Adriane Boyd, Jirka Hana, Lionel Nicolas, Detmar Meurers, Katrin Wisniewski, Andrea Abel, Karin Schöne, Barbora Stindlová, and Chiara Vettori. 2014. The MERLIN corpus: Learner language and the CEFR. In *LREC'14*, pages 1281–1288.

[Chiarcos2012] Christian Chiarcos. 2012. Interoperability of corpora and annotations. In *Linked Data in Linguistics*, pages 161–179. Springer.

[Dobric2015] Nikola Dobric. 2015. Quality measurements of error annotation-ensuring validity through reliability. *The European English Messenger. Volume 24.1*, pages 36–42.

[8] `https://www.clarin.eu/resource-families/L2-corpora`; see also `https://office.clarin.eu/v/CE-2018-1202-L2-corpora-report.pdf`

[9] `https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html`

[10] `https://vlo.clarin.eu/`

[11] Thanks to an anonymous reviewer for this observation.

[Granger and Paquot2018] Sylviane Granger and Magali Paquot. 2018. Towards standardization of metadata for l2 corpora. presentation at the workshop on interoperability of second language resources and tools, 6–8 dec 2017, gothenburg, sweden. `https://sweclarin.se/swe/workshop-interoperability-l2-resources-and-tools`.

[Granger et al.2002] S. Granger, E. Dagneaux, and F. Meunier. 2002. *International Corpus of Learner English*. UCL, Louvain.

[Ide and Pustejovsky2010] Nancy Ide and James Pustejovsky. 2010. What does interoperability mean, anyway? Toward an operational definition of interoperability for language technology. In *Proceedings of the Second International Conference on Global Interoperability for Language Resources*. Hong Kong, China.

[Ivaska2014] Ilmari Ivaska. 2014. The corpus of advanced learner Finnish (LAS2): database and toolkit to study academic learner Finnish. *Apples: journal of applied language studies*.

[Lindström Tiedemann et al.2018] Therese Lindström Tiedemann, Jakob Lenardić, and Darja Fiser. 2018. L2 learner corpus survey – Towards improved verifiability, reproducibility and inspiration in learner corpus research. In *Abstracts from the CLARIN Annual Conference 2018, Pisa, Italy*. CLARIN.

[Lozano2009] C. Lozano. 2009. Cedel2: Corpus escrito del español L2. In Carmen M. Bretones Callejas et al., editor, *Applied Linguistics Now: Understanding Language and Mind*, pages 197–212. Universidad de Almería, Almería, Spain.

[Stemle et al.Submitted] Egon W. Stemle, Adriane Boyd, Maarten Janssen, Therese Lindström Tiedemann, Nives Mikelić Preradović, Alexandr Rosen, Dan Rosén, and Elena Volodina. Submitted. Working together towards an ideal infrastructure for language learner corpora. In *Proceedings of the 4th Learner Corpus Research Conference*.

[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.

[Volodina et al.2016] Elena Volodina, Ildikó Pilán, Ingegerd Enström, Lorena Llozhi, Peter Lundkvist, Gunlög Sundberg, and Monica Sandell. 2016. Swell on the rise: Swedish learner language corpus for European reference level studies. *LREC'16*, pages 206–212.

Tweak Your CMDI Forms to the Max

Rob Zeeman
KNAW HuC DI
rob.zeeman@di.huc.knaw.nl

Menzo Windhouwer
CLARIN ERIC/KNAW HuC DI
menzo.windhouwer@di.huc.knaw.nl

1 Introduction

Metadata records created and provided via the Component Metadata Infrastructure (CMDI) can be of high quality due to the possibility to create a metadata profile tailored for a specific resource type. However, this flexibility comes with a cost: it's harder to create a metadata editor that can cope well with this diversity. In the Dutch CLARIAH project the aim is to create a user-friendly CMDI editor, which is able to deal with arbitrary profiles and can be embedded in the environments of the various partners. Already a few CMDI editors have been created, e.g., Arbil [Withers 2012], CMDI-Maker [CLASS 2018] and COMEDI [Lyse *et al* 2015]. Of these Arbil is not supported anymore and CMDI-Maker only supports a limited number of profiles. COMEDI can handle arbitrary CMDI profiles, but it comes with its own dedicated environment and stays very close to the profile, which makes certain technical limitations of CMDI still leak into the end user's experience. An example is the lack of multilingual labels for elements in the CMDI profile specifications. In this abstract CLARIAH's CMDI Forms (CCF; [KNAW HuC DI 2018a]) is introduced. It supports CMDI 1.2 and can handle any CMDI profile, but also allows various tweaks (usually small adjustments) to enhance usability. CMDI Forms can also be embedded, by a set of plugins, into a specific environment. The next sections will describe these features in more depth.

2 CLARIAH CMDI Forms

CCF is a web-based application, implemented in PHP, JavaScript (jQuery) and CSS. In the basic workflow (see Figure 1) a CMDI profile is, optionally, merged by the backend with its tweaks and then converted into JSON. The JSON is used to construct an entry form within the user's browser. When a metadata record is finished, JSON is send back to the backend where it's converted into a valid XML CMDI record.

2.1 Basic CMDI support

To support the core of CMDI CCF can deal with nested components, elements and attributes and their various value schemes. All of these basic building blocks can be rendered in the entry form for the end user. Also, resources can be selected and associated with specific components. Other meta information, e.g., the creator and creation date of the record, is provided by the backend when creating the XML CMDI record.

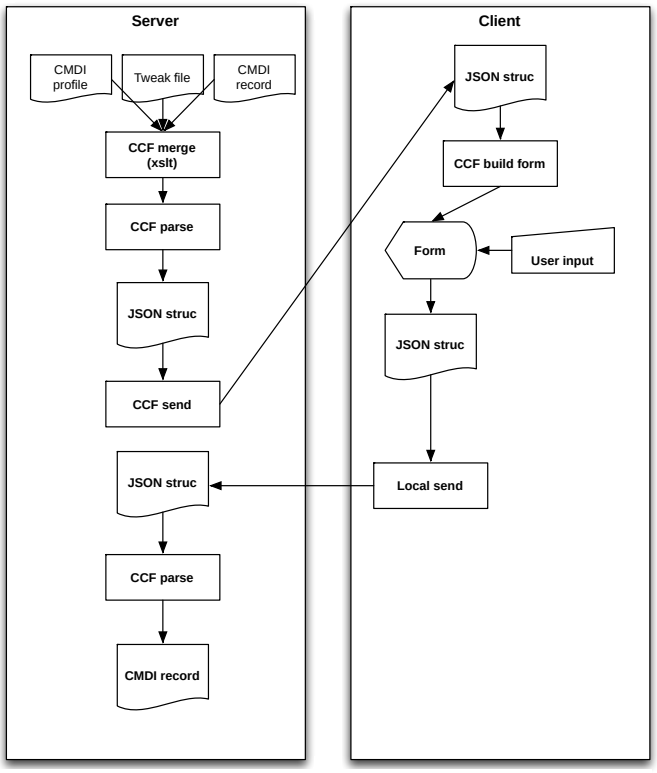


Figure 1. CCF basic workflow

2.2 CMDI 1.2 support

The latest version of CMDI, 1.2 [CMDI Task Force 2016, CLARIN ERIC 2017], added optional new functionality, which can greatly improve the user experience of editing CMDI records. CCF supports this new functionality out-of-the-box.
External vocabularies from the CLARIN vocabulary service CLAVAS [Meertens Institute 2018] can be accessed via CLAVAS's autocomplete API, and CCF uses this so an end user can reduce the number of items to choose from with a few keystrokes.
A number of auto value derivation rules are supported:

1. *now*: will set the value to the current date and/or time;
2. *fileSize*: will set the value to the file size of the associated file, i.e., the nearest resource associated with an ancestor component;
3. *fileModification*: will set the date and/or time of the associated file;
4. *default:* will set the default value for the field;
5. *fallback*: will set the value if the user does not enter any.

Various cues for tools are also supported:

1. *displayOrder*: can be used to change the order of elements and components, e.g., the mandatory and most used elements can be placed at the top of the form for less scrolling; it is even possible to mix elements and components, which is not possible in CMDI profiles;
2. *hide*: hide the component, element or attribute; can be safely used with optional parts of the profile or should be combined with default or fallback auto values to prevent validation problems;
3. *resource*; marks a component to which a resource (proxy) can be associated;
4. *inputWidth*: set the width for an input field, e.g., to allow enough space to enter a long title;
5. *inputHeight*: set the height for an input field, e.g., to allow enough space to enter a description of multiple lines of text.

However, as these cues cannot be specified in the Component Registry yet, they are always added via the tweak file, whose role and functionality is described in the next section.

2.3 Tweaking a CMDI form to the max

Any CMDI profile can, optionally, be tweaked in CCF. The tweaks include basic CMDI features, i.e., it is possible to override constraints, as long as the original constraints are not violated, e.g., an optional element can become mandatory or a string element can be limited by a closed vocabulary in the tweak file. Basically, the original CMDI profile and the tweaks are merged into a derived CMDI profile whose instances, the CMDI records created with CCF, will also be valid instances of the original CMDI profile.
Any of the CMDI 1.2 features, auto values and cues, can also be added to the tweaks, which is handy if one is not the owner of a CMDI profile but still wants to use these new features.
The tweak file also allows adding some additional multilingual information, which will not fit nicely in a cue for tools. One can use labels for translation and/or enhanced readability of component, element and attributes names, and also for known, i.e., vocabulary items, values.
Minimizing errors can be obtained with a build-in form validation. Basic validation is done according to the defined constraints of the (derived) CMDI profile. On top of this type of validation conditions can be defined to ensure consistency of the data entered in the form. For instance, if a resource is not digital, no values referring to disk type and file size can be filled in.

This an example of a tweak file (trimmed here and there to limit the size):

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<ComponentSpec xmlns:clariah="http://www.clariah.eu/" xmlns:cue="http://www.clarin.eu/cmdi/cues/1">
    <Header>
        <ID>clarin.eu:cr1:p_1440426460262</ID>
    </Header>
    <Component name="MeertensCollection">
        <clariah:label xml:lang="nl">Meertens collectie</clariah:label>
        <Component name="CoreCollectionInformation">
            <clariah:label xml:lang="nl">Core-collectie informatie</clariah:label>
            <Element name="creator">
                <clariah:label xml:lang="nl">Auteur</clariah:label>
            </Element>
            <Element name="title" cue:displayOrder="2">
                <clariah:label xml:lang="nl">Titel</clariah:label>
            </Element>
            <Element name="description" CardinalityMin="1" cue:inputWidth="80" cue:inputHeight="5">
                <clariah:label xml:lang="nl">Beschrijving</clariah:label>
            </Element>
            <Element name="language">
                <clariah:label xml:lang="nl">Taal</clariah:label>
                <ValueScheme>
                    <Vocabulary URI="https://hdl.handle.net/11459/CLAVAS_810f8d2a-6723-3ba6-2e57-41d6d3844816" ValueProperty="notation"/>
                </ValueScheme>
            </Element>
            <Element name="collectionID" cue:displayOrder="1">
                <clariah:label xml:lang="nl">Collectie ID</clariah:label>
            </Element>
            <Element name="category">
                <clariah:label xml:lang="nl">Categorie</clariah:label>
                <ValueScheme>
                    <Vocabulary>
                        <enumeration>
                            <item>
                                <clariah:label xml:lang="nl">MI Algemeen</clariah:label>
                                <clariah:value>MI General</clariah:value>
                            </item>
                            <item>
                                <clariah:label xml:lang="nl">Dialectologie/taalvariatie</clariah:label>
                                <clariah:value>Dialectology/language variation</clariah:value>
                            </item>
                            ...
                            <item>Realia</item>
                        </enumeration>
                    </Vocabulary>
                </ValueScheme>
            </Element>
        </Component>
        <Component name="Inventory">
            <Component name="CoreResourceInformation" cue:resource="yes">
                <Component name="TechnicalMetadata">
                    <clariah:label xml:lang="nl">Technische metadata</clariah:label>
                    <Component name="Size">
                        <Element name="number" cue:inputWidth="5">
                            <AutoValue>fileSize</AutoValue>
                        </Element>
                    </Component>
                </Component>
            </Component>
        </Component>
    </Component>
</ComponentSpec>
```

So, the tweak file mimics directly the structure of the original CMDI profile. The component and elements are identified by their names, which is mandatory. But only the tweaked part of the profile is duplicated, other XML attributes or elements from the profile are only needed when they are overruled, e.g., *description* has become mandatory and *language* got associated with a CLAVAS vocabulary. The *number* element in the *Size* component has an auto value rule to make sure that the actual file size of the associated resource is filled in. A directly related cue can also be seen, i.e., the *coreResourceInformation* component is marked as a place to associate a resource (proxy). Also, the mandatory *collectionID* and *title* element have *displayOrder* cues to make them appear first in the form (see Figure 2). Last but not least, labels are added to make it possible to render the whole form in Dutch.

3 Embedding the CLARIAH CMDI Forms core in your environment

The CCF core can be integrated into existing web-based applications. It consists of the server and client side components depicted in Figure 1. For embedding in one's own environment one class has to implemented, which takes care of loading an existing CMDI record and saving a new or updated CMDI record and the associated new resources. Various other behaviors can be configured, e.g., the default language for the form and the set of mandatory or optional languages for multilingual elements or attributes. This part of the CCF development is currently test driven in the context of the FLAT repository [Trilsbeek *et al* 2016] instance at the Meertens Institute. It forms the core of its Collection Management Interface.

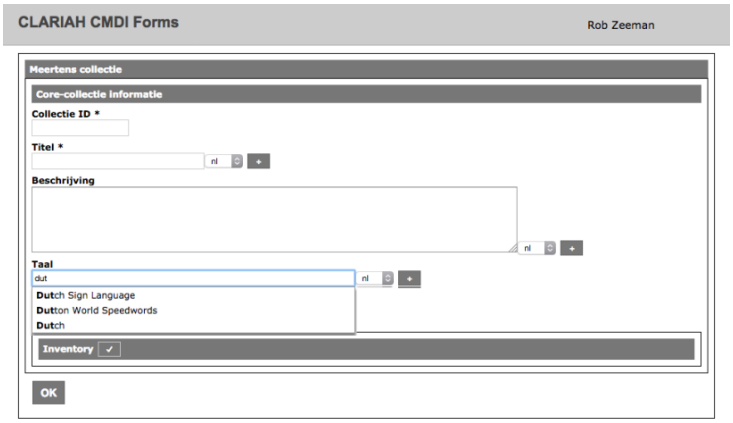


Figure 2. CCF in action

Next to FLAT, work is also underway to let the core of CCF provide the edit interface for data in Timbuctoo [KNAW HuC DI 2018b]. Timbuctoo works with linked data, which means that the step from CMDI to and from JSON (see the CCF basic workflow in section 2) will become adaptable as well, i.e., can be replaced by a procedure which translates linked data to, and from, the same JSON structures.

4 Conclusions

The CLARIAH CMDI Forms application provides a new CMDI editing environment for records based on arbitrary profiles, but with extensive possibilities, based on CMDI 1.2 features and some extensions, to tweak the profile information for a maximum user-friendly editing experience. In addition to this CCF can be embedded in other environments.

References

[CLARIN ERIC 2017] CLARIN ERIC. 2017. *CMDI 1.2*, https://www.clarin.eu/cmdi1.2 Accessed April 29, 2018.

[CLASS 2018] CLASS - Cologne Language Archive Services. 2018. *CMDI-Maker*, http://cmdi-maker.uni-koeln.de/ Accessed April 4, 2018.

[CMDI Task Force 2016] CMDI Task Force. 2016. *CMDI 1.2 specification*. CE-2016-0880, CLARIN ERIC, Utrecht, The Netherlands.

[KNAW HuC DI 2018a] KNAW Humanities Cluster – Digital Infrastructure. 2018. *CLARIAH CMDI Forms*, https://github.com/knaw-huc/clariah-cmdi-forms Accessed September 7, 2018.

[KNAW HuC DI 2018b] KNAW Humanities Cluster – Digital Infrastructure. 2018. *Timbuctoo*, https://timbuctoo.huygens.knaw.nl/ Accessed September 7, 2018.

[Lyse *et al* 2015] G. I. Lyse, P. Meurer, K. De Smedt. 2015. COMEDI: A component metadata editor. In *Selected Papers from the CLARIN 2014 Conference*.

[Meertens Institute 2018] Meertens Institute. 2018. *CLAVAS: vocabulary service*, https://vocabularis.clarin.eu/clavas/ Accessed September 10, 2018.

[Trilsbeek *et al* 2016] P. Trilsbeek, M. Windhouwer. 2016. *FLAT: A CLARIN-compatible repository solution based on Fedora Commons*. At the CLARIN Annual Conference. Aix-en-Provence, France, October 26 - 28, 2016.

[Withers 2012] P. Withers. 2012. Metadata Management with Arbil. In *Proceedings of the workshop Describing LRs with Metadata: Towards Flexibility and Interoperability in the Documentation of LR*, Istanbul, May 22, 2012.

CLARIN Data Management Activities in the PARTHENOS Context

Marnix van Berchum
Huygens ING, KNAW
Amsterdam, Netherlands
marnix.van.berchum@huygens.knaw.nl

Thorsten Trippel
University of Tübingen
Tübingen, Germany
thorsten.trippel@uni-tuebingen.de

Abstract

Data Management is one of the core activities of all CLARIN centres providing data and services for the academia. In PARTHENOS, European initiatives and projects in the area of the humanities and social sciences assembled to compare policies and procedures. One of the areas of interest is data management. The data management landscape shows a lot of proliferation, for which an abstraction level is introduced to help centres, such as CLARIN centres, in the process of providing the best possible services to users with data management needs.

1 Introduction

Data management is the activity of creating, providing, maintaining and archiving research data over all stages of the research data life cycle (see Pennock, 2007). CLARIN centres working with data, operating repositories, and providing services to their users all work in the area of data management.

Each certified CLARIN-B centre provides some services (see Wittenburg et al 2013-2018, paragraphs 1a and 1d) and deals with the processes and technology required for data management. CLARIN does not operate independently. To avoid duplication of work, CLARIN makes use of open interfaces and compatibility layers to other systems. All certified CLARIN centres provide a level of trust documented by certification authorities (such as the Core Trust Seal, www.coretrustseal.org) and others to make the requirements and processes of data management transparent. Almas et al. (2016) describe the current diverse situation of data management and lay out possible ways for further development, including the establishment of policies. Independent of the documentation of current practices, all parties involved recognize the need to follow procedures and guidelines early on in the process, starting with a data management plan (DMP) ideally before data are created. The creation of DMPs is becoming more and more a requirement by research funders as well. Consequently CLARIN centres can provide assistance to their users in the process of DMP creation. To do this efficiently, they need to have an understanding on the requirements of funders for DMPs.

To synchronize and harmonize activities, CLARIN ERIC is part or the European PARTHENOS project (www.parthenos-project.eu). In this paper, we describe the current situation with regards to data management in CLARIN and PARTHENOS, the data policy implementation as required by funding organization and according to best scientific practice. We outline an abstraction level that is under discussion between partners in PARTHENOS and that can be applied by CLARIN centres.

2 Current situation with regards to data management in PARTHENOS and CLARIN

Amongst the PARTHENOS partners no common data management practices exist. The disciplines represented in the project – defined as 'the broad sector of Linguistic Studies, Humanities, Cultural Heritage, History, and Archaeology' – all have a different history with regard to data management. Some have archiving experience for objects and artefacts but less with digital born data (e.g. archeology), others are completely new to the field of data management.

Within PARTHENOS, CLARIN represents the disciplines dealing with all forms of language and text data. Although, as a well established research infrastructure, CLARIN has a long background in the handling (use, sustainably maintain, provide) of data, there is no overarching, fixed policy concerning the management of these data. This is contrary to the CLARIN Centre Assessment procedure which requires a finished, or rather 'at least initiated', Core Trust Seal application, implying several technical and organisational requirements for the CLARIN Centres (see CTS, 2017 in the references). These include requirements on backup solutions, persistent identification, minimal metadata, licence recommendation, and access restriction implementation. Individual CLARIN centres operating repositories may have additional requirements regarding data formats or have specific depositing agreements in place (see CCA, undated, in the references). Some CLARIN Centres even provide Data management plans as a service to scholars (see Trippel & Zinn (2015) or published handbooks and documentation (see for example Herold & Lemnitzer (2012), chapter 2) as reference for users.

Despite the organisation around a core technical infrastructure of tools and resources that imply ideas of data management, CLARIN has not yet recommended centralised data policies on its partners. From this background, CLARIN joined the activities by PARTHENOS, which has as one of the goals the design of Data Management Plan templates, which serve larger communities. For this purpose, PARTHENOS reviewed existing, discipline independent DMP templates provided by research funders.

3 Data Policy Implementation

Current policy of most funders and research organizations is to rely on FAIR data (see FAIR, undated, in the references), i.e. data needs to be Findable, Accessible, Interoperable and Re-usable. CLARIN is dedicated to the FAIR principles documented by some infrastructure components: the VLO and FCS are good examples of making data in the CLARIN world findable, for details on CLARIN's efforts and mission with regards to FAIR, see de Jong, et al. (2018). With resolvable PIDs the data becomes accessible – even if access is restricted CLARIN provides access mechanisms with the Identity Provider and Shibboleth architecture. Interoperability is achieved by the utilization of standards such as the ISO TC 37 SC 4 endorsed standards or TEI. However, interoperability is dependent on the selection of the flavour of these standards and the support by software. The Language resource Switchboard (see Zinn, 2017) is a tool to provide opportunities to interoperability of data. Interoperability and Reusability are not only addressed by technology, but also by policies, legal restrictions, infrastructure components etc. Re-usability has the additional requirement that researchers are also allowed to reuse data and have access to the tools to do this. These policies and decisions need to be documented for data to become interoperable and reusable.

Reconstructing of FAIR relevant documentation with finished data sets is virtually impossible – especially when third parties were involved and hold rights for data and software. To avoid that, these aspects are documented before data is assembled or in the processes of data creation. As all of this is part of the data management plan, the data management plan is the key to successfully implement a FAIR data policy.

The situation with regards to the central nature of DMPs for a FAIR data policy is currently only partially reflected by the support for data management plans. In fact, there is obvious proliferation in communities and funding institutions with regards to their requirements for DMPs. The requirements range from no formal requirements besides providing DMPs (e.g. German Research Foundation) to detailed templates (e.g. Horizon 2020, see the Guidelines on FAIR Data Management in Horizon 2020 (2016) in the references). For the United Kingdom, Jones (2012) already provided a summary of eight (national) funding organizations and their requirements. Some academic institutions have their own requirement of good scholarly practice, for example at the University of Edinburgh (see Research Data Service in the references), providing additional requirements. To ease the situation the Data Curation Centre (DCC) in the UK provides a website with an interactive template for various funders, called DMPonline (see "DMPonline" (2010-2018) in the references). Though this is open source software which is extensible and many data management plans show significant overlap, DMPonline cannot cover all funders, disciplines and data centre related requirements. Despite this – conceptual –

shortcoming, the DMPs created with such a template have the benefit of documenting awareness of scholars in the data management process, which is a requirement for long-term digital preservation.

Within PARTHENOS a detailed report was created, documenting the proliferation, complexity and issues with regards to data management (see Hollander et al., 2017, Chapter 3). In the same document (ibid. Appendix III, Section 9.2) a unified template was created, which failed the usability test because of its complexity as the variety of descriptive levels including funders, data centres and disciplinary requirements are not sufficiently distinguished and the scholar is at loss with the template.

4 Abstraction for Data Management Plans: Data Management Protocols

For the creation and reuse of elements of the DMP templates, an abstraction layer could be helpful. Such an abstraction layer for DMPs takes common components of DMPs into account that depend on the funders' requirements and the requirements of data centres. This abstraction layer is termed DMP protocol, which could also be seen as a template for DMP templates used by funders and data centres alike.

The DMP protocol conceptually is a template for creating DMP templates and could be used by funders and data centres alike. The DMP protocol takes the DMP requirements of (1) a funder (2) a discipline (3) a data type (4) a data centre and (5) a DMP protocol to generate (6) a DMP template to be filled in for a concrete research project within a discipline and with a specific type of data, utilizing a specific data centre and tailored to the project reviewing needs of a funder. Though at present this DMP protocol is a conceptual model, it opens up the way to an implementation that helps to ease the DMP process by omitting unnecessary questions to a researcher and directing the attention to essential issues that are indeed project specific.

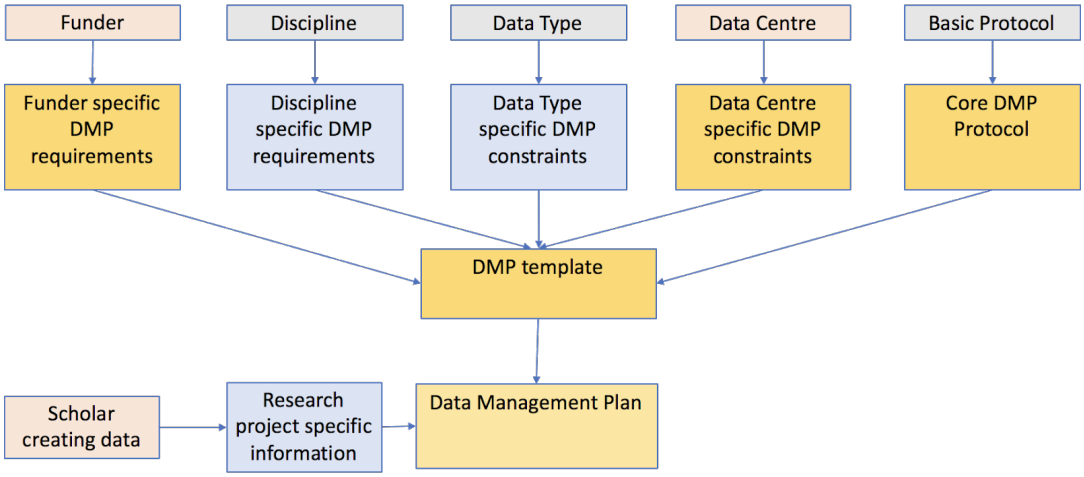


Figure 1 Schematic structure of a data management protocol for creating DMP templates and DMPs

The funder specific requirements, data centre specific constraints and core protocol in this model are defined by policies of institutions, while the discipline specific requirements and data types are based on practices and have to be extrapolated by experts on the field and in data modelling, hence they are intrinsic to the discipline and data type. Similarly, the research project information directly relates to the project, which is defined by the scholar, hence this information is intrinsic to the project. It seems obvious that there is a difference in the level of formalization possible on the input side. Formalising the policies of data centres and funders together with a core protocol are institutional decisions and can be achieved, if the institutions decide to do so. The discipline and data type specifics are harder to formalize as they rely on implicit best practices in a discipline, interpretation and formalization.

5 Future Work: CLARIN implementation of the DMP Protocol using CMDI

The abstraction model for data management plans shows the five different components going into a template plus the project specific information. To implement these, it would be possible to define components or partial documents for each and integrate it by means of for example standard XML technologies (XML include). As the documents are not necessarily prefilled, another option would be to utilize ISO/DIS 24622-2 based CMDI to define (CMDI-)components for each, such as a funder specific component, a discipline specific component, a data type specific component, etc. The result would be a CMDI-Profile dependent on all the components, which results in an XSchema for a DMP for each of these. A sample profile would be the developmental DMP-Protocol profile (CLARIN component registry ID p_1527668176067, see https://catalog.clarin.eu/ds/ComponentRegistry/rest/registry/1.x/profiles/clarin.eu:cr1:p_1527668176067/xsd). Using standard CLARIN technology such as the COMEDI CMDI editor (Lyse et al., 2015), a DMP could then be written as a CMDI document, to be transformed into a DMP in a layouted format, for example based on HTML.

This implementation might look awkward as it is reusing a technology that is used for data descriptions elsewhere. However, the details specified in the DMP should also be included in the metadata description of the research data created within a project, hence the CMD-modelled information of a DMP will be reused and updated in the metadata for the research data. Other parts of the DMP are not part of the metadata created in the project, such as details on DMP budgets, repository assessment information, etc. The CMDI implementation of DMPs looks promising and should be further explored. As the competence and technology is available within CLARIN, the CLARIN researchers within PARTHENOS will work on that.

6 Conclusion

There is a gap between disciplinary requirements and wishes concerning data management, funder requirements, and ideas by interdisciplinary initiatives such as PARTHENOS. The latter project would be the right platform to compare and assess the different data management practices across the humanities, but it is still complicated to reach common, generic data management policies or services. As a disciplinary infrastructure CLARIN should define a default policy for DMP for its centres and require that in the Centre Assessment Process. This policy could feed into the work of PARTHENOS.

The abstraction of the DMP protocol (paragraph 4) should be taken into account when CLARIN defines the default policy. The CLARIN template should cover on a general level the disciplinary and data type specific aspects of the DMP. Each centre can fill in their data centre specific needs. Funder requirements would have to be left out at present, as there is too much of a variety there. However, a template for CLARIN could build on top of the HORIZON 2020 template as a starting point.

Reference

Almas, B.; Bicarregui, J.; Blatecky, A.; Hill, S.; Lannom, L.; Pennington, R.; Stotzka, R.; Treloar, A.; Wilkinson, R.; Wittenburg, P.; Yunqiang, Z. (2016): Data Management Trends, Principles and Components - What Needs to be Done Next? Research Data Alliance (RDA). Available at http://hdl.handle.net/11304/7721971a-23f3-4eed-8df8-739ff0f2bc6e.

CCA (undated). Assessment procedure. Available at https://www.clarin.eu/content/assessment-procedure.

CTS (2017). Core Trustworthy Data Repositories Extended Guidance: Core Trustworthy Data Repositories Requirements for 2017–2019, Extended Guidance. Version 1.0: October 2017. Available at https://www.coretrustseal.org/wp-content/uploads/2017/01/20171026-CTS-Extended-Guidance-v1.0.pdf.

Peter Doorn (ed., 2018) 'Science Europe Guidance Document Presenting a Framework for Discipline-specific Research Data Management': D/2018/13.324/1, available at: https://www.scienceeurope.org/wp-content/uploads/2018/01/SE_Guidance_Document_RDMPs.pdf

DMPonline. (2010-2018). Digital Curation Centre (DCC). Available at https://dmponline.dcc.ac.uk/.

FAIR (undated): THE FAIR DATA PRINCIPLES. FORCE11. Available at https://www.force11.org/group/fairgroup/fairprinciples.

Guidelines on FAIR Data Management in Horizon 2020 (2016). Available at http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf.

Herold, A.; Lemnitzer, L. (2012): CLARIN-D User Guide. BBAW, Berlin. Available at http://media.dwds.de/clarin/userguide/userguide-1.0.1.pdf.

Hollander, H. et al. (2017) Report on Guidelines for Common Policies Implementation. Project Deliverable D 3.1. Available at http://www.parthenos-project.eu/Download/Deliverables/D3.1_Guidelines_for_Common_Policies_Implementation.pdf.

ISO/DIS 24622-2 (2018) Language resource management -- Component metadata infrastructure (CMDI) -- Part 2: The component metadata specification language. Draft International Standard.

De Jong, F.; Maegaard, B.; de Smedt, K. Fišer; van Uytvanck, D. (2018): CLARIN: Towards FAIR and Responsible Data Science Using Language Resources. In N. Calzolari, et al. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). Miyazaki, Japan. Available at http://www.lrec-conf.org/proceedings/lrec2018/pdf/575.pdf .

Jones, S. (2012): Summary of UK research funders' expectations for the content of data management and sharing plans. Digital Curation Centre (DCC). Available at http://www.dcc.ac.uk/sites/default/files/documents/resource/policy/FundersDataPlanReqs_v4%204.pdf.

Lyse, G. I.; Meurer, P.; De Smedt, K. (2015): COMEDI: A component metadata editor. In: Selected Papers from the CLARIN 2014 Conference, October 24-25, 2014 (8), Soesterberg, The Netherlands, 82-98. Available at https://www.clarin.eu/sites/default/files/cac2014_submission_13_0.pdf.

Pennock, M. (2007): Digital Curation: A life-cycle approach to managing and preserving usable digital information". In: Library & Archives Journal, (1). Available at http://www.ukoln.ac.uk/ukoln/staff/m.pennock/publications/docs/lib-arch_curation.pdf.

Research Data Service. Website of the University of Edinburgh's research data services for local data management services., The University of Edinburgh. Available at https://www.ed.ac.uk/information-services/research-support/research-data-service.

Trippel, T.; Zinn, C. (2015): DMPTY - A Wizard For Generating Data Management Plans. In: Selected Papers from the CLARIN Annual Conference 2015, October 14–16, 2015, Wroclaw, Poland, (123), 71-78. Available at http://www.ep.liu.se/ecp/123/006/ecp15123006.pdf.

Wittenburg, P. et al. (2013-2018) CLARIN B Centre Checklist, Version 6, Last Update: 2018-02-07, Status Approved by the Centre Committee, https://office.clarin.eu/v/CE-2013-0095-B-centre-checklist-v6.pdf.

Zinn, C. (2017) The CLARIN Language Resource Switchboard. CLARIN Annual Conference 2016. https://www.clarin.eu/sites/default/files/zinn-CLARIN2016_paper_26.pdf.

Integrating language resources in two OCR engines to improve processing of historical Swedish text

Dana Dannélls, Leif-Jöran Olsson
Språkbanken
University of Gothenburg, Sweden
{dana.dannells/leif-joran.olsson}@svenska.gu.se

Abstract

We are aiming to address the difficulties that many History and Social Sciences researchers struggle with to bring in non-digitized text into language analysis workflows. In this paper we present the language resources and material we used for training two Optical Character Recognition engines for processing historical Swedish text written in Fraktur (blackletter). The trained models, resources and dictionaries are freely available and accessible through our web service, hosted at Språkbanken, to enable users and developers easy access for extraction of historical Swedish text that are only available in images for further processing.

1 Introduction

The lack of availability of digital text restricts many History and Social Sciences researchers for carrying out parts of their actual research. This limitation also puts some restrictions on the possibility to enrich the texts with linguistic annotations for further processing and discovery. To achieve some of the proposed goals and increase the success of infrastructures like CLARIN we would like to remove this threshold and scarcity. One way to accomplish this is by training Optical Character Recognition (OCR) models to correctly recognise digitized text and enable accurate language analysis and processing. In this paper, we present our attempt to improve the recognition results for processing historical Swedish text of two free OCR engines: Tesseract and Ocular. For each engine we trained a language model using dictionaries and wordlists compiled to cover various time periods of written Swedish. The models and wordlists have been made available through our web service licenced under an open CC-BY license. The service currently allows users to access and run the OCR engines for interpretation of images containing historical Swedish text. A machine-to-machine API will also become available shortly.

2 Tesseract and Ocular

2.1 Tesseract

Tesseract is an open source engine that is currently developed by Google. The latest Tesseract (version 4.0) engine is based on Long Short Term Memory (LSTM) Recurrent Neural Networks (Smith, 2007).[1]

The engine comes with some pre-trained OCR models for various languages and facilities for training new languages and character sets for purposes of improving the existing models. Language modeling is done primarily through dictionaries. Training of models is easily established by following the detailed specification of the training process which is available through the Tesseract's web page.

As a part of the project *A free cloud service for OCR* (Borin et al., 2016), we trained a model for processing Swedish blackletter. During the training process, we retained the model starting with the pre-trained language model that is available for handling Swedish blackletter and added specific wordlists and lists of words with frequencies extracted from Swedish material (Section 3.2).

[1] <https://github.com/tesseract-ocr/>

Data	Description	Entries
Swedberg	18th century dictionary	16,158
Dalin	19th century dictionary	509,924
1734 års lag	corpus	61,220
Runeberg	selected set of titles	711,080

Table 1: Dictionaries and wordlists covering the 18th and 19th centuries.

2.2 Ocular

Ocular is an open source engine (Garrette and Alpert-Abrams, 2016). The engine does not come with any pre-trained OCR models for Swedish. The only model that was available before this experiment is the one provided by us, and that is accessible in eXist-db app ocular-ocr.[2]

Ocular language modeling is done primarily through creating n-gram models on token and word level in addition to creating the font model from the transcribed gold standard and its corresponding images. Secondly dictionaries can be used during model training as well as for other parts of the post-processing. We trained Ocular on a smaller set of documents than those we used for training Tesseract. This is because, at the time of training several pages were not possible to transcribe due to shortcomings that are being addressed in making the neural net setup more robust. Similar to Tesseract, we experimented with different dictionaries during the post-processing stage.

3 Language resources and material for processing historical Swedish text

3.1 Processing historical text

The specific task of processing historical text and in particular, blackletter font should not to be underestimated. Printed texts in blackletter often give poor OCR results because character and line boundaries may be faulty. The distinguished shape of the font lead to common OCR errors. Some of the problems of processing blackletter font are addressed by Furrer and Volk (2011). Blackletter characters are normally confused with other characters, such as 'n' and 'u' with 'll'. Many errors could be avoided if the model is trained with sufficient amount of data and suitable dictionaries (Breuel et al., 2013).

3.2 Dictionaries and wordlists

One of the challenges with OCR engines that are trained with language resources is finding full coverage dictionaries and wordlists that cover the vocabulary of the processed material. Full coverage dictionaries and wordlists are essential for post correction of OCR errors because missrecognition of characters in a word can be corrected if the missrecognized word appears in the lexica. Consider for example the word 'trää', which was incorrectly recognized as 'trciä', where 'ä' has been misplaced with 'ci'. In another example 'krciget' was incorectly recognized as 'krijget, in this example 'ji' was misplaced with 'ci'. These kind of errors can be difficult to correct with simple replacement rules. Although there has been research that shows that large dictionaries might not always lead to improvements if the accuracy of the OCR is high (Smith, 2011), we wanted to explore this simple method with our lexical resources.

We experimented with dictionaries and wordlists that were compiled in the OCR project (Borin et al., 2016). These comprise data which was extracted from: (1) Swedberg, a dictionary over 18th century (Borin and Forsberg, 2011), (2) Dalin, a dictionary over 19th century, covering the morphology of late modern Swedish, (3) the corpus 1734 års lag,[3] and (4) a selected set of proofread works from the Projekt Runeberg.[4] An overview of these resources is presented in Table 1, all are available from our web page.[5]

[2] <https://github.com/ljo/exist-ocular-ocr/>
[3] <https://spraakbanken.gu.se/swe/resurs/lag1734>
[4] <http://runeberg.org/>
[5] <https://spraakbanken.gu.se/eng/ocr>

Engine	Character accuracy	Word accuracy
Tesseract 3.0.5.01	**83.55%**	**65.55%**
Tesseract 3.0.5.01 + wordlists	80.50%	59.90%
Tesseract 4.0 + wordlists	82.03%	64.08%
Ocular 0.3	57.80%	52.67%
Ocular 0.3 + wordlists	62.93%	59.97%

Table 2: OCR error evaluation results at character and word levels from three engines that were trained with and without wordlists.

3.3 Training material

Both Tesseract and Ocular were trained on a selected set of documents that we produced as a part of the OCR project (Borin et al., 2016). This set comprises a randomly selected samples of 199 pages, which were transcribed, i.e. double-keyed, by an external company. The font of the material is in blackletter, together with the transcription, the material is available for download.

3.4 Testing material

The Library of University of Gothenburg, UB, has scanned and OCR-generated historical Swedish texts in blackletter from 1800 for training and evaluation. The OCR quality is not good enough to make it usable but gives a baseline with 25.7 percent word accuracy. For evaluation purposes we selected a small set of this material, which we manually transcribed. In total, the material comprises 10,718 characters and 2,152 words.[6]

4 Word and character accuracy of Tesseract and Ocular

The results reported in this section were obtained with the evaluation toolkit that has been developed at the Information Science Research Institute (ISRI) (Bagdanov et al., 1999).[7] We evaluated the results with respect to character and word accuracy levels. The OCR error evaluation results of the three tested engines are specified in Table 2. Ocular is early in its development stages, but shows promising results even though it was only trained on a small set of documents.

The results show that Tesseract 3.0.5.01 reaches a character accuracy of 83.55% and a word accuracy of 65.55%, which is the highest results we have reached for processing historical print. For this specific test material (described in Section 3.4) the results are slightly better than the results measured for Tesseract 4.0 and considerably higher compared to the results given by Ocular. Interestingly, the highest score was achieved without wordlists.

Even though the results look promising, about 35% of the words in the test material are missrecognized. A closer look at the results shows that the majority of missrecognized words occur because of spelling variations, for example occurrences of 'w' instead of 'v', in words such as 'qwinnors', 'qwarts', 'swarade' that are not covered in the dictionaries. Another major problem is incorrect character recognition, more notably, occurrences of duplicates, such as 'rr', 'nn', 'll' in the beginning or the end of a word. Hyphenated words have not been joined in the transcriptions, something that is reflected in the evaluation where many word parts are singled out.

5 Conclusions

We presented the results of two OCR engines which have been trained on Swedish blackletter text and processed with dictionaries and wordlists extracted from Swedish language resources and corpora. Although the evaluation is based on a very small amount of test data, it shows some promising results.

[6] <http://hdl.handle.net/10794/swedish-blackletter-ocr-evaluation-material-2018>
[7] <http://code.google.com/p/isri-ocr-evaluation-tools>

In the nearest future we are planning to exploit these resources further with larger evaluation sets and optimize the wordlists to different materials. The language resources presented in this paper are freely available for download under CC-BY license.

Acknowledgements

The research presented here was supported by Swe-Clarin, a Swedish consortium in Common Language Resources and Technology Infrastructure (CLARIN) financed by the Swedish Research Council for the period 2014–2018 (grant agreement Dnr 821-2013-2003).

References

A.D. Bagdanov, Stephen V. Rice, and T.A. Nartker. 1999. The OCR Frontiers Toolkit. Version 1.0. Technical report, Information Science Research Institute.

Lars Borin and Markus Forsberg. 2011. A diachronic computational lexical resource for 800 years of Swedish. In *Language technology for cultural heritage*, pages 41–61. Springer:Berlin.

Lars Borin, Gerlof Bouma, and Dana Dannélls. 2016. A free cloud service for OCR / En fri molntjänst för OCR. Technical report, Department of Swedish, University of Gothenburg. `https://gupea.ub.gu.se/bitstream/2077/42228/1/gupea_2077_42228_1.pdf`.

T. M. Breuel, M. Al Azawi A. Ul-Hasan, and F. Shafait. 2013. High performance OCR for printed English and Fraktur using LSTM networks. In *International Conference on Document Analysis and Recognition*.

Lenz Furrer and Martin Volk. 2011. Reducing OCR Errors in Gothic-Script Documents. In *Proceedings of Language Technologies for Digital Humanities and Cultural Heritage Workshop*, pages 97–103.

Dan Garrette and Hannah Alpert-Abrams. 2016. An unsupervised model of orthographic variation for historical document transcription.

Ray Smith. 2007. An overview of the Tesseract OCR engine. In *Proc. of the 9^{th} Int. Conference on Document Analysis and Recognition (ICDAR)*, pages 629–633. IEEE Computer Society.

Ray Smith. 2011. Limits on the application of frequency-based language models to ocr. In *Proceedings of the 2011 International Conference on Document Analysis and Recognition*, ICDAR 11, pages 538–542, Washington, DC, USA. IEEE Computer Society.

Looking for hidden speech archives in Italian institutions

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

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

Abstract

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 are presented and discussed. A huge amount of speech archives is especially preserved among researchers: the most part of the resources is not accessible and legal issues are generally not deeply addressed. The great majority of the respondents would agree in storing their archives in national repositories, if any.

1 Introduction

Very few censuses 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 census was made under the Ministry of Cultural Heritage aegis 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 censuses 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 in the net, 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 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. 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 census of institutional archives, that is a census of the existing speech archives deposited in (and by) institutions and associations;
ii) draft a census 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*; §3 reports the sample that answered to the survey together with the main results; §4 addresses some conclusion remarks and underlines the urgent need to find an Italian repository to host these materials.

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

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). Other formal and informal networks were targeted (e.g. Analisi dell'Interazione e della Mediazione, 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 (three open questions were also included and for which it was impossible to predict or suggest options). 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 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 contained the actual questionnaire consisting of 19 questions. The first question 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 18 questions, was devised in order to obtain a rough description and quantification of audiovisual resources (also with respect to accessibility and legal issues). One last question asked the respondents if they were aware of the existence of the CLARIN European infrastructure;
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 and due to space limitations, 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 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, 149 respondents took part in the survey: 130 participants completed the survey, 17 opted-out and 2 only suggested other contacts.

[2] The survey (available at https://goo.gl/8uHYK1) started on February 20th, 2018 and will be kept active until October, 2018. This will allow the authors to continue the census by reaching more respondents.

Since for most of the questions the participants were allowed to select multiple responses and eventually to specify further responses on an extra field, we also provide the number of items to which the percentages refer to, where required.

3.1 Spoken resources and their scientific domains

With reference to the scientific domains to which the resources belong to, the most mentioned by the respondents are, in decreasing order: *Oral History* (n = 53), *Phonetics & Phonology* (n = 35), *Dialectology* (n = 30), *Anthropological Linguistics* (n = 30), *Sociolinguistics* (n = 9), *Applied Linguistics* (n = 8), *Ethnomusicology* (n = 8), *Sociology* (n = 5), *Language Acquisition* (n = 4), *Speech Technology* (n = 2).[3]

After grouping the same responses 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. The majority of the participants we were able to reach indicated *Linguistics* (57.6%) and *Oral History* (32.1%) as core domains for their resources, with a minor portion of them indicating *Ethnomusicology* (4.8%), *Sociology* (2.4%) and *Speech Technology* (1.2%); *Other* & NA's (1.8%).

3.2 Type of resources involved

When collecting speech in the different domains the spoken productions can be recorded as a uni-modal signal (e.g. Audio only) or as a bi-modal signal (e.g. Audiovideo). This consideration led the authors to include this distinction in the survey and for which the respondents chose in 13.1% of the cases Audiovideo only, in 40% of the cases both Audio and Audiovideo, and the remaining 46.9% Audio only.
The asked distinction is anything but trivial as it has direct consequences both on the quality and on the size (and eventually format) of the resources.

In our survey we further asked the respondents to indicate of what type of resources they were in possession of. As much as 70.7% of the resources were mentioned to be of digital nature (e.g. *.wav, *.Mp3, *.eaf, *.TextGrid, *.txt etc.), 26.4% of analogue nature (tapes, forms etc.).

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

As far as the format of additional data available besides the speech resources is concerned, we categorized the 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* (n = 55) and *non-binary* (n = 85), while for 26 respondents no additional files are available. Under the first group, which we categorized as *binary* files the most common ones listed are *.doc (23.1%), *.pdf (4.7%). For the *non-binary* files the most listed are *.txt (23.1%), PRAAT's *.TextGrid (15.4%) and ELAN's *.eaf files (7.1%). Despite declaring the presence of additional files for the speech resources, 26 respondents did not specify any format.

3.4 Accessibility and maintenance issues

Almost half of the resources listed in our survey (48.5%) is barely accessible. Only 10% of the resources is accessible and available, 3.1% is partially accessible, 36.9% is available upon request, 0.8% is available upon request and only for selected parts (NA's = 0.8%).

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 53%) fall within the category which we defined as casual workers (e.g. workers without a permanent position nor a perma-

[3] The rest of the mentioned *Linguistic* subfields (categorized as *Other*) put together totals 19.
[4] Due to the possibility the respondents had to fill in the "*Other, please specify*" option when indicating the scientific domain or 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 reduce the sparseness of the data.

nent affiliation to an institution). Only 35.4% of the remaining respondents declared a permanent position and affiliation (for example to universities or other public institutions), while 6.2% did not provide any information (the remaining 5.4%, 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 relates to ethics and legal issues, which are addressed by the respondents only in 46.2% of the cases. This has unavoidable effects especially on the accessibility and reusability of such resources and represents something all the subjects involved in the creation and collection of future resources should be aware of.

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

An unexpectedly surprising result emerging from our survey is that only 31.5% 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. 64.6%, more than half of our respondents) who would agree in storing their archives and their speech resources in national repositories. 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 consciousness, 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 forthcoming annual conference of the Italian Speech Sciences association (AISV) to be held in February 2019, will be devoted to speech archives and the Executive Director of CLARIN will be giving a keynote lecture.

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.

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

Setting up the PORTULAN / CLARIN repository

Luís Gomes **Frederico Apolónia** **Ruben Branco** **João Silva** **António Branco**

NLX-Group, University of Lisbon, Portugal
{luis.gomes,frederico.apolonia,ruben.branco,jsilva,ahb}@di.fc.ul.pt

Abstract

This paper aims at sharing the lessons learned at setting up a CLARIN repository based on the META-SHARE software, which we have just used to develop the PORTULAN / CLARIN centre.

This paper documents the changes and extensions to META-SHARE that were needed to fulfil the CLARIN requirements for becoming a B-type centre.

The main purpose of this paper is to serve as a one-stop guide for teams pondering or having decided to adopt META-SHARE software for setting up their own CLARIN repositories in the future.

1 Introduction

In order to set up the CLARIN repository of the Portuguese network, PORTULAN / CLARIN, following other national networks, we decided to use META-SHARE software to set up our repository software, extending it with the functionalities required by CLARIN for B-centre approval.

These extensions address a number of aspects that we will be reporting on in this paper, namely: metadata harvesting (Section 2), single sign-on (Section 3), persistent identifiers (Section 4) and the user interface (Section 5).

META-SHARE is a network of repositories for sharing and exchanging language data and tools. It is also the name of the software[1] running on these repositories and we will refer to it as MS for short.

2 Metadata Harvesting

The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) is an application-independent protocol specification[2] for metadata harvesting. The OAI-PMH specification describes the communication between *repositories* and *harvesters*.

A repository must implement this protocol to be approved as a CLARIN B-centre. Furthermore, the metadata harvested from the repository must comply with a profile in the Component MetaData Infrastructure (CMDI) CLARIN registry.

A MS metadata profile already exists in the CLARIN CMDI registry and an implementation of the OAI-PMH protocol was kindly provided to us by the executive staff of CELR at University of Tartu.[3] We made minor tweaks to this implementation in order to make it work with our version of MS, which is slightly more recent (we are using the 3.1.1 branch instead of the 3.0 one). Then, we sent an email to <harvester@clarin.eu> announcing our OAI-PMH endpoint to the Virtual Language Observatory (VLO) and asking to be harvested by the alpha VLO instance.

After the harvesting was concluded we browsed the PORTULAN / CLARIN resource collection in the VLO alpha instance, and we checked the correctness and completeness of the metadata.

[1] https://github.com/metashare/META-SHARE
[2] http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm
[3] https://keeleressursid.ee/en/center/people

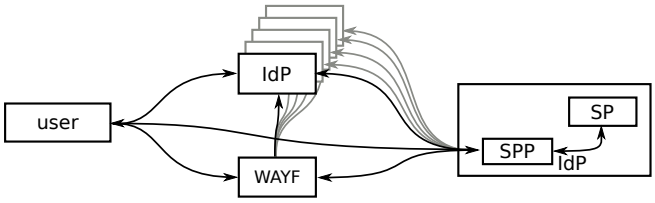


Figure 1: Using a Service Provider Proxy (SPP) to overcome PySAML2 limitations in terms of the number of IdPs it is able to handle.

3 Single Sign-On with SAML2

The Security Assertion Markup Language version 2 (SAML2) is an open standard XML-based framework that enables user authentication within a federation of web Service Providers (SPs) and Identity Providers (IdPs) through Single Sign-On (SSO). In practice, once users have authenticated themselves to a federated IdP, they will thereafter remain authenticated for any service provided by the federated SPs until they explicitly logout.

MS is implemented in Django and djangosaml2[4] integrates the PySAML2 library as a Django application and authentication backend, making it easy to integrate MS as a SAML2 SP. Unfortunately, the PySAML2 library is unable to cope with a large number of IdPs, and when we tried to load the metadata file of the CLARIN federation IdPs a "too many open files" exception was raised (apparently PySAML2 opens many auxiliary files to process a single large metadata XML). Thus, to work around this PySAML2 limitation, we set up a Service Provider Proxy (SPP) using SimpleSAMLphp[5].

As depicted in Figure 1, the proxy behaves as an IdP to our SP (the only IdP that the SP must known about) and behaves as a SP to the IdPs in the CLARIN federation. The local copy of the federation IdPs metadata XML is synced to the master metadata file[6] on a daily basis.

When a user tries to access a service provided by a federated SP, the SP needs to contact the IdP of the user's home institution/organization to authenticate the user and obtain identity attributes such as the user's name and email address.

The number of IdPs in a federation such as CLARIN may be over one thousand and thus a special service called *WAYF* or *Discovery Service* is needed to allow the users to *easily* select the IdP of their home institution from such a large list. Instead of simply presenting a list containing all IdPs in the federation, a good WAYF service will try to automatically detect the users' institution based on several parameters such as the IP address and geo-location of the computer. Also, the user's choice of IdP will be memorized by the WAYF service and recalled upon subsequent visits.

Since we have deployed SimpleSAMLphp as a SPP, we could also use its WAYF implementation. However, we have configured our SPP to redirect users to the CLARIN WAYF service[7] instead, which is based on DiscoJuice[8] and we find it to provide a more user friendly interface than the WAYF of SimpleSAMLphp.

4 PIDs

The use of Persistent Identifiers (PIDs) for resources is a requirement for B-centre repositories.

Fortunately, MS has a metadata field for storing externally-assigned identifiers such as the PID of a resource. Thus far, we have not automated the generation of PIDs for resources deposited in the repository, but we have plans for integrating MS with ePIC[9] through its RESTful API, allowing easy generation of a PID from the metadata editor interface.

[4] `https://pypi.org/project/djangosaml2/`
[5] `https://simplesamlphp.org/`
[6] `https://infra.clarin.eu/aai/prod_md_about_spf_idps.xml`
[7] `https://www.clarin.eu/content/clarin-central-discovery-service`
[8] `http://discojuice.org/`
[9] `http://www.pidconsortium.eu/`

5 Changes to the User Interface

MS project originally used the Blueprint CSS framework. However, since this framework was last updated on 14th of May 2011 and is missing a lot of modern website design features such as responsive design, we decided to change the MS interface to a more modern and robust alternative. As CLARIN has made available[10] an implementation of the interface guidelines as a Bootstrap theme, we decided to use this theme as the basis of our work on revitalizing the MS interface.

Bootstrap is easy to use, well documented and supports responsive design. This has made it very popular and is currently supported by a large community. Because the aforementioned CLARIN theme was built upon Bootstrap version 3, we are using that version instead of the newest version 4.

The CLARIN human interface guidelines[11] specify, among other things, the general page layout, the typography that should be used and the colour palette to be followed. With respect to the layout, MS already provides a very solid foundation but some changes are required to make PORTULAN / CLARIN comply with the guidelines.

The landing page of the repository allows searching and browsing the resources by means of a search box and a list of "Filter by" tags which implement a faceted search. Following the CLARIN guidelines that suggest secondary navigation to appear on the right-hand side of the page, these "Filter by" tags were moved to the right, as shown in Figure 2 (a).

The search box was sized down and moved to the upper right corner of the page (b).

In MS, the user had to hover the entry title to see a description of the resource, but we opted to make the description always visible, although clipped to a maximum of 300 characters, requiring less effort from the user to get to important information (c).

Furthermore, different resource and media types were conveyed to the user by adorning resource titles with different icons, which requires the user to hold a memory mapping between icons and their meaning. We replaced these icons with their textual meaning in a key-value structured presentation, which we find easier to assimilate, as shown in (d), (e) and (f) in the figure.

To minimize scrolling on today's wide screen monitors, we resized the page header to take up less vertical space, as shown in (g).

6 Conclusion

The repository of the PORTULAN / CLARIN centre is ready to enter production. We keep doing tests and rounding some sharp edges, but the main requisites for a B-centre repository seem to be ensured.

Ideally, MS should be ported to make use of more recent and supported versions of its dependencies, in particular Django, but the effort required is too high at this point. Therefore, to mitigate possible security flaws in MS and its outdated dependencies we decided to keep our repository running within a virtualized environment accessed through a reverse HTTP proxy. Despite the disadvantage of not being actively developed and supported, MS has the advantage (for us) of being implemented in Python and Django, which allows us to reuse knowledge of this language and framework in other components of the CLARIN / PORTULAN centre.

The next main step will be the CoreTrustSeal quality assessment procedure.

Acknowledgements

We thank the executive staff of CELR at University of Tartu[12], especially Neeme Kahusk, for providing us with their MS implementation of the OAI-PMH protocol and SAML2 configuration. We thank Esmeralda Pires from the FCCN (Portuguese Foundation for National Scientific Computation) for helping us with SAML2 questions.

[10] https://github.com/clarin-eric/base_style
[11] https://office.clarin.eu/v/CE-2016-0794-CLARINPLUS-D3_1.pdf
[12] https://keeleressursid.ee/en/center/people

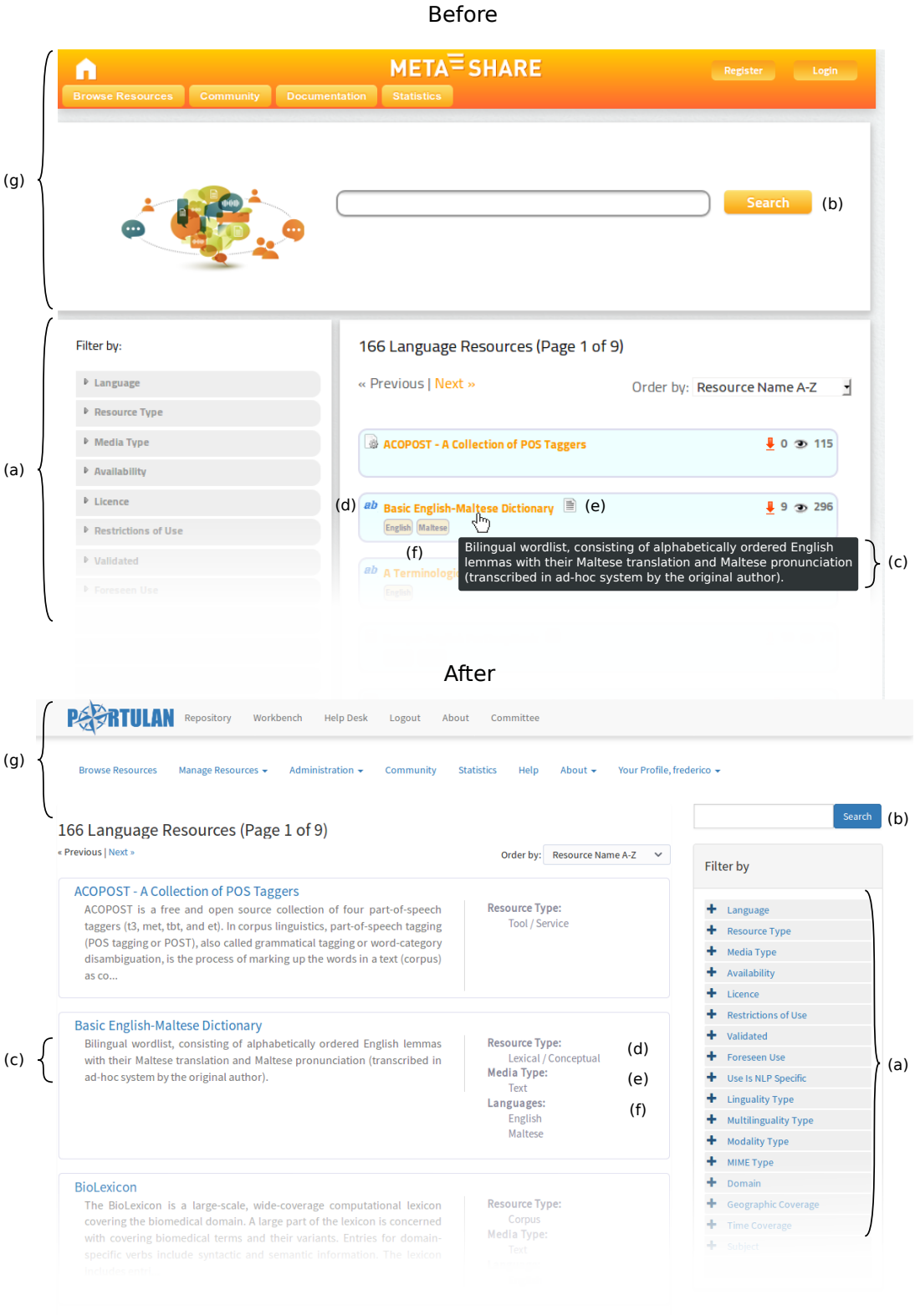


Figure 2: Repository search/browse page before and after our changes.

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 illustrate the usage of LaMachine 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.

1 Introduction

Software is a key deliverable and a vital component for research in projects such as those under the CLARIN umbrella. It is software that provides researchers the instruments to yield for their research; 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.

LaMachine incorporates software providing different types of interfaces[1] that typically address different audiences. Whilst we attempt to accommodate both technical[2] and less-technical audiences[3], 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* of LaMachine chosen, it 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 for production servers in for example CLARIN centres. We demonstrate how LaMachine can create a fully functioning and standalone research environment for text mining and NLP for Dutch texts in a use case project at the Healthcare Inspectorate.

2 Architecture

Being an open-source NLP software distribution, LaMachine is constrained to Unix-like platforms; this primarily means Linux, but also BSD and, with some restrictions, macOS. Cygwin[4] is not tested or supported. However, virtualisation technology enables deployment on a wider range of platforms, including Windows. The focus of the LaMachine distribution stands in contrast with mobile platforms (Android/iOS/etc), native Windows/mac desktop software, or certain interface types in general such as classical desktop GUI applications or mobile 'apps', all of which fall beyond our scope.

[1] Command line interfaces, programming interfaces, web-user interfaces, webservices.
[2] Data scientists, DevOps, system administrators, developers.
[3] The wider researcher community, particularly the Humanities; also educational settings.
[4] A unix environment on Windows

All software that is incorporated in LaMachine must 1) bear some relevance to NLP, 2) be under a recognised open-source license, 3) be deposited in a public version controlled repository[5] and 4) have a release protocol (with semantic versioning) using the proper technology-specific channels.

LaMachine is a *meta distribution* as it can be installed in various contexts. At its core, LaMachine consists of a set of machine-parsable instructions on how to obtain, build (e.g. compile from source), install and configure software. These are implemented using Ansible[6]. 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 given and only needs to know which repositories to use. Similarly, there are different programming-language-specific ecosystems providing their own repositories, such as the Python Package Index for Python, CRAN for R, CPAN for Perl, Maven Central for Java. LaMachine again relies on those to pull and install software from and never forks, archives, or modifies the software in any way. In doing so, we compel participating software projects to adhere to well-established distribution standards and ensure the software is more sustainable towards the future (van Gompel et al., 2016). Moreover, we ensure that LaMachine never becomes a prerequisite for the software but merely a courtesy or convenience.

LaMachine provides ample flexibility that allows it to be deployable in different contexts. First of all there is flexibility with regard to the target platform, where we support several major GNU/Linux distributions (Debian, Ubuntu, CentOS, RedHat Enterprise Linux, Fedora, Arch Linux), as well as macOS (although with more limitations). Second, there is flexibility with regard to the form, where we support *containerisation* through Docker[7], *virtualisation* through Vagrant and VirtualBox[8], direct remote provisioning through Ansible (for production servers), or an installation that is either global to the machine or local in a custom directory for a specific user (using `virtualenv`). Pre-built docker containers and virtual machine images with a limited selection of participating software are regularly uploaded to the Docker Hub and Vagrant Cloud, respectively. The different flavours all offer a different degree of seperation from the host OS, where Virtual Machines are completely virtualised, Docker Containers still share the kernel with the host OS, and the machine-specific installation flavour actually compiles against the machine's distribution itself and thus offers the least amount of overhead.

Installation of LaMachine begins with a single bootstrap command[9]. It can interactively query the users for their 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 customized by the user. The user may also opt for installing the latest releases, the more experimental development versions of the software, or specific custom versions (to facilitate scientific reproducibility). The bootstrap procedure detects and installs the necessary prerequisites automatically and eventually invokes Ansible to perform the bulk of the work. Figure 1 provides a schematic view.

LaMachine also aims to harmonise the metadata of all installed software, by converting metadata from upstream repositories, i.e. the repositories where tool providers deposit their software, to a common standard called CodeMeta [10] (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. This in turn enables other tools to do proper service discovery and provenance logging.

Leveraging this metadata, LaMachine comes with a webserver that offers a portal website with access and overview of all installed tools, including web services and web applications. It also comes with a Jupyter Lab[11] environment which provides a web-based Integrated Development 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 data science community nowadays.

[5] e.g. Github, Gitlab, Bitbucket, provided the repository is public
[6] `https://www.ansible.com`
[7] `https://www.docker.com`
[8] `https://vagrant.org`, `https://www.virtualbox.org`
[9] See `https://proycon.github.io/LaMachine`
[10] `https://codemeta.github.io/`, described in JSON-LD
[11] `https://jupyter.org/`

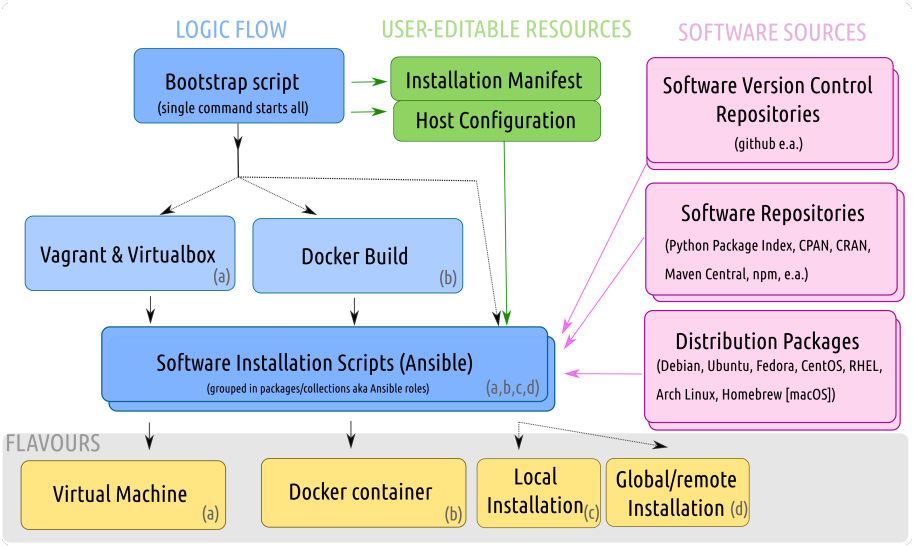


Figure 1: A schematic representation of the LaMachine architecture

3 Software

LaMachine exists since May 2015 and has been used extensively ever since by numerous users, in 2018 version 2 was released which was a significant rewrite. LaMachine was initially conceived as the primary means of distribution of the software stack developed at CLST, Radboud University Nijmegen. It therefore includes a lot of our software. A full list of included software goes beyond the scope of this overview; we will merely mention some CLARIN-NL/CLARIAH-funded tools: ucto (a tokeniser), Frog (an NLP suite for Dutch), FoLiA (Format for Linguistic Annotation, with assorted tools), FLAT (a web-based linguistic Annotation tool), PICCL (an OCR and post-OCR correction pipeline) and CLAM. However, this project is not limited to one research group and is open to participation by other software providers, especially those also in CLARIAH and the upcoming CLARIAH PLUS project. We already include some relevant software by other CLARIAH partners. Moreover, LaMachine incorporates a large number of renowned tools by external international parties, offering most notably a mature Python environment with renowned scientific modules such as scipy, numpy, scikit-learn, matplotlib, nltk, spacy, pytorch, keras, gensim, tensorflow, and many others, but also R, Java and tools such as Stanford CoreNLP and Kaldi.

4 Case study

We participated in a small Dutch national project titled *"Text mining for Inspection: an exploratory study on automatic analysis of health care complaints"* [12] led by IQhealthcare[13], 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[14]) 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.

[12] `https://bit.ly/2N2AICS`
[13] `http://www.iqhealthcare.nl/nl/`
[14] `https://www.landelijkmeldpuntzorg.nl`

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. 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.

5 Conclusion & Future work

The recent release of LaMachine v2, which constituted a full rewrite, has opened up LaMachine to outside contribution. Contributor documentation has been written, and at this stage, we greatly welcome external participants to join in. Use cases as the example in section 4 contribute to thorough testing and running of LaMachine in less ideal circumstances such as nested VM constructions and offline usage.

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.

Development of LaMachine presently takes place in collaboration with the CLARIAH WP3 Virtual Research Environment (VRE) project[15], which has higher ambitions in accommodating the researcher and connectivity of data and services, and transcends also those of the CLARIN Language Resource Switchboard (Zinn, 2016). An important part of our future focus will therefore be on interoperability with the higher-level tools emerging from the VRE efforts, but also with other parts of the CLARIN infrastructure; single-sign on authentication being a notable example here.

Acknowledgement

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 all project partners: the Dutch Health Inspectorate, IQhealthcare, and Tim Voets for their valuable contributions and help in the ZonMw project.

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 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*.

[15] `https://github.com/meertensinstituut/clariah-wp3-vre`

XML-TEI-URS: using a TEI format for annotated linguistic resources

Loïc Grobol
Lattice / ALMAnaCh
Paris, France
loic.grobol@ens.fr

Frédéric Landragin
Lattice
Paris, France
frederic.landragin@ens.fr

Serge Heiden
IHRIM
Lyon, France
slh@ens-lyon.fr

Abstract

This paper discusses XML-TEI-URS, a recently introduced TEI-compliant XML format for the annotation of referential phenomena in arbitrary corpora. We describe our experiments on using this format in different contexts, assess its perceived strengths and weaknesses, compare it with other similar efforts and suggest improvements to ease its use as a standard for the distribution of interoperable annotated linguistic resources.

1. Related works

XML-TEI-URS[1], introduced in (Grobol, Landragin, et al. 2017) is an annotation format inspired by the URS (Unit-Relation-Schema) metamodel developed for Glozz (Widlöcher and Mathet 2012) with a concrete serialization in TEI mark-up complying with the latest recommendations (TEI P5 v3.3.0). The original intent of this format was to provide a way to annotate reference phenomena, and particularly coreferences and anaphora, but it proved versatile enough for a larger class of annotations, as in (Grobol, Tellier, et al. 2018), where it is used for dependency syntax annotations. By design, it is not meant to be a ground-breaking new format, but rather a concrete realisation — within the limits of a standard serialization — of an abstract model proved to be sensible for coreference.

XML-TEI-URS is by no mean the first attempt at devising a general-purpose linguistic annotation format. There already exist several such formats, with wide range of uses, both in Corpus Linguistics and in Natural Language Processing, for example the tabular format used by BRAT (Stenetorp et al. 2012) or the XML-based formats used by GATE (Cunningham et al. 2013), MMAX2 (Müller and Strube 2006) or Glozz. But those formats are mostly tied to those specific annotation softwares — even when they express theoretically sound annotation models — and are susceptible to change along with their needs, with no guarantee of backward compatibility or notification of evolution. Consequently, they can only be thought of as *de facto* standards, whose use for perennial storage of linguistic resources could be problematic.

Conversely, as described in (Grobol, Landragin, et al. 2017), most of the annotated corpora for coreference use ad-hoc formats, that are usually well-suited to this single phenomenon, but do not support extension to other kind of annotations. The most common way to add other types of annotations (such as syntactic ones) in these resources is to use hybrid formats, such as in the tabular format used for the CoNLL-2012 corpus (Pradhan et al. 2012), which uses two different and incompatible types of parenthesized expressions for syntax and coreference annotations. One of the downsides of this approach is the data preparation overhead it imposes on the development NLP systems, a tedious and error-prone process, with scarce opportunities for reuse.

2. Experiments

Our experiments so far with XML-TEI-URS have been the following ones:

1. Porting the ANCOR corpus (Muzerelle et al. 2014) coreference annotations to XML-TEI-URS, first as a proof-of-concept for (Grobol, Landragin, et al. 2017).

2. Enriching ANCOR with syntactic annotations for (Grobol, Tellier, et al. 2018), which had us first convert it to a suitable input format for automatic parsers (Universal Dependencies CoNLL-U (Nivre et al. 2016)), then convert the resulting syntactic analysis back to XML-TEI-URS, with an additional annotation layer that describes the relations between syntax and coreference.

3. Integrating XML-TEI-URS in the URS annotation plugin of the TXM platform (Heiden 2010).

2.1. XML-TEI-URS for coreference: ANCOR

When it comes to file formats, ANCOR has a tumultuous story: it is composed of three different oral corpora, that were originally distributed in the native format of the transcription tool Transcriber (Barras et al. 1998). Its coreference annotations were then added in Glozz as if the data were raw texts, thus ignoring the existing XML structure (the consistency between the two layers was enforced manually), and finally integrated in it, using a non-standard ad-hoc format. This combination made the exploitation of this corpus tedious at best, and information-destructive in some cases (e.g. when entity mention crossed utterances borders). It was clear from the beginning of our work that there was a need for a better format — or at least one that was easier to use.

The initial conversion of ANCOR to XML-TEI-URS has actually been done at the same time as the definition of the format, which probably made the development of the necessary software tools more time-consuming than under other circumstances. Reflecting on that experience, we find that most of our difficulties came from the shortcomings of the original format, and from our efforts to enforce data consistency by correcting the errors that are inevitably present in any corpus of a significant size. All in all, the initial conversion took us no more than a few weeks, with some later refinements to meet unforeseen needs revealed by our actual use of the resulting corpus.

The resulting corpus is much easier to use than the original one, particularly thanks to the choice of completely stand-off annotations using a reference word segmentation (which are not mandatory for XML-TEI-URS, but heartily encouraged). The most welcomed advantage of this choice is that it allows to completely ignore the existence of annotations for preprocessing that does not take them into account (which is obviously harder with inline annotations) e.g. extracting the raw text of the corpus to run third-party tools on it is completely transparent.

The expressiveness of the TEI format also allowed annotations that were not possible in the original format, e.g. entity mentions spanning several utterances, or parallel and overlapping utterances.

2.2. XML-TEI-URS for syntax: ANCOR-AS

As stated in (Grobol, Tellier, et al. 2018), most automatic coreference detection systems use rich syntactic knowledge, which implies a need for corpora that hold both types of annotations. Existing corpora usually use one of three main strategies: use a ad-hoc hybrid format that incorporates the two types of annotations (as in CoNLL-2012), keep one version of the corpus for each type (as in the NER version of the French Treebank (Sagot et al. 2012)) or base one type on the other (as in the PCC, Polish Coreference Corpus (Ogrodniczuk et al. 2015)). In our context, none of these solutions was satisfactory. The most satisfactory would have been the option chosen for the PCC, but it requires mutually consistent annotation layers, which was not the case with automatic syntactic annotations.

Instead, we took advantage of the unobtrusive nature of standoff XML-TEI-URS annotations by totally ignoring existing coreference annotations at first when adding syntactic annotations, and only linking the two types of annotation in a third layer. The main obstacle in that process was that the word segmentation we used in the original version was not necessarily the same as the one given by the automatic parser. This issue was dealt with by adding correcting elements inspired by the then-current draft of (ISO 2017), that link between the surface forms of the raw corpus and the syntactic words used by the parser, in e.g. expansions (*du*→*de le* in French) and multi-word units.

Apart from this technical issue, the conversion between formats, from XML-TEI-URS to CoNLL-U and back was relatively straightforward, here again thanks to the use of reference to a word segmentation, for instance to clean up the parser inputs from easily detected disfluencies — thus improving its performances — while keeping them available in the raw text of the final resource.

That said, the final resource expresses the main drawback of the format: it is very heavy, far more than the corresponding CoNLL-U annotation, mostly because of our rather crude use of feature structure. Future versions of the resource will try to address this issue, most notably by a judicious use of MAF (ISO 2006) feature libraries, but a certain heaviness of TEI formats will always be unavoidable. In the meantime, we tried to mitigate this heaviness by keeping syntactic annotations in separated files, using the prefixed id facilities offered by the TEI to link them to the source files, which would have made sense in any case: since these syntactic annotations are not gold-standard, keeping them separated preserves the integrity of the manual coreference annotations of ANCOR.

2.3. Democrat and TXM platform

Since (Grobol, Landragin, et al. 2017), a progressive move towards using XML-TEI-URS as the format for the final version of the Democrat project corpus (Landragin 2016) is underway. In accordance, support for this format has been added to the URS annotation plugin developed in the context of this project for the TXM open-source platform. Full support for importing from and exporting to stand-off XML-TEI-URS is currently available, along with cross-corpus transfer of annotation between different corpora, as long as the tokens targeted by the annotations are present in both of them.

Integrating XML-TEI-URS to TXM was not too hard, thanks to the similarities with TXM internal format, which already used token-based stand-off annotations for lemmas and part-of-speech. The integration of XML-TEI-URS to TXM has been beneficial in reducing data duplication in version management, since several versions of a corpus can share the same outsourced annotation file as long as the token ids are constant. Conversely, several annotation sets can refer to the same corpus, which allows concurrent annotation by different annotators and keeping track of several versions of the same annotation set.

3. Conclusion and perspectives

Since its initial development, we have used XML-TEI-URS in several different contexts, and so far, it has lived up to our expectancies: it provides a standard, versatile and generally easy to use format for linguistic annotations. However, it comes with a certain heaviness that is probably linked to the TEI characteristics. This somewhat degrades human-reading experience for our prototype corpus, even though it had no impact on machine reading.

In the context of CLARIN, since the current guidelines advocate the use of TEI XML formats for textual data, we believe that XML-TEI-URS might serve as a basis — or at least an inspiration — for future linguistic resources with annotations going beyond the default set of attributes hardcoded into the current TEI guidelines (`lemma`, `pos` and friends). We are very much open to further developments or refinements of this format to better suit the needs of the community.

4. Acknowledgements

This work is part of the "Investissements d'Avenir" overseen by the French National Research Agency ANR-10-LABX-0083 (Labex EFL), and is also supported by the ANR DEMOCRAT (Describing and Modelling Reference Chains: Tools for Corpus Annotation and Automatic Processing) project ANR-15-CE38-0008.

References

[Barras et al. 1998] Claude Barras, Edouard Geoffrois, Zhibiao Wu, and Mark Liberman. 1998. Transcriber: a Free Tool for Segmenting, Labeling and Transcribing Speech. In *Proceedings of the First International Conference on Language Resources and Evaluation (LREC)*. European Language Resources Association (ELRA). Granada, España, May 1998.

[Cunningham et al. 2013] Hamish Cunningham, Valentin Tablan, Angus Roberts, and Kalina Bontcheva. 2013. Getting More Out of Biomedical Documents with GATE's Full Lifecycle Open Source Text Analytics. *PLOS Computational Biology*, 9.2, Feb. 2013: 1–16.

[Grobol, Landragin, et al. 2017] Loïc Grobol, Frédéric Landragin, and Serge Heiden. 2017. Interoperable annotation of (co)references in the Democrat project. In Harry Bunt, editor, *Thirteenth Joint ISO-ACL Workshop on Interoperable Semantic Annotation*. ACL Special Interest Group on Computational Se-

mantics (SIGSEM) and ISO TC 37/SC 4 (Language Resources) WG 2. Montpellier, France, Sept. 2017.

[Grobol, Tellier, et al. 2018] Loïc Grobol, Isabelle Tellier, Éric De La Clergerie, Marco Dinarelli, and Frédéric Landragin. 2018. ANCOR-AS: Enriching the ANCOR Corpus with Syntactic Annotations. In *LREC 2018 - 11th edition of the Language Resources and Evaluation Conference*. Miyazaki, Japan, May 2018.

[Heiden 2010] Serge Heiden. 2010. The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. In *24th Pacific Asia Conference on Language, Information and Computation (PACLIC)*, pages 389–398. Institute for Digital Enhancement of Cognitive Development, Waseda University. Sendai, Japan, Nov. 2010.

[ISO 2006] ISO/TC 37/SC 4. 2006. *ISO 24610-1:2006 Language resource management – Feature structures – Part 1: Feature structure representation*. Reference. Geneva, CH: International Organization for Standardization, Apr. 2006.

[ISO 2017] ISO/TC 37/SC 4/WG 2. 2017. *ISO AWI 24617-9 Language resource management – Part 9 Reference Annotation Framework (RAF)*. Reference. Geneva, CH: International Organization for Standardization, 2017.

[Landragin 2016] Frédéric Landragin. 2016. Description, modélisation et détection automatique des chaînes de référence (DEMOCRAT). *Bulletin de l'AFIA*, 92, 2016: 11–15.

[Müller and Strube 2006] Christoph Müller and Michael Strube. 2006. Multi-level annotation of linguistic data with MMAX2. In Sabine Braun, Kurt Kohn, and Joybrato Mukherjee, editors, *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, pages 197–214. Peter Lang. Frankfurt a.M., Germany, 2006.

[Muzerelle et al. 2014] Judith Muzerelle et al. 2014. ANCOR Centre, a Large Free Spoken French Coreference Corpus: Description of the Resource and Reliability Measures. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA). Reykjavík, Ísland, May 2014.

[Nivre et al. 2016] Joakim Nivre et al. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In Nicoletta Calzolari et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA). Portorož, Slovenia, May 23–28, 2016.

[Ogrodniczuk et al. 2015] Maciej Ogrodniczuk, Katarzyna Głowińska, Mateusz Kopeć, Agata Savary, and Magdalena Zawisławska. 2015. *Coreference in Polish: Annotation, Resolution and Evaluation*. Walter De Gruyter, 2015.

[Pradhan et al. 2012] Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes. In *Proceedings of the joint Conference on Empirical Methods in Natural Language Processing (EMNLP) and Computational Natural Language Learning (CoNLL)*, pages 1–40. CoNLL '12. Association for Computational Linguistics. Jeju, Korea, 2012.

[Sagot et al. 2012] Benoît Sagot, Marion Richard, and Rosa Stern. 2012. Annotation référentielle du Corpus Arboré de Paris 7 en entités nommées. In Georges Antoniadis, Hervé Blanchon, and Gilles Sérasset, editors, *Traitement Automatique des Langues Naturelles (TALN)*. Volume 2 - TALN. Actes de la conférence conjointe JEP-TALN-RECITAL 2012. Grenoble, France, June 2012.

[Stenetorp et al. 2012] Pontus Stenetorp et al. 2012. BRAT: a Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107. Association for Computational Linguistics. Avignon, France, Apr. 2012.

[TEI P5 v3.3.0] TEI consortium, editor. 2018. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Version 3.3.0. TEI P5: Guidelines for Electronic Text Encoding and Interchange. Jan. 31, 2018. URL: `http://www.tei-c.org/Guidelines/P5` (visited on 04/24/2018).

[Widlöcher and Mathet 2012] Antoine Widlöcher and Yann Mathet. 2012. The Glozz Platform: A Corpus Annotation and Mining Tool. In *Proceedings of the 2012 ACM Symposium on Document Engineering*, pages 171–180. DocEng '12. ACM. Paris, France, 2012.

Visible Vowels: a Tool for the Visualization of Vowel Variation

Wilbert Heeringa
Fryske Akademy
Ljouwert, The Netherlands
wheeringa@fryske-akademy.nl

Hans Van de Velde
Fryske Akademy
Ljouwert, The Netherlands
HVandeVelde@fryske-akademy.nl

Abstract

Visible Vowels is a web app for the analysis and visualization of acoustic vowel measurements: f0, formants and duration. The app is a useful instrument for research in linguistics. The app combines user friendliness with maximum functionality and flexibility, using a live plot view.

1 Introduction

Researchers in phonetics, sociolinguistics, dialectology, forensic linguistics, speech pathology, language acquisition, pyscholinguistics and neurolinguistics study the acoustic characteristics of vowels, measuring vowel formants, fundamental frequency, duration and other acoustic variables. Next, the measurements are visualized by graphs in order to find patterns that reflect particular external (e.g., region, age, gender, language background, pathological vs. non-pathological) or internal (e.g., word stress, following segment) factors. While programs like Praat, Speech Filing System (SFS) and WaveSurfer are frequently used for making acoustic measurements, most researchers use programs like PLOTNIK, NORM and VOIS3D or the R packages `vowels` and `phonR` for the visualization of these measurements. However, these packages do not meet the demands of most researchers in the field, due to a lack of features, flexibility and user-friendliness.

PLOTNIK was developed by Labov (2011) for the study of vowel variation and change in North American English and plots the vowels in F1/F2 space. Any selection of vowels can be visualized. Subsets of vowels can be highlighted within the larger set. Procedures for speaker normalization of formant frequencies are available. Means, standard deviations and the significance of difference between means can be calculated. Vowel systems of different speakers can be compared, by overlaying their vowel plots. PLOTNIK runs only on Macintosh operating systems.

The R package `vowels` includes "procedures for the manipulation, normalization, and plotting of phonetic and sociophonetic vowel formant data." (Kendall and Thomas, 2015). This package is the backend for the vowel normalization and plotting suite NORM. The NORM web application offers the possibility to run a number of normalization techniques on formant data and to quickly compare the results. The web application NORM runs on any platform and does not require any knowledge of the programming language R. However, NORM is less flexible than the `vowels` package, therefore, the authors of NORM encourage users to use their R package `vowels` rather than using NORM.

VOIS3D offers the possibility to normalize formant frequencies and duration. Additionally, it provides an analytic geometric solution for assessment of spectral overlap. Normalized scatter for two vowel distributions is modeled as two best-fit ellipses oriented at angles with respect to the F1 and F2 axes. The output of the metric is an overlap fraction, which represents the region shared by both best-fit ellipses. This tool runs only on Windows operating systems (Wassink, 2006). An important limitation of both PLOTNIK and NORM is that they only offer visualization of formants, not of duration and f0 of vowels. VOIS3D, however, can visualize vowel duration variation.

A package related to the `vowels` package is the `phonR` package (McCloy, 2016) that can be used to visualize trajectories with an unlimited number of measure points. Additionally, IPA glyphs, ellipses

showing degree of confidence in the location of the mean of each vowel/group and convex hulls representing the outline of the vowel space can be added to the vowel plot. The degree of encroachment or overlap between vowel categories can be calculated and plotted by means of a heat map.

In Section 2 we present Visible Vowels, a web application that aims to combine user friendliness with maximum flexibility and functionality. The app is a useful instrument for research in phonetics, sociolinguistics, dialectology, forensic linguistics, and speech-language pathology. Visible Vowels includes a large deal of the functionality of the aforementioned packages, but offers also new functionalities such as the measurement of long-term formants and comparison of speakers with Huckvale's ACCDIST metric (Huckvale, 2004). Different from earlier vowel plot programs, Visible Vowels uses a live view, i.e. each time the user changes something in the settings, the plot shown in the viewer is immediately adjusted accordingly. This makes the comparison of, for example, different normalization techniques extremely easy. Being web-based, Visible Vowel runs on any platform. Future lines of development are presented in Section 3.

2 Program

The web app Visible Vowels is implemented in the programming language R (R Core Team, 2017). For drawing graphs functions from the packages `ggplot2` (Wickham, 2009), `plot3D` (Soetaert, 2017) and `ggdendro` (De Vries and Ripley, 2016) are used. The user interface has been built using the packages `shiny` (Chang et al., 2017) and `shinyBS` (Bailey, 2015). Visible Vowels is freely available at `https://visiblevowels.org`. The app is tested in Firefox, Google Chrome, Opera, Edge and Safari. A standalone version can be used by installing the package `visvow` in R. The app consists of seven tab panels: 'Load file', 'Contours', 'Formants', 'Dynamics', 'Duration', 'Explore' and 'Help'. In the panels 'f0', 'Formants', 'Duration' and 'Explore' the user can set the size and font of the axis ticks and labels and the size of the graph as a whole. Data files can be downloaded either as tab-delimited text file or as Office Open XML file. Graphs can be saved in five different formats (SVG, PDF, EPS, JPG, PNG).

2.1 Load file

In this panel, a spreadsheet file that contains the measurements is loaded. Spreadsheets are usally made in Microsoft Excel, therefore a file with Excel's default file extension ('.xlsx') can be used without the need to export this file to a text file. The spreadsheet should contain the speaker labels, vowel labels, categorical variables (for example, region, gender, etc.) and duration. Next a set of five variables –'time', 'f0', 'F1', 'F2' and 'F3'– may be repeated as *many times* as the user wishes. More explanation about the table format is provided in the 'Help' panel. In this panel an example spreadsheet is provided as well.

2.2 Contours

In this panel the contours of f0, F1, F2 or F3 can be visualized using the time points selected by the user.

Multiple line graphs can be generated, where each contour represents a category of a categorical variable. A plot can also be divided in panels, where each panel shows a graph with a contour for a value of a categorical variable. When multiple line graphs are combined with multiple panels, it is, for example, possible to compare contours of vowels [i] and [a] for older and younger speakers, having a panel for vowel [i] and a panel for vowel [a], and with two contours in each panel, for each age group one contour. Hertz (Hz) values can be converted to bark, ERB, ln, mel and ST values and exported to a data file.

2.3 Formants

In this panel vowels are plotted either in two- or three-dimensional space, using combinations of F1, F2 and F3. It is possible to average categories of the categorical variables. For example, for 10 vowels, three regions and two genders the plot will show $10 \times 3 \times 2 = 60$ plot symbols in three different colors (region in the example) and two different shapes (gender). Additionally, long-term formants can be shown: per speaker formant values are averaged over the vowels, and the points in the graph represent speakers instead of vowels. When multiple five-column sets are included in the input table, vowel trajectories can

be drawn. Given *n* sets, the user can include all time points or any subset of points, with a minimum of two subsets.

Hz values can be converted to bark, ERB, log and mel. There are 14 vowel normalization methods available that are divided in four groups: formant-ratio normalization, range normalization, centroid normalization and log-mean normalization. We added a new normalization method 'Heeringa & Van de Velde 2018' that is similar to the normalization method of Watt and Fabricius (2002), but it uses all the points the convex hull is made up of. Some methods require f0 and/or F3 measurements. The normalization methods can be used on the basis of any scale (Hz, bark, etc.) except for the methods that use a log transformation, which can only be applied to Hz values.

In 2D graphs the different colors of the plot symbols represent the categories of one categorical variable, and the shapes represent the categories of another categorical variable. Shapes can be circles, triangles, squares etc, but when 'vowel' is chosen, the vowel labels are shown instead. It is also possible to show multiple panels, each graph corresponding with a value of a categorical variable. Thus vowel systems of, for example, different regions can easily be compared.

Groups as defined by categorical variables can be made more visible by drawing spokes between each vowel and the gravity center of the group to which it belongs. Ellipses showing the degree of confidence in the location of the mean of each group can be shown for any confidence level. Convex hulls can be drawn in order to delineate the outline of the vowel space per category of a categorical variable. By drawing a convex hull around the vowel space, the effects of the different normalization methods can be compared due to the live plot view.

In 3D graphs vertical spikes can be added in order to see more clearly the *x*,*y* location of the points and to enhance the three-dimensional effect. The user can change the angles with respect to respectively the *x*-axis and the *z*-axis in order to view the graph from different perspectives.

2.4 Dynamics

Vowels measurements change over time during the vowel segment. In order to measure vowel dynamics, we implemented two methods: trajectory length and spectral rate of change. Fox and Jacewicz (2009) used these methods on the basis of F1 and F2 measurements. In Visible Vowels the user can choose any subset of f0, F1, F2 and F3, including the individual variables or all of the variables.

Using either multiple line graphs or grouped bar graphs it is possible to visualize vowel dynamic measurements for different values of a categorical variable, where each line in a line graph or 'sub bar' in a grouped bar graph represents a category. In a grouped bar graph, per category of the first categorical variable, bars are shown for all categories of the second categorical variable. Additionally, multiple panels can be shown, where each panel corresponds with a value of a categorical variable.

2.5 Duration

Duration can be visualized in the same way as vowel dynamics, using multiple line graphs or grouped bar graphs and panels as well. Duration measurements can be normalized by means of Lobanov's *z*-score transformation (Wang and Van Heuven, 2006).

2.6 Explore

In this panel distances between speakers on the basis of their vowel systems can be calculated using the ACCDIST metric of Huckvale (2004) which we extended by offering the possibility to include F3 (in addition to F1 and F2), and to compare groups of speakers that are defined according to one or more categorical variables. The results are visualized by either dendrograms or multidimensional scaling plots. The user can choose from five cluster algorithms and four multidimensional scaling routines.

2.7 Help

The help panel provides an 'about' section, describes the format of the input file and provides an example spreadsheet. A document is provided that explains how methods for averaging and measuring long-term formants, scale conversion, speaker normalization of formants, speaker normalization of duration,

measuring vowel dynamics and the ACCDIST metric are implemented in Visible Vowels. Additionally, in the help panel references for the R packages that are used in the program are provided.

3 Future work

Currently, Visible Vowels is available as a web app and a standalone version. In the latter version we will implement clickable maps, in which a user can click on a vowel in the F1/F2 space and hear the recording. We welcome other suggestions for extensions and further improvements. Visible Vowels will be included in the CLARIAH Virtual Research Environment (VRE). The CLARIAH VRE for linguistics will be hosted at the Meertens Institute/Humanities Cluster in Amsterdam and will integrate selected tools and services for linguistic research developed and supported in the Dutch CLARIN NL and CLARIAH projects. Finally, we plan to develop 'Visible Consonants, a similar tool for consonants.

Acknowledgements

We thank Vincent van Heuven and several other users of Visible Vowels for their valuable suggestions which we were happy to implement.

References

Eric Bailey. 2015. Package 'shinyBS': Twitter Bootstrap Components for Shiny. version 0.61.

Winston Chang, Joe Cheng, J.J. Allaire, Yihui Xie, and Jonathan McPherson. 2017. Package 'shiny': Web Application Framework for R. version 1.0.1.

A. De Vries and B. D. Ripley. 2016. ggdendro: Create Dendrograms and Tree Diagrams Using 'ggplot2'. https://CRAN.R-project.org/package=ggdendro. R package version 0.1-20.

R. A. Fox and E. Jacewicz. 2009. Cross-dialectal variation in formant dynamics of American English vowels. *Journal of the Acoustical Society of America*, 126(5):2603–2618. https://doi.org/10.1121/1.3212921.

M. Huckvale. 2004. ACCDIST: a Metric for Comparing Speakers' Accents. In *Proceedings of the International Conference on Spoken Language Processing, Jeju, Korea, Oct 2004*, pages 29–32.

Tyler Kendall and Erik R. Thomas. 2015. Package 'vowels': Vowel Manipulation, Normalization, and Plotting. version 1.2-1.

W. Labov, 2002. *A PLOTNIK tutorial*.

W. Labov. 2011. Plotnik. https://www.ling.upenn.edu/ wlabov/Plotnik.html. version 10.

Daniel R. McCloy. 2016. phonr: Tools for Phoneticians and Phonologists. https://cran.r-project.org/web/packages/phonR/index.html. R package version 1.0-7.

R Core Team. 2017. R: A language and environment for statistical computing. URL: `http://www.R-project.org/`.

K. Soetaert. 2017. plot3d: Plotting Multi-Dimensional Data. https://CRAN.R-project.org/package=plot3D. R package version 1.1.1.

H. Wang and V. J. Van Heuven. 2006. Acoustical analysis of English vowels produced by Chinese, Dutch and American speakers. In Jeroen van de Weijer and Bettelou Los, editors, *Linguistics in the Netherlands 2006*, volume 23, pages 237–248. https://doi.org/10.1075/avt.23.23wan. John Benjamins, Amsterdam.

A.B. Wassink. 2006. A geometric representation of spectral and temporal vowel features: Quantification of vowel overlap in three varieties. *Journal of the Acoustical Society of America*, 119(4):2334–2350. https://doi.org/10.1121/1.2168414.

D. J. L. Watt and A. H. Fabricius. 2002. Evaluation of a technique for improving the mapping of multiple speakers. *Leeds Working Papers in Linguistics and Phonetics*, 9:159–173.

H. Wickham. 2009. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, New York.

ELEXIS - European lexicographic infrastructure

Miloš Jakubíček
Lexical Computing
Brno, Czech Republic
milos.jakubicek@sketchengine.eu

Iztok Kosem
Jozef Stefan Institute
Ljubljana, Slovenia
iztok.kosem@ijs.si

Simon Krek
Jozef Stefan Institute
Ljubljana, Slovenia
simon.krek@ijs.si

Sussi Olsen
University of Copenhagen
Denmark
saolsen@hum.ku.dk

Lene Offersgaard
University of Copenhagen
Denmark
leneo@hum.ku.dk

Bolette Sandford Pedersen
University of Copenhagen
Denmark
bspedersen@hum.ku.dk

Abstract

This paper describes the establishing ELEXIS lexicographic infrastructure, a research infrastructure financed by the European Union through the H2020 funding scheme. We present the project as a whole in terms of its target audience and stakeholders as well as its key parts. We outline the components of the infrastructure, both those already implemented and those to be implemented in the course of the project (2018-2022). Close collaboration with CLARIN is supported by the Integration and Sustainability Committee.

1 Introduction

In 2013, the European lexicographic community was brought together in the European Network of e-Lexicography (ENeL) COST action.[1] This initiative was set up to improve the access for the general public to scholarly dictionaries and make them more widely known to a larger audience. In the context of this network, a clear need has emerged for a broader and more systematic exchange of expertise, for the establishment of common standards and solutions for the development and integration of lexicographical resources, and for broadening the scope of application of these high quality resources to a larger community, including the Semantic Web, artificial intelligence, NLP and digital humanities. At the end of the COST action, the initiative had been successfully transformed into a H2020 infrastructure project – European Lexicographic Infrastructure (ELEXIS).[2]

The objectives emphasised in ELEXIS are the following: the infrastructure will (1) foster cooperation and knowledge exchange between different research communities in lexicography in order to bridge the gap between lesser-resourced languages and those with advanced e-lexicographic experience; (2) establish common standards and solutions for the development of lexicographic resources; (3) develop strategies, tools and standards for extracting, structuring and linking of lexicographic resources; (4) enable access to standards, methods, lexicographic data and tools for

[1] www.elexicography.eu
[2] http://www.elex.is

scientific communities, industries and other stakeholders; (5) promote an open access culture in lexicography, in line with the European Commission recommendation on access to and preservation of scientific information. This paper focuses on objective 4 in particular as the activities related to these objectives have already begun.

2 Virtual access: Sketch Engine and Lexonomy

First parts of the infrastructure in terms of virtual access provision have already been implemented by providing access to Sketch Engine, a leading corpus management system [1], and Lexonomy, a simple web-based dictionary writing system. The access has been established as of April 1st, 2018 through the single-sign-on facilities of the eduGAIN network of academic identity federations. Thereby anybody academic with an institution affiliated in eduGAIN can gain access to the services without the need of further registration.

During the first month of availability, more than 100 European universities have arranged for access and delegated a responsible local administrator that manages a local quota of 1 billion words that users have available for corpus building. Besides that, they gain access to over 500 preloaded corpora hosted at https://sketchengine.eu as well as all analytic functions of Sketch Engine, such as concordancing, building wordlists, compiling word sketches, thesauri or automatic dictionary drafting.

The last is loosely connected to the Lexonomy dictionary writing system, allowing for bidirectional information connectivity between Sketch Engine and Lexonomy: in the push model, Sketch Engine exports a fully automatically created dictionary draft into Lexonomy for post-editing, whereas in the pull model, an editor retrieves additional information (such as corpus examples or collocation candidates) from Sketch Engine from within the Lexonomy interfaces.

3 Trans-national access

Even though lexicography has a long history of international research conferences, it has traditionally been a research area with limited knowledge exchange outside of each lexicographical institution, and in many cases lexicographic data has only been accessible to researchers from the institution who created the data and held the copyright. This is partly due to the commercial aspect of lexicography, lexicographical data used to be good business, but also to the fact that easy accessibility to restricted data requires that an effort is put into facilitating and controlling the availability of and access to the data, which again requires time and money not easily found in the budgets of the lexicographic projects.

An important objective of the ELEXIS project is therefore to stimulate knowledge exchange between lexicographical research facilities, infrastructures and resources throughout Europe and thus benefit mutually from the vast experience and expertise that exist in the community.

Inspired by other EU projects, e.g. EHRI[3], RISIS[4], InGRID[5], and sobigdata[6], ELEXIS has chosen to help overcome this gap by offering trans-national access activities in form of visiting grants that enable researchers, research groups and lexicographers to work with lexicographical data which are not fully accessible online or where professional on the spot expertise is needed in order to ensure and optimise mutual knowledge exchange, and to gain knowledge and expertise by working with lexicographers and experts in NLP and artificial intelligence.

We expect that the trans-national access activities will have a long-term impact specifically but not only for lesser-resourced languages and will boost the network and infrastructure of the European lexicographic community and facilitate future collaboration and knowledge exchange.

The objectives of the ELEXIS trans-national activities are listed below:

[3] https://ehri-project.eu/ehri-fellowship-call-2016-2018
[4] http://datasets.risis.eu/
[5] http://www.inclusivegrowth.eu/visiting-grants
[6] http://www.sobigdata.eu/access/transnational

- to offer opportunities to researchers or research teams to access research facilities with an excellent combination of advanced technology and expertise
- to support training of new specialists in the field of e-lexicography in order to conduct high-quality research and ensure sustainability of infrastructure
- to ensure support for excellent scholarly research projects and innovative enterprises and also support the complex multi-disciplinary research
- to encourage the integrative use of technology and methodologies
- to improve the overall services available to the research community
- to exchange knowledge and experience and to work towards future common projects and objectives
- to create an interdisciplinary community, collaborating on activities that are fully or partially of relevance to the proposed work.
- to create knowledge at the interaction between academia and societal knowledge

The transnational activities of ELEXIS consist of visiting grants of 1 to 3 weeks for researchers to experiment with and work on lexicographical data in a context of mutual knowledge exchange with the hosting institutions. Five visiting grants will be launched twice a year during the entire project period amounting to 35-40 grants in total. Researchers and lexicographers within the EU member states and associated countries will be invited to apply for free-of-charge access to and support of one of the providers of lexicographical infrastructures. These 11 infrastructures/lexicographical institutions will receive visiting researchers:

- ELEXIS-SL: Institut Jozef Stefan (JSI, Slovenia)
- ELEXIS-NL: Institute for Dutch Language (INT, The Netherlands)
- ELEXIS-AT: Austrian Academy of Sciences (OEAW,Austria)
- ELEXIS-RS: Belgrade Center for Digital Humanities (BCDH, Serbia)
- ELEXIS-BG: Institute of Bulgarian Lanugage Lyubomir Andreychin (IBL, Bulgaria)
- ELEXIS-HU: Hungarian Academy of Sciences (RILMTA, Hungary)
- ELEXIS-IL: K-Dictionaries (KD, Israel)
- ELEXIS-DK: Det Danske Sprog- og Litteraturselskab, University of Copenhagen (DSL/UCPH, Denmark)
- ELEXIS-DE: Trier Center for Digital Humanities (TCDH, Germany)
- ELEXIS-EE: Institute for Estonian Language (EKI, Estonia)
- ELEXIS-ES: Real Academia Española (RAE, Spain)

During the grant visits, the hosting institutions will provide support in terms of lexicographical and IT man power expertise. The calls for applications will include descriptions of the infrastructures and the lexicographical resources, tools, and expertise that are made available. Researchers and lexicographers interested in visiting a particular infrastructure should make a motivated application describing their background, the purpose of the visit, etc.

The trans-national activities represent a way of ELEXIS to enable access to restricted data not formerly available outside of the hosting institutions to researchers from other institutions and countries. However, the results of research conducted in the trans-national activities will be available under open access licences enabling the international lexicography community to become acquainted with previously inaccessible resources.

The first call for visiting grants was launched June 2018, deadline August 6, and received 10 applications of which 5 were accepted. The planning of the first visits has just been initiated.

4 Collaboration with CLARIN

Although ELEXIS already consists of a large group of partners and, by offering grants for transnational research visits, puts a great deal of effort sharing knowledge of lexicographic resources broadly, it also has as a main objective to collaborate with the infrastructures CLARIN and DARIAH. In ELEXIS, a dedicated work package focuses on integration and sustainability by means of an "Integration and Sustainability Committee (ISC)", which has already been formed and is meeting face-to-face as a co-located event of the CLARIN Annual Conference this year.

The ISC counts 13 members: the Executive Director of CLARIN, the Director of DARIAH, members involved in both national CLARIN consortia and ELEXIS, work package leaders from ELEXIS, and representatives appointed by CLARIN.

The committee is currently working on a common ground document to achieve a common understanding of the initiatives they will work on and of other ways to facilitate collaboration.

This strong commitment from ELEXIS to ensure sustainability and collaboration with other infrastructures right from the beginning of the project, will hopefully lead to the best possible opportunities for building on expertise, services and repositories already established in CLARIN and DARIAH. Possibilities for collaboration lie in working on standards, tools, and sharing of resources. Reuse of already established technical services such as federated identity systems, metadata sharing and repositories is also expected.

5 Conclusion

The ELEXIS infrastructure aims at creating a sustainable European lexicographic environment, building upon existing tools and methodologies, while combining and improving them beyond the state-of-the-art. Within the coming four years, it will foster cooperation between researchers in lexicography and natural language processing, and provide easy-to-access services to academic audience across the European Union.

Reference

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, Vít Suchomel. The Sketch Engine: ten years on. Lexicography, 1: 7-36, 2014.

Měchura, M. B. (2017) 'Introducing Lexonomy: an open-source dictionary writing and publishing system' in Electronic Lexicography in the 21st Century: Lexicography from Scratch. Proceedings of the eLex 2017 conference, 19-21 September 2017, Leiden, The Netherlands.

Sustaining the Dictionary of Southern Dutch Dialects (DSDD): a case study for CLARIN and DARIAH.

Jacques Van Keymeulen
Ghent University, Belgium
Jacques.Vankeymeulen@ugent.be

Sally Chambers
Ghent University, Belgium
Sally.Chambers@ugent.be

Veronique De Tier
Ghent University, Belgium
Veronique.Detier@ugent.be

Jesse de Does
The Dutch Language Institute (INT), Netherlands
Jesse.Dedoes@ivdnt.org

Katrien Depuydt
The Dutch Language Institute (INT), Netherlands
Katrien.Depuydt@ivdnt.org

Tanneke Schoonheim
The Dutch Language Institute (INT), Netherlands
Tanneke.Schoonheim@ivdnt.org

Roxane Vandenberghe
Ghent University, Belgium
Roxane.Vandenberghe@ugent.be

Lien Hellebaut
Ghent University, Belgium
Lien.Hellebaut@ugent.be

Abstract

In this paper, we report on an ongoing project, the Dictionary of the Southern Dutch Dialects (DSDD), funded by the Research Foundation Flanders (FWO). The DSDD is based on three dictionaries of the Flemish, Brabantic and Limburgian dialects. The project aims to aggregate and standardise the three comprehensive dialect lexicographic databases into one integrated dataset. The project, which started in January 2017, is organised in three phases: i) design and preparation, ii) implementation and iii) exploitation. The Ghent University DSDD team (Department of Dutch Linguistics/Ghent Centre for Digital Humanities) works closely together with the Dutch Language Institute (INT) who are responsible for the technical development and sustainability of the DSDD linguistic data infrastructure. During the project period (2017-2020), 3-4 research use cases will be developed to test the applicability of the newly aggregated DSDD for digital scholarship. At a later stage, the DSDD database can be linked with other dialect data in Belgium and the Netherlands. Within Flanders, work is underway to strengthen collaboration between DARIAH and CLARIN. As the DSDD is already working with both infrastructures, the DSDD is in a unique situation to benefit from CLARIAH.

1 Introduction

In 2016, the Research Foundation Flanders (FWO) funded a medium-scale research infrastructure project, the Database of Southern Dutch Dialects (DSDD)[1]. The aim of the DSDD project is to aggregate and standardise three comprehensive lexicographic dialect databases of the Brabantic[2], the Limburgian[3] and the Flemish Dialects[4] into one integrated dataset for the Southern Dutch Dialect area. The DSDD will contain several million lexicographic tokens, which will be made available via a Virtual Research Environment for digital lexicographical research, enabling new research questions,

[1] Database of Southern Dutch Dialects (DSDD): http://www.ghentcdh.ugent.be/projects/database-southern-dutch-dialects-dsdd.
[2] Dictionary of the Brabantic Dialects: https://e-wbd.nl.
[3] Dictionary of the Limburgian Dialects: http://e-wld.nl.
[4] Dictionary of the Flemish Dialects: https://e-wvd.be.

particularly in the fields of geo-visualisation, qualitative lexicology and dialectometry, to be answered. Furthermore, DSDD is a pilot project of DARIAH-BE and is also interoperable with CLARIN data standards[5].

2 History

The Flemish, Brabantic and Limburgian dialect databases were used to compile three dialect dictionaries (WBD, WLD, WVD)[6] which describe the vocabulary of the traditional dialects in the southern part of the Dutch language area (see Figure 1).

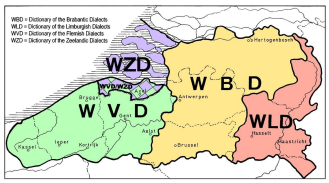


Figure 1: Dialect regions in the Southern Dutch Language Area

The vocabulary of these dialects was compiled as part of an international, inter-university project. The three dictionaries are arranged by onomasiological concept, rather than as an alphabetical listing as in a semasiological dictionary. They are therefore to be considered as geographically-orientated inventories of word usage, rather than as dictionaries proper. Completed between 2005 and 2018[7], the dictionaries were set-up in parallel, according to the same approach conceived by Weijnen (Nijmegen) [8], to enable the data to be aggregated. However, this original objective, until the start of DSDD, has not yet been achieved. To that effect, a consortium of linguists, computer scientists, digital humanities experts and geographers was established in 2016, to set-up the DSDD project.

3 Methodology

The DSDD project is organised in three phases: i) design and preparation, ii) implementation and iii) exploitation. As part of the first phase, the majority of the work focuses on data preparation; in particular selecting the concepts for standardisation and alignment.

3.1 Concept selection and standardisation

Despite the aim to develop the dictionaries according to the same approach, as they were compiled over a period of several years, at different places (Nijmegen, Leuven, Ghent) and by different editors, a large number of inconsistencies arose over time (Van Keymeulen, 2004). In order to create the aggregated DSDD database, standardisation activities had to be carried out before the selected concepts could be aligned. To investigate these problems, 1500 concepts from different thematic volumes of the dictionaries were selected to make an inventory of the issues to be resolved.
The principle behind the alignment is to first select which concept will be used as the title of the new dictionary article in the aggregated dataset. Each dictionary article (concept) consists of: a definition, a source list, a list of keywords (dialect entries in normalised spelling) and lexical variants (different dialectical expressions of the same concept) and location codes. All the related keywords and their lexical variants can then be aligned with the relevant 'DSDD concept'. For some concepts, the mapping is quite straightforward, e.g. where concepts and keywords are the same in all three dictionaries. However, for other concepts, different editorial choices have been made between the

[5] https://www.clarin.eu/content/standards-and-formats.
[6] WBD: Dictionary of the Brabantic Dialects, WLD: Dictionary of the Limburgian dialects, WVD: Dictionary of the Flemish dialects
[7] The WBD, the 'mother' of the two other projects, was finished in 2005; the WLD in 2008. They were compiled at the University of Nijmegen and Leuven. The WVD started 12 years later in 1972 at the Ghent University and will be finished by the end of 2018.
[8] For further details about this approach see: Van Keymeulen, J. (1992).

dictionaries. For example; one dictionary may have created two separate concepts, e.g. 'ant' and 'black ant', whereas in another dictionary there is a single concept, 'ant'. The challenge is how to integrate these concepts within the DSDD. Where inconsistencies appear at the 'lower' levels of the dictionary (e.g. keyword, lexical variant), the situation becomes even more complex.
The aim of this inventory was to identify and document the different types of issues that occurred in the data, including details of how they had been resolved. In this way, if similar issues arise at a later stage of the alignment, the editorial choices made can be referred to and reused, to increase the consistency across the dataset. Following this evaluation and standardisation phase, the 'real' concept alignment could begin.

3.2 Data import and concept alignment

In the original project proposal, it was anticipated that a concept-alignment tool would need to be developed. However, thanks to the collaboration with the Dutch Language Institute (INT) via CLARIN, their existing Lex'it tool could be adapted for the DSDD project. Lex'it is a rapid application development platform for lexicographical and other data production work benches developed at INT. It is used in the DSDD project for creating links between different entities (e.g. concepts, keywords, lexical variants) in the three dialect dictionaries.
For the concept alignment phase, firstly, the lexicographical data needed to be exported from the three databases. These exports, data-dumps, then needed to be 'imported' into Lex'it for alignment. To ensure the integrity of the underlying (source) datasets, they were stored separately within Lex'it and then merged into an integrated 'DSDD dataset'. Alignment has been carried out at concept level for approximately 2300 concepts. The next step will be to check these alignments at the level of keywords and lexical variants.

4 Enabling the DSDD for digital scholarship

Now that the concept alignment is underway, the preparation for the exploitation phase can begin. The aim of this phase is to develop a Virtual Research Environment (VRE) for digital lexicographical research, which will: a) make the newly aggregated DSDD available via a user-friendly website and b) enable the DSDD for digital scholarship. An Application Programming Interface (API) will be developed to enable the export of the aggregated DSDD dataset for analysis with existing digital research tools. Furthermore, the possibility of publishing the DSDD dataset as Linked Open Data will be explored. To evaluate the applicability of the newly aggregated DSDD for digital scholarship, some research use cases will be developed, for example in the fields of geo-visualisation, qualitative lexicology and dialectometry, based on existing research questions from the DSDD consortium, such as:

1. What systematic lexico-geographical patterns do the southern Dutch dialects show? Do they coincide with the traditional ones, based on phonology? (see De Vriendt 2012). Are there geographical patterns in semantics?
2. How far can the geographical spread of dialectology concepts be explored through the automatic generation of heatmaps from linked geocoordinates? Is it possible to automatically detect the homogeneity (or heterogeneity) of a particular dialectical concept using segmentation and clustering techniques?
3. How can cluster analysis be used to explore the linkage (and visualisation) of linguistic data with synchronic and diachronic extralinguistic data of all kinds?
4. Lexical diversity is defined as the number of different words or expressions that exist to refer to a particular concept. What is the relationship between concepts and their lexical diversity? Why does lexical diversity differ dramatically between different concepts? What are the factors that explain the differences between concepts? (Franco, 2017).

To facilitate the design of the research use cases, an interdisciplinary workshop is being planned in 2019 to bring together researchers from linguistics, digital humanities and cartography to explore how

recent advances in geo-visualisation, geo-humanities, web-mapping and digital cartography, can be applied to dialectology.

5 Sustaining the DSDD

The sustainability of the integrated DSDD dataset and related website has been considered from the outset. It is intended that the DSDD will be made available as a sustainable service beyond the lifetime of the project. To facilitate this, a data management plan for the DSDD is in the process of being created in liaison with the Ghent University Faculty Library of Arts and Philosophy. This will be regularly reviewed and where necessary updated throughout the project. The DSDD website and related research tools will be hosted at the Dutch Language Institute (INT) as part of the CLARIN infrastructure.

The DSDD is also a DARIAH-BE affiliated project. Once developed, it will be offered as a Belgian contribution to DARIAH-EU and as a result will be made available via the DARIAH Marketplace, which will considerably increase the visibility at the international level. In Flanders, the current funding phase of DARIAH comes to an end in December 2018. For the subsequent funding phase, DARIAH has joined forces with CLARIN, to establish a CLARIAH Open Humanities Services Infrastructure for Flanders. The aim of CLARIAH in Flanders is to unite and extend the existing portfolio of services for digital scholarship in the Arts and Humanities offered by the DARIAH-VL Virtual Research Environment Service Infrastructure (VRE-SI), with digital language data and tools offered through the CLARIN-VL portal. As the DSDD is already working with both infrastructures, the DSDD is in a unique situation to benefit from CLARIAH.

6 Future Research

The DSDD should be seen as a first phase in the inter-linking of Dutch dialect data. Subsequent phases include: the integration of the Dictionary of the Zealandic Dialects (WZD)[9](phase 2), the integration of the DSDD into the planned cumulative database for all onomasiological dictionaries of the entire Dutch area (phase 3), which could be linked with the databases of the alphabetical amateur dialect dictionaries in Belgium and the Netherlands to create a complete overview of the dialect vocabulary of the Dutch speaking area (phase 4). The Ghent Linguistics Department is currently working with volunteers on this Word Database of the Dutch Dialects[10], consisting of Flemish regional and local dialect dictionaries (Van Keymeulen en De Tier (2010)/(2013)). A comparable project is being undertaken in The Netherlands at the Meertens Institute[11]. At a later stage, the dialect data can be linked to the historical dictionaries of the Dutch platform[12] of the Dutch Language Institute (INT) (phase 5). The last phase (6) could be the linking with other dialectological databases concerning phonology, morphology, syntax and speech recordings.[13]

7 Conclusion

The DSDD will contain several million lexicographic tokens, that can be linked with other dialect data. Additionally, the Department of Dutch Linguistics at Ghent University also works on dialect projects that can at a later stage be combined with the DSSD to create a comprehensive platform for dialects in the Netherlands and Belgium. Working closely with both the DARIAH and CLARIN infrastructures will help to ensure the sustainability of DSDD and other digital dialectological research projects for future generations.

[9] WZD: https://www.zeeuwsedialect.nl
[10] Woordenbank: https://www.woordenbank.be
[11] eWND: http://www.meertens.knaw.nl/ewnd/
[12] See http://gtb.inl.nl/.
[13] Examples include: http://www.dialectzinnen.ugent.be, https://www.dialectloket.be/ (Geluid - Stemmen uit het Verleden) - https://research.flw.ugent.be/en/projects/parsed-corpus-southern-dutch-dialects - FAND - MAND - SAND.

References

De Tier, V. & J. Van Keymeulen. (2010), Software Demonstration of the Dictionary of the Flemish Dialects and the pilot project Dictionary of the Dutch Dialects. In: Dykstra, A & T. Schoonheim (eds.), Proceedings of the XIV Euralex International Congress. Fryske Akademy, Leeuwarden. blz. 620-627 (issued on CD-ROM).

De Vriendt, F. (2012), Tools for Computational Analyses of Dialect Geography Data. PhD Radboud University Nijmegen.

Franco, K. (2017), Concept features and lexical diversity. A dialectological case study on the relationship between meaning and variation. PhD KU Leuven.

Van Keymeulen, J. (1992), *De algemene woordenschat in de grote dialectwoordenboeken (WBD, WLD, WVD): een methodologische reflectie*. Onuitgegeven proefschrift, Ugent, Vakgroep Nederlandse Taalkunde

Van Keymeulen, J. (2004), Trefwoorden en lexicale varianten in de grote regionale woordenboeken van het zuidelijke Nederlands (WBD, WLD, WVD). In: De Caluwe J, G. De Schutter, M. Devos en J. Van Keymeulen, Taeldeman, man van de taal, schatbewaarder van de taal. Vakgroep Nederlandse Taalkunde UGent – Academia Press, Gent (2004); 1111 blz.; blz. 897-908.

Van Keymeulen, J. & V. De Tier (2010), Pilot Project: A Dictionary of the Dutch Dialects. In: Dykstra, A & T. Schoonheim (eds.), Proceedings of the XIV Euralex International Congress. Fryske Akademy, Leeuwarden. blz. 754-763 (issued on CD-ROM).

Van Keymeulen, J. & V. De Tier (2010), Towards the completion of the Dictionary of the Flemish Dialects. In: Dykstra, A & T. Schoonheim (eds.), Proceedings of the XIV Euralex International Congress. Fryske Akademy, Leeuwarden. blz. 764-773. (issued on CD-ROM).

Van Keymeulen, J. & V. De Tier (2013), The Woordenbank van de Nederlandse Dialecten (Wordbase of the Dutch Dialects). In: Kosem, I., J. Kallas, P. Gantar, S. Krek, M. Langemets & M. Tuulik (eds.). *Electronic lexicography in the 21th century: thinking outside the paper. Proceedings of the eLex 2013 conference. 17-19 October 2013, Tallinn. Estonia*. Ljubljana/Tallinn: Trjina, Institute for Applied Slovene Studies/Eesti Keele Instituut. p. 261-279.

SweCLARIN – Infrastructure for Processing Transcribed Speech

Dimitrios Kokkinakis
Department of Swedish
U of Gothenburg, Sweden
first.last@gu.se

Kristina Lundholm Fors
Department of Swedish
U of Gothenburg, Sweden
first.last@gu.se

Charalambos Themistocleous
Department of Swedish
U of Gothenburg,, Sweden
first.last@gu.se

Abstract

In this paper we describe the spoken language resources (including transcriptions) under development within the project "Linguistic and extra-linguistic parameters for early detection of cognitive impairment". The focus of the present paper is on the resources that are being produced and the way in which these could be used to pursue innovative in dementia prediction, an area in which more scientific investigations are required in order to research develop additional predictive value and improve early diagnosis and therapy. The language resources need to be thoroughly annotated and analyzed using state-of-the-art language technology tools and for that purpose we apply Sparv, a corpus annotation pipeline infrastructure which is part of the SweCLARIN toolbox. Sparv is offering state-of-the-art language technology as an e-research tool for analyzing and processing various types of Swedish corpora. We also highlight some of the difficulties in working with speech data and suggest ways to mediate these.

1 Introduction

Recent research has suggested that analysis of speech and language may lead to the discovery of sensitive behavioural biomarkers of dementia that can be obtained through non-invasive methods. Automated analysis techniques from Natural Language Processing (NLP) can provide objective measures of dementia since a range of markers can be automatically derived from spoken language and modelled in machine learning frameworks. For instance, a potential, early-stage diagnostic marker for neurodegenerative diseases, such as Alzheimer's disease, is the onset of language disturbances which is often characterized by subtle word-finding difficulties, impaired spontaneous speech, slight speech hesitancy, object naming difficulties, phonemic errors, simplified syntax or overuse of certain pronouns and empty words. Such information can be used for several purposes, such as *identifying cognitive decline, late-life dementia prediction*, *classifying dementia status* or *improving dementia screening tests.*

The main goal of our project is improving screening tools for dementia. For this purpose, we use connected speech (both the speech signal and its transcription), since it can provide valuable information in a non-invasive and easy-to-assess way for determining the severity of language impairment. Moreover, detailed linguistic analysis of such data requires a reliable infrastructure that can be applied to annotate the transcribed spoken input in multiple ways. Therefore, for the linguistic analysis we utilize Sparv (Borin et al., 2016) which is an integrated part of the SweCLARIN toolbox.

In this paper we provide information about the elicitation of spoken language for research purposes. Thereafter, we give a brief description of our speech acquisition procedure, which results in audio recordings and transcriptions of several language tasks. We then describe our data analysis pipeline, focusing on Sparv. Finally, we discuss future research and some of the difficulties of working with speech data in an infrastructure developed for text, and suggest ways of accommodating these difficulties.

2 Background

Controlled elicitation of spontaneous speech and narrative discourse samples from healthy individuals as well as from individuals with a variety of mental and cognitive disabilities, at various stages, have been developed and studied for several decades (see for example Bryant et al., 2016; Mueller et al.,

2018). These include story (re)tellings, conversational interviews, description of a single picture of picture or picture sequences etc. There are several ways to approach collecting spontaneous spoken data (monologue and conversational) from participants in a prospective study. For monologue data, a single picture or a picture sequence description has been extensively used (de Lira et al., 2011), while for a more conversational-like data, a task-oriented dialogue can be one approach. In such a task, the participants collaborate through spoken interaction to solve a problem such as planning a route through a map of a scene, visiting a selection of places while bypassing others (e.g. a zoo; Wilson et al., 1998). The aim is to gather language data that is naturalistic and spontaneous while at the same time comparable between participants and over time.

Systematic forms of language assessment can play a crucial role in the clinical diagnosis of several psychiatric illnesses and neurodegenerative diseases, and it has emerged as a valuable resource for describing and even quantifying the degree and severity of e.g. cognitive decline, aphasia or schizophrenia. Moreover, recent advances in NLP and automatic speech recognition (ASR) make it possible to develop objective and precise instruments for automated or semi-automated analysis of spoken language (Pakhomov, et al., 2010). In a recent review, Boschi et al. (2017) discuss the crucial role that language assessment has in the clinical diagnosis of several neurodegenerative diseases. Since there is no publicly available Swedish speech database of a relevant population, who perform at least some of the aforementioned tasks, we designed a data collection procedure specifically targeting a population with mild[1] and subjective[2] forms of cognitive decline (Gauthier et al., 2006) and healthy, age matched controls (HC).

3 Population, Data Collection and Transcription

All samples are produced by Swedish speakers after obtaining written consent approved by the local ethics committee. Our population consists of 90 participants: 36 healthy controls, 23 persons with SCI and 31 persons with MCI. Table 1 provides some descriptive statistics about the population. Education differs significantly between the groups, and a post-hoc LSD test reveals that the SCI group is significantly more educated than both the HC group ($p = 0.001$) and the MCI group ($p = 0.026$). The MMSE score differs significantly between the three groups, and post-hoc LSD tests show that scores in the MCI group are significantly lower than in the HC group ($p < 0.0001$) and the SCI group ($p < 0.0001$), whereas there is no difference between the SCI group and the HC group.

	HC (*n=36*)	SCI (*n=23*)	MCI (*n=31*)
Age (years)	67.9 (7.2)	66.3 (6.9)	70.1 (5.6)
Education (years)	13.2 (3.4)	16.1 (2.1)	14.1 (3.6)
Sex (F/M)	23/13	14/9	16/15
MMSE (/30)	29.6 (0.61)	29.5 (0.90)	28.2 (1.43)

Table 1. Age, education, sex and MMSE score for the participants. The MMSE (MiniMental State Exam) is a test of general cognitive ability, where max is 30 and a score of ≤24 is been proposed for cognitive impairment; a score between 25-27 indicates possible cognitive impairment which should be further evaluated (Palmqvist et al., 2013).

Recordings were made on two occasions (rounds). The first recording round was carried out in the second half of 2016 and early 2017 and the second round, a repetition of the first plus new tasks, is being made in 2018. After considering the different types of tasks described in the previous section, we decided to incorporate and, consequently, record, the following tasks: a *picture description task*, the 'Cookie theft' (Goodglass et al., 2001) (round 1 and 2), a *read aloud task* (round 1 and 2), a *complex planning task* (round 2), a *map task* (round 2) and a *semantic verbal fluency task*, category 'animals' (round 2). The audio capture of the narrative data were collected using a Zoom digital recorder and the resulting audio files were saved and stored as uncompressed audio in .wav 44.1 kHz with 16-bit resolution. The

[1] MCI: a prodromal state of dementia and a transition between normal aging and early dementia, in which a person has minor problems with cognition but these are not severe enough to warrant a diagnosis of dementia.

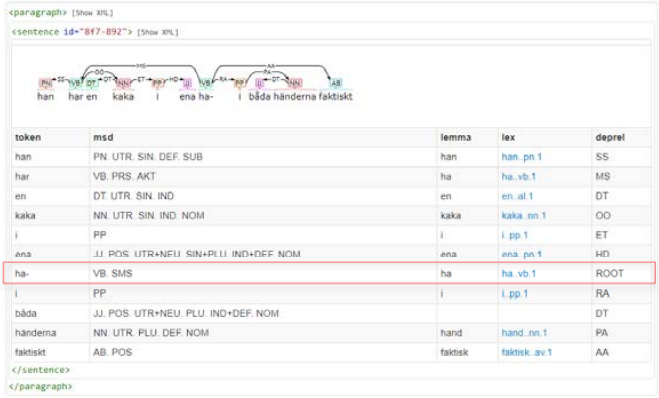
[2] SCI: a common diagnosis in elderly people, sometimes suggested to be associated with e.g. depression, but also a risk factor for dementia and is mainly characterized by an individual's subjective experience of cognitive decline.

recordings are carried out in an isolated environment at the University lab in order to avoid noise. A speech pathologist and/or a computational linguist were present during the recording sessions, providing all subjects with identical instructions according to a predefined protocol.

The total length of the recordings in the first round is 2 h 35 min for the picture descriptions, 1 h 31 min for the first reading task and 1 h 14 min for the second reading task. The transcriptions (both orthographic with standardized spelling and one with maintained spoken language phenomena, such as partial words) during the first round were made manually by professional transcribers; while the transcriptions of the second round will be made automatically with a speech-to-text system ("THEMIS-SV") that is being developed in the current project (Themistocleous & Kokkinakis, 2018). All alignments during the second round will be manually verified and corrected when necessary, while the alignment will be made in Praat (Boersma & Weenink, 2016). THEMIS-SV processes the recordings and returns an output with three tiers: the utterance tier, the word tier, and the vowels' and consonants' tier. The system will be evaluated and compared to manual transcriptions. The automatic segmentation of speech enables targeted acoustic measurements, such as consonant spectra, formant frequencies of vowels, fundamental frequency, pauses, speech rate, etc. and other acoustic measurements that have been known to differentiate between the different types of language disorders.

The transcriptions require some pre-processing before using automatic annotation tools such as Sparv. In written text, sentences and clauses are typically delimited by punctuation marks, whereas in spoken language, boundaries are indicated by for example pauses and prosodic patterns. The transcribers were therefore asked to identify appropriate segmentation points and to manually add a full-stop based on a set of guidelines. That was a necessary step since most NLP tools require sentence boundaries, and good automatic machine learning models for the task are not available for Swedish. During transcription a number of other phenomena are also annotated, such as filler words (words or sounds that a speaker utters while thinking about what he/she is going to say next such as *uhm*), corrections and false starts (where speakers begin a sentence but change their plan of what they want to say and continue different), and also non-verbal vocalizations such as laughing and coughing. By tagging these phenomena we were able to easily exclude them when analysing the transcriptions with the SweCLARIN[3] infrastructure.

4 Automatic Linguistic Annotation



token	msd	lemma	lex	deprel
han	PN. UTR. SIN. DEF. SUB	han	han..pn.1	SS
har	VB. PRS. AKT	ha	ha..vb.1	MS
en	DT. UTR. SIN. IND	en	en..al.1	DT
kaka	NN. UTR. SIN. IND. NOM	kaka	kaka..nn.1	OO
i	PP	i	i..pp.1	ET
ena	JJ. POS. UTR+NEU. SIN+PLU. IND+DEF. NOM	ena	ena..pn.1	HD
ha-	VB. SMS	ha	ha..vb.1	ROOT
i	PP	i	i..pp.1	RA
båda	JJ. POS. UTR+NEU. PLU. IND+DEF. NOM			DT
händerna	NN. UTR. PLU. DEF. NOM	hand	hand..nn.1	PA
faktiskt	AB. POS	faktisk	faktisk..av.1	AA

</sentence>
</paragraph>

Figure 1. Processing results using Sparv and the analysis in a table format. Here the truncated token *ha-* is erroneously pos-annotated as VB ([compound] verb) instead of noun which also produces an erroneous dependency annotation.

For the linguistic annotation of the transcriptions we use standard tools for the automatic processing of Swedish texts provided in Sparv (Borin et al., 2016), a major infrastructure for Swedish processing and part of the SweCLARIN. Sparv consists of several NLP tools for instance, tools for lexical and compound analysis, part-of-speech tagging, lemmatization and syntactic analysis. Sparv comes with a web interface[4] and Figure 1 shows a screenshot of the test sentence *han har en kaka i ena ha- i båda händerna*

[3] https://sweclarin.se/eng.
[4] https://spraakbanken.gu.se/sparv.

faktiskt 'he has a cookie in one ha- in both hands actually'. The user can view the resulting analysis in a table, as in the image above. When the user hovers over abbreviations used for e.g. dependency labels, an expanded description is shown. The tools are connected in a pipeline for easy and fast processing; moreover, various processing tools can be deactivated, which enables users to exclude some part of the annotation if they wish and use their own annotation instead. For illustrative purposes we have not excluded the fragment *ha-* from Figure 1, which is erroneously annotated as verb and also negatively influences the dependency results (the relation *ROOT* which is assigned to the wrong token). However, we manually corrected such systematic errors which also highlights the need for pre- or postprocessing of spoken data. These errors arise since transcriptions have distinctive characteristics not commonly found in (well-formed) written texts, which are the texts usually used for building NLP tools. Moreover, we have not performed any formal evaluation of the obtained annotations.

5 Exploitation, Future Directions and Conclusions

The corpus we have presented represents a valuable resource for early detection of a neurodegenerative diseases and we have outlined how these resources might be exploited for our research agenda. Several studies based on both the spoken signal and its transcription have been completed; see for instance Lundholm Fors et al. (2018) and Fraser et al. (2017; 2018), and more are planned for the near future. Our national project, that develops the corpus and machine learning models described in this paper, ends in 2019. After that we plan to open up part of the resources: this will include the transcriptions, which currently are encoded using XML, and demographic information. Spoken data will become more easily accessible as automatic speech recognition continues to improve, and the need for time-intensive manual transcription decreases. With this added wealth of linguistic data, we see a need for language tools to be adapted to accommodate both written and (transcriptions of) spoken language. A first step would be to add automatic sentence segmentation to Sparv, and another would be the possibility to handle interrupted utterances and corrections; features that are typical of spoken language. While removing them in a pre-processing step works, by doing this we are removing characteristics of the data that if analysed within the framework, could add further nuance to the linguistic analysis.

Acknowledgements
This work has received support from *Riksbankens Jubileumsfond* - The Swedish Foundation for Humanities and Social Sciences, through grant agreement no: NHS 14-1761:1 as well as SweCLARIN.

References

Paul Boersma and David Weenink. 2016. *Praat: doing phonetics by computer* [Computer program]. Version 6.0.19, retrieved in Aug. 2016 from <http://www.praat.org/>.

Lars Borin, et al. 2016. Sparv: Språkbanken's corpus annotation pipeline infrastructure. *The 6th Swedish Language Technology Conference (SLTC)*. Umeå University.

Veronica Boschi, et al., 2017. Connected Speech in Neurodegenerative Language Disorders: A Review. *Front. Psychol.* 8:269.

Lucy Bryant, Alison Ferguson and Elizabeth Spencer. 2016. Linguistic analysis of discourse in aphasia: A review of the literature. *Clinical Linguistics & Phonetics*. 30:7, 489-518. DOI: 10.3109/02699206.2016.1145740.

Kathleen C. Fraser, Kristina Lundholm Fors, Dimitrios Kokkinakis and Arto Nordlund. 2017. An analysis of eye-movements during reading for the detection of mild cognitive impairment. *EMNLP*. Pp 1027–1037. Denmark.

Kathleen C. Fraser, et al. 2018. Improving the Sensitivity and Specificity of MCI Screening with Linguistic Information. 2nd RaPID workshop. Pp. 19-26. Miyazaki, Japan.

Serge Gauthier, et al. 2006. Mild cognitive impairment. *The Lancet* 367 (9518), 1262–1270.

Harold Goodglass, E. Kaplan and B. Barresi. 2001. *Boston Diagn Aphasia Exam*. Phil: Lippencott, Williams & Wilkins.

de Lira, J.O. et al. 2011. Microlinguistic aspects of the oral narrative in patients with Alzheimer's disease. *International Psychogeriatrics* Apr;23(3):404-12. doi: 10.1017/S1041610210001092.

Kristina Lundholm Fors, Kathleen C. Fraser and Dimitrios Kokkinakis. 2018. Automated Syntactic Analysis of Language Abilities in Persons with Mild and Subjective Cognitive Impairment. *The 29th Med Info Eur.* Sweden.

Kimberly D. Mueller, Bruce Hermann, Jonilda Mecollari and Lyn S. Turkstra. 2018. Connected speech and language in mild cognitive impairment and Alzheimer's disease: A review of picture description tasks. *J of Clin and Exp Neuropsychology*.

Serguei Pakhomov, et al. 2010. A computerized technique to assess language use patterns in patients with frontotemporal dementia. *J Neuroling.* 23(2):127–144.

Sebastian Palmqvist, B. Terzis, C. Strobel and A. Wallin. 2013. MMSE-SR: Mini Mental State Examination - Svensk Rev.

Charalambos Themistocleous and Dimitrios Kokkinakis. 2018. THEMIS-SV: Automatic classification of language disorders from speech signals. *The 4th European Stroke Organisation Conf.* Gothenburg, Sweden.

Barbara A Wilson, et al. 1998. The development of an ecologically valid test for assessing patients with a dysexecutive syndrome. *Neuropsych Rehab,* 8, 213–228.

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 Anlaysis) tools and the associated CHAT transcription format. 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 5TB of richly annotated media. (MacWhinney, 2000).

	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

Currently, most users interact with TalkBank data by using the CLAN program. CLAN consists of a set of tools for annotating media with CHAT, playing back the media with time-stamped annotations, and for 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 spend time reading the CLAN manual and becoming acquainted with CHAT annotations. This can be a barrier to those who may not be interested in individual transcripts, but who wish to derive general statistics and patterns of features defined across the annotated media. CLAN is mostly tuned for working with single corpora, and is not intended to query the entire TalkBank corpus to extract general patterns and statistics. 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), then browsing/downloading the media and annotations.

Without prior knowledge of how TalkBank is structured and what corpora exist within each, it is difficult to find or be aware that particular resources exist. TalkBankDB provides a single online interface to query across all of TalkBank to find names of relevant corpora and links to media and transcripts. For example, a query for the Spanish language yields a list of media 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), 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 effectively allows users to define new corpora based on features they define (Figure 1 and Figure 2).

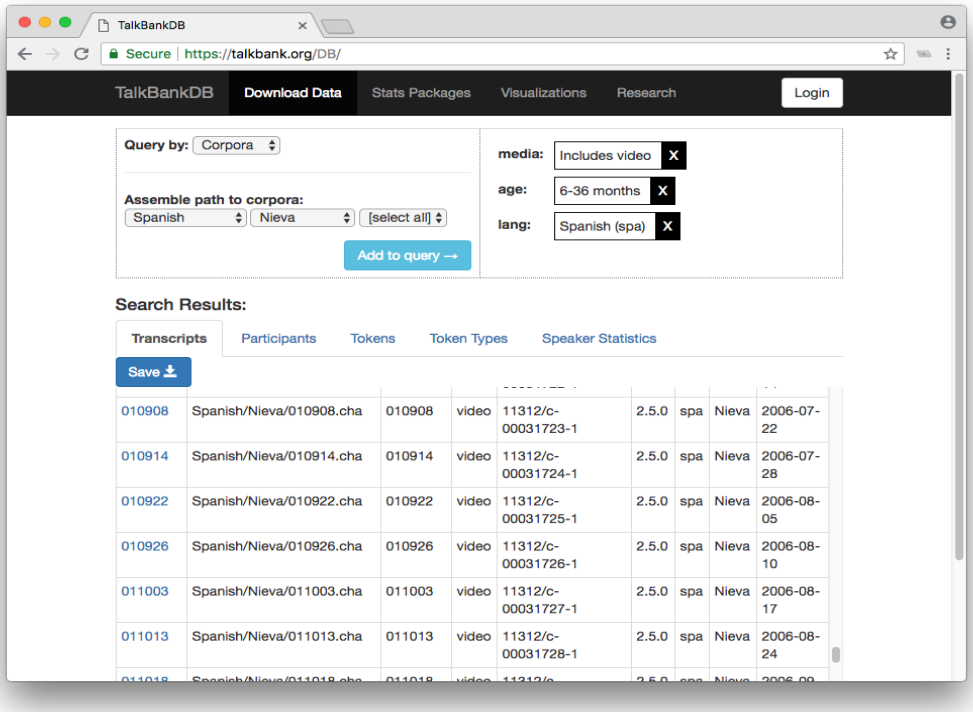


Figure 1. A query yields a table of all matching documents with metadata for each, allowing the user to further refine the query.

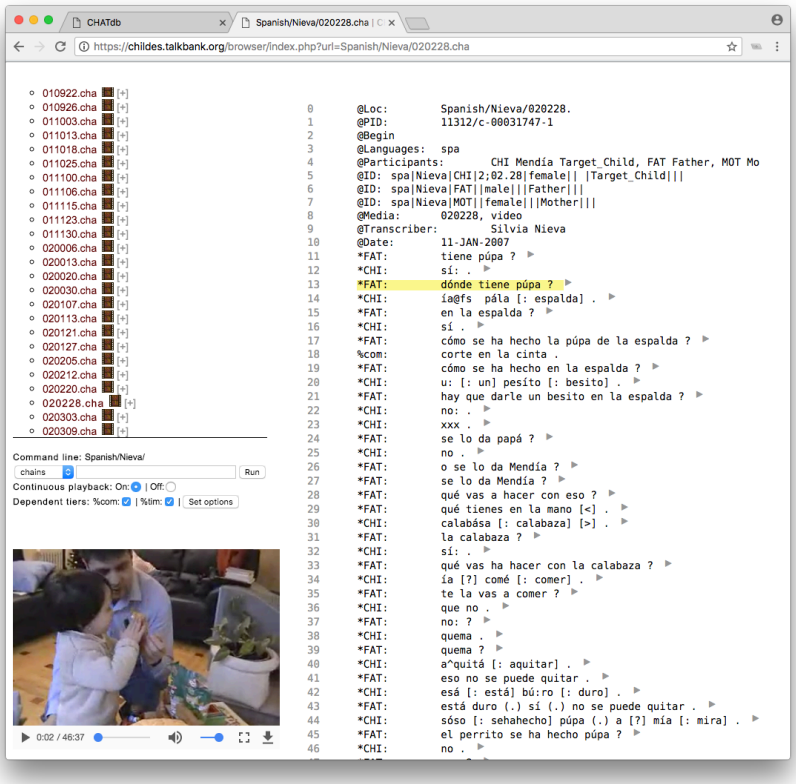


Figure 2. Clicking on the name of the transcript loads the corresponding media and annotations in the browser and allows for direct playback of the media.

In addition to using TalkBankDB to easily find media with specific features across TalkBank, researchers can query to get statistical summaries of the annotated media. 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 at http://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 their published R library: get_transcripts(), get_participants(), get_tokens(), get_types(), get_utterances(), and get_speaker_statistics(). For the web interface, they employ 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 (at http://gandalf.talkbank.org:8080) to offer a web interface to extract and analyse variables from the transcripts. However, this facility is based on a complete 140GB spreadsheet encoding of all data in CHILDES. TalkBankDB differs from these facilities by creating a JSON database in MongoDB, thereby reducing storage to about 8GB, and markedly improving responsivity.

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, we eliminate the need of 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, we can round-trip from JSON to the original CHAT formatted transcript.

Since much of the data and metadata contained within the TalkBank CHAT transcripts are changing, using a common relational database like mySQL with a strict tabular schema is not as suitable as something with the flexibility of a document database. The effort to pre-set a clear schema with a normalized relational database provides little benefit and 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 it makes scaling to increasing data demands easy by allowing the database to scale out across multiple inexpensive machines through "sharding" of the database. 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 scaling-up strategy is not always possible, and can one day 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 public 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 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 users. 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.

5 Related Work

The design and scope of TalkBankDB has been influenced by our work with several related projects, including SketchEngine, EXMARaLDA, MTAS, ANNIS, CQL, and Alpheios, as well as the childes-db and LuCiD Toolkit projects mentioned earlier. In its current shape, TalkBankDB does not yet provide the full functionality of these systems. However, it is our goal to implement the functions of these various systems for use with TalkBank multi-tier annotations and multimedia corpora.

6 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 one "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 and morphologies can easily be adapted for inclusion in TalkBankDB.

7 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/

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

[MacWhinney 2014] MacWhinney, B. (2014). The childes project: Tools for analyzing talk, volume ii: The database. Psychology Press.

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. (in prep). childes-db: a flexible and reproducible interface to the Child Language Data Exchange System (CHILDES). Manuscript in preparation.

L2 learner corpus survey – Towards improved verifiability, reproducibility and inspiration in learner corpus research

Therese Lindström Tiedemann
University of Helsinki, Finland
therese.lindstromtiedemann@helsinki.fi

Jakob Lenardič
University of Ljubljana
Ljubljana, Slovenia
jakob.lenardic@ff.uni-lj.si

Darja Fišer
University of Ljubljana
Ljubljana, Slovenia
darja.fiser@ff.uni-lj.si

Abstract

We present a survey of the second language learner corpora available within CLARIN. The survey provides a test of the ease of finding these corpora through the VLO and the extent of the metadata and documentation which users have included. Based on this we suggest some ways of improving the usefulness of the VLO and making more linguists aware of what CLARIN provides. Furthermore, we suggest that in addition to collecting data and metadata, a bibliographical database of research using and documenting work on second language learner corpora should be collaboratively maintained.

1 Introduction

Learner corpus research has strongly established itself as a discipline in the last couple of decades. At the Teaching and Language Corpora (TaLC) conference in 1994, only the odd learner corpus was mentioned, and the organisers had thought that the focus of the whole conference would rather be on teaching linguistics with corpora (McEnery 2018). Twenty years later the proceedings from TaLC instead focus on learner corpora, both the 'well-established' and smaller 'private' corpora (McEnery 2018: xvii, Brezina & Flowerdew 2018), leading Brezina and Flowerdew (2018: 1) to quite rightly claim that today 'corpora play a crucial role in second language (L2) research and pedagogy'.

A large number of second language (L2) learner corpora have been compiled, and although most of them are still mainly for English as a target language or with English as the (main) L1 of the learners, corpora for several other languages are becoming increasingly available. The Centre for English Corpus Linguistics at the University of Louvain (UCL)[1] has compiled a rather comprehensive online list of learner corpora with links and names of contacts and basic information about the size, medium, etc.[2] However, many corpora are still being compiled without much awareness about what is already available and many existing corpora are still kept internally within research groups or even by individual teachers. Obviously, the L2 research community would gain in many ways from making it easier to find information about the corpora that already exist, and preferably also making them searchable or even downloadable.

Researchers and corpus developers alike could benefit from learning about the design of existing learner corpora and from being able to study their content and compare that to material collected by other groups. If they had more information about what there already is, they could also try to make new resources more comparable; e.g., by agreeing on a common error annotation schema, or considering making an extension to a corpus rather than starting a completely different corpus, thereby moving towards bigger, more comprehensive and more representative learner corpora. Furthermore, making corpora easily available means that learner corpus research contributes to reproducibility and

[1] https://uclouvain.be/en/research-institutes/ilc/cecl
[2] https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html

verifiability of research. The most promising way towards this goal seems to be through a research network like CLARIN focused on providing language data to researchers.

The goal of the paper is to present the results of a recent survey of second language learner corpora and give an overview of the existing corpora. Furthermore, we wish to emphasise the need to make learner corpora more easily findable and to make documentation about the corpora more widely available to facilitate the reuse of the corpora, but also to inspire people who would be interested in compiling their own corpora. We present a brief overview of the corpora (section 2) which are currently available within CLARIN before moving on to look at improvements which we believe should be made to the VLO in order to make this particular kind of corpora more easily findable and usable (section 2.3). Finally, we finish with a summary and some thoughts about future work (section 3).

2 Survey of L2 learner corpora

Learner corpora are generally still quite small and they are unlikely to reach the billions of tokens that are already commonplace for many reference corpora. This is because data have to be collected from several different learners who have to give their consent and the data has to be at least partly manually processed (transcribed, error annotated, anonymised etc.). All of this requires considerable resources, and convincing learners to contribute is not always easy. This is why a more efficient approach towards bigger corpora would be to join efforts and pool resources, which requires documenting the existing corpora, producing best practices and compiling a bibliography of papers on corpus construction as well as studies based on L2 learner corpora in a way that is easily accessible to the research community and easy to contribute to by the entire community.

2.1 Setup of the survey

In this paper, we present the first step in this direction. We have surveyed the existing repositories and curated lists with the goal of compiling a comprehensive overview of the available L2 corpora in the CLARIN infrastructure and beyond. By searching the Virtual Language Observatory (VLO)[3] and the UCL list of learner corpora worldwide[4] as well as through input from participants at the CLARIN-sponsored Workshop on Interoperability of L2 resources and tools[5] that took place between 6–8 December 2017 in Gothenburg, Sweden, we have identified 180 L2 corpora in the CLARIN countries, some of which do not seem to be available through any channel, which is why they are not discussed in the remainder of this paper.

We divide the identified corpora into CLARIN and non-CLARIN corpora. We consider a corpus to belong to CLARIN if it is listed in the VLO or in one of the national repositories. In addition to *learner corpora proper* according to Granger's definition as 'electronic collections of (near-) natural foreign or second language learner texts assembled according to explicit design criteria' (Granger 2008: 338), we have also taken into account the so called *peripheral learner corpora* (Nesselhauf 2004: 128), e.g. texts being read aloud, or tasks which entail less natural language such as translation or tasks that restrict the learner much more in their language use, *databases* (Gilquin 2015: 10) that entail both natural, near-natural and experimental material.

2.2 Overview of the results

The results of our survey, summarized in Table 1, show that **34 L2 learner corpora** are available within the CLARIN infrastructure. To a large extent (91%) they can be found in VLO. However, the survey has also identified many corpora that remain outside CLARIN (but within the CLARIN member states) and there are of course also a large number of L2 learner corpora in the world that are completely outside the CLARIN network. Based on these results we can conclude that CLARIN already plays an important role in L2 corpus research but is still lacking a comprehensive coverage of the existing corpora of this type and that if more corpora were deposited CLARIN could help us make L2 corpora more comparable and promote further use of the existing corpora. Hence we believe that CLARIN should become more

[3] https://vlo.clarin.eu/
[4] https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html
[5] https://sweclarin.se/swe/workshop-interoperability-l2-resources-and-tools

proactive in contacting projects working on second language acquisition and language departments in general to inform them of CLARIN and the benefits thereof.

	Written	Spoken	Mixed / multi-modal	TOTAL
CLARIN English L2	5	4	1	10 (29%)
CLARIN Other L2	13	7	4	24 (71%)
Listed in the VLO	15	11	5	31 (91%)
Available for download or querying	12	10	1	23 (68%)
Available through a CLARIN repository	8	7	/	15 (44%)
Specify the L1 language(s)	7	6	2	15 (44%)
Include metadata on size	16	6	2	24 (71%)
Include metadata on annotation	4	5	/	9 (26%)
Include licence	16	9	3	28 (82%)
TOTAL CLARIN	18	11	5	34 (19%)
TOTAL NON-CLARIN				146 (81%)
TOTAL				180

Table 1: Number of known L2 learner corpora in the CLARIN infrastructure at the time the report was compiled.

Less than half of the 34 L2 corpora (44%) are available directly through CLARIN repositories or concordancers. In terms of metadata, only 44% corpora specify the L1 background of the learners – while this is in some cases likely to be because the participants come from a great variety of L1 backgrounds (one corpus lists 54), others simply do not make this explicit. Metadata on size fares better (available for 71% corpora) although it is inconsistently listed with respect to what kind of size is provided. One spoken corpus provides only the word length of each individual transcription while another only gives the total temporal size of the recordings, even though it also includes transcriptions. Furthermore, information on the annotation levels is even less frequent (26% of the corpora) and varies widely from corpus to corpus. Only about half of the 9 corpora for which information on annotation is available display L2-specific markup in addition to the general annotation layers such as PoS-tagging or lemmatisation. Finally, license information is included for most corpora (82%), which is a welcome result given the nature of the content.

The more recent corpora increasingly focus on the L2 languages other than English and even though internationally we see that most L2 corpora still cover English, 71% of the L2 corpora in the CLARIN infrastructure primarily contain data in languages other than English. This is an invaluable contribution to the research community which has been asking for data for other L2s.

2.3 Identified issue and recommendations

Finding L2 corpora. As there is no designated facet for *learner corpora* in the VLO search facility, it is difficult to find learner corpora unless the key phrase "learner corpus" is part of the title. Additionally, a researcher who looks for a learner corpus often has to simply infer the nature of the language used in the corpus. As an example, *MERLIN Written Learner Corpus for Czech, German, Italian 1.1* (Wisniewski et al. 2018), which is available through the LINDAT repository (as well as listed in the VLO), lacks overt information that would clearly disambiguate which of the three languages used in the corpus are the speakers' L1 and L2 languages.

As a tentative solution, we suggest that a special facet be added to the VLO by means of which the authors of the corpus could specify if the data are primarily produced by native or non-native speakers. Ideally, researchers should be able to use the VLO to find learner corpora with e.g. Swedish as the target language, or corpora with learners who have German as their L1, or longitudinal corpora.

Metadata and documentation. The range and availability of corpus metadata in the VLO varies a lot and calls for standardisation and recommendations which take into account the specific needs of the L2 researcher community. In addition, some thought should be put into how this can be made searchable to some extent through the VLO as stated above.

The importance of carefully considering the range of metadata encoded in the corpus and covered in corpus descriptions is discussed by several authors (e.g. Burnard 2005, Gilquin 2015). One of the most common types of metadata is the target language (L2) and the first language (L1) of the learners. However, there is some variation in how L1 is recorded in corpora: is it simply the learner's L1(s) or are the learner's parents' L1s that is listed? Similarly, there is variation in how other parts of the metadata are understood and listed, such as how proficiency is generally listed and how this is judged: is it based on the learner's general proficiency (how?) or on the text(s) included in the corpus? Moreover, is it clear how the material was collected and how it is to be interpreted? Gilquin (2015) underlines the need of trying to standardise metadata in order to make corpora more comparable, but there are many complicating factors (cf. Stemle et al. accepted) which all point to the need to work together on finding a way of standardising this as far as possible, while taking enough national legal perspectives and different linguistic perspectives into account.

We suggest that CLARIN provide a set of clear guidelines on how to present metadata related to content. In the case of learner corpora, this would mean that the authors/curators should be encouraged to describe their corpora in a consistent manner, by always providing the same types of information, such as the following: (i) whether the corpus contains data in the speakers' L1 or L2 language (or possibly both), which is especially important information because the current language facet does not distinguish between the two; (ii) the L1 backgrounds of the speakers; (iii) the speakers' age; (iv) the type of learning tasks that constitute the corpus (e.g., essays written in the context of a certain language-proficiency examination); (v) the number of speakers involved in the learning task; and (vi) detailed information on the learners' proficiency level.

It should be emphasised that a majority of the surveyed learner corpora are very inconsistently documented in this respect, so we believe that drafting and implementing a set of guidelines such as the ones just proposed would lead to a better overall presentation of the content of the current and prospective corpora in the infrastructure. Additionally, we suggest that authors/curators take special care to tag the crucial metadata, such as the target language, the L1s, other L2s, the medium, the proficiency levels included (CEFR-levels), the type of the collection setting, and so forth.

3 Conclusions

Based on the survey presented in this paper, it is clear that even though there are many L2 learner corpora already integrated in the CLARIN infrastructure, there are still many existing valuable resources yet to be added. The survey has also uncovered the need for fine-tuning the VLO search functionalities for querying L2 corpora more efficiently. Crucially, a consensus on the L2 corpus metadata and a comprehensive implementation of the consensus is urgent as there are big discrepancies in the documented metadata. Furthermore, the documentation of many corpora in the CLARIN infrastructure is incomplete; for example, many corpora lack references to articles which treat the design of the corpus or those that report on the research conducted on the basis of its data. This not only means that it is hard to ascertain how earlier corpora have been designed, but it also means that linguists may find it difficult to use the corpora in their research since there is not enough information about the collection context and the type of material included to make it useful from the point of view of research methodology.

In communications with the L2 research community, it has been revealed that many researchers in this field are still unfamiliar with the CLARIN infrastructure and how it could support their work. Consequently, many of the researchers collect data to carry out research tasks on their own without considering how their data could be used by a larger research community. This highlights the need for CLARIN to become more involved with this community and to inform the researchers about the available services (esp. the depositing services). We also suggest that research groups should be contacted by national CLARIN consortia to organise information sessions about the benefits of using the infrastructure, which would also provide the research groups with an invaluable opportunity to present their work to CLARIN and for both sides to realise how they could cooperate. Additionally, presentations of research based on data deposited in the CLARIN infrastructure should be circulated more among linguists through emails, blog entries, etc.

Acknowledgment

The work reported here has received funding from the *Riksbankens jubileumsfond P17-0716:1* project (first author), and (through CLARIN ERIC) from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 654119 for the project *PARTHENOS: Pooling Activities, Resources and Tools for Heritage E-research Networking, Optimization and Synergies* (second and third authors). We would also like to thank the anonymous reviewers for their helpful comments, and all the CLARIN ERIC User Involvement representatives and National Coordinators (as well as everyone else) who have contributed to the original survey of the L2 corpora.

References

[Brezina and Flowerdew 2018a] Vaclav Brezina and Lynne Flowerdew. 2018a. Introduction. In: Brezina and Flowerdew (eds), pp. 1–2.

[Brezina and Flowerdew 2018b] Vaclav Brezina and Lynne Flowerdew (eds). 2018b. *Learner Corpus Research – new perspectives and applications.* London and New York: Bloomsbury.

[Burnard 2005] Lou Burnard. 2005. Metadata for corpus work. In: M. Wynne. (ed.) Developing Linguistic Corpora: A Guide to Good Practice. Oxford: Oxbow Books, pp. 30–46. Available at http://ota.ox.ac.uk/documents/creating/dlc/ [Accessed 24 April 2018]

[Gilquin 2015] Gaëtanelle Gilquin. 2015. From design to collection of learner corpora. In: S. Granger, G. Gilquin & F. Meunier (eds.) The Cambridge Handbook of Learner Corpus Research. Cambridge: CUP. pp. 9–34.

[Granger 2008] Sylviane Granger. 2008. Learner corpora in foreign language education. In: N. Van Deusen-Scholl & N. H. Hornberger (eds.) Encyclopedia of Language and Education, vol. 4: Second and Foreign Language Education. New York: Springer. pp. 337–51.

[Lenardič, Lindström Tiedemann and Fišer 2018] Jakob Lenardič, Therese Lindström Tiedemann and Darja Fišer. 2018. Overview of L2 corpora and resources. https://office.clarin.eu/v/CE-2018-1202-L2-corpora-report.pdf

[McEnery 2018] Tony McEnery. 2018. Preface. In: [Brezina and Flowerdew 2018b].

[Nesselhauf 2004] Nadja Nesselhauf. 2004. Learner corpora and their potential in language teaching. In: J. Sinclair (ed.) *How to Use Corpora in Language Teaching*. Amsterdam: Benjamins, pp. 125-52

[Stemle et al. accepted] Egon W. Stemle, Adriane Boyd, Maarten Jansen, Therese Lindström Tiedemann, Nives Mikelić Preradović Alexandr Rosen, Dan Rosén, Elena Volodina. Accepted. Working together towards an ideal infrastructure for language learner corpora.

[Volodina et al. accepted] Elena Volodina, Maarten Jansen, Therese Lindström Tiedemann, Silje Ragnehildsveit, Kari Tenfjord, Koenraad de Smedt. Accepted. Interoperability of Second Language Resources and Tools. CLARIN annual conference 2018, Pisa, Italy.

[Wisniewski et al. 2018] Katrin Wisniewski, Andrea Abel, Kateřina Vodičková et al. 2018. *MERLIN Written Learner Corpus for Czech, German, Italian 1.1*. Eurac Research CLARIN Centre. http://hdl.handle.net/20.500.12124/6.

DGT-UD: a Parallel 23-language Parsebank

Nikola Ljubešić
Department of Knowledge Technologies
Jožef Stefan Institute, Slovenia
nikola.ljubesic@ijs.si

Tomaž Erjavec
Department of Knowledge Technologies
Jožef Stefan Institute, Slovenia
tomaz.erjavec@ijs.si

Abstract

We present DGT-UD, a 2 billion word 23-language parallel parsebank, comprising the JRC DGT parallel corpus of European law parsed with UD-Pipe. The paper introduces the JRC DGT corpus, details its annotation with UD-Pipe and discusses its format under the two CLARIN.SI web-based concordancers and its repository. An analysis is presented that showcases the utility of the corpus for comparative multilingual research. The corpus is meant as a shareable CLARIN resource, useful for translators, service providers, and developers of language technology tools.

1 Introduction

Recently two powerful concordancers, KonText[1] and noSketch Engine[2] have been added to the CLARIN.SI supported services. The two concordancers share their back-end, namely Manatee (Rychlý, 2007), a reimplementation of CQP (Christ, 1994), meaning that they also share their file formats, as well as compiled corpora. Furthermore, this format is used by a large number of other Manatee and CQP based installations of concordancers all over the world.

To expand the number of corpora offered, stress-test the implementation of KonText and noSketch Engine, and experiment with the idea of sharing corpora among CLARIN partners, we are producing several new corpora to add them to the concordancers, and, further, to the CLARIN.SI repository. This paper documents the first large and generally useful addition, namely the 23-language parallel DGT-UD parsebank (automatically parsed corpus), containing over 2 billion words.

2 The JRC DGT corpus

The JRC DGT corpus (Steinberger et al., 2012; Steinberger et al., 2014) is the translation memory of Acquis Communautaire (European Union law), first made publicly available in 2007 by the European Commission's Directorate-General for Translation (DGT) and the Joint Research Centre (JRC) of the European Commission, with the goal of "fostering the European Commission's general effort to support multilingualism, language diversity and the re-use of Commission information".[3]

The corpus contains texts in 24 languages aligned on the sentence level and is distributed as a set of TMX files identified by the EUR-Lex number of the underlying documents. The corpus contains proof-read texts and carefully checked translations by professional translators, does not contain duplicate sentences[4], is highly multilingual, large, available under a very permissive licence, and regularly updated. On the down-side, it contains only one type of texts, i.e. European law.

To create the DGT-UD corpus we took all the releases of JRC DGT, up to 2017. At the time of writing, the data for 2018 has also been released, which shows the usefulness of the resource as it will keep on growing, as well as the necessity of publishing updates of the UD-DGT corpus.

[1] https://www.clarin.si/kontext/, source available from https://github.com/ufal/lindat-kontext

[2] https://www.clarin.si/noske/, source available from https://nlp.fi.muni.cz/trac/noske

[3] https://ec.europa.eu/jrc/en/language-technologies/dgt-translation-memory

[4] This also means that included texts can contain gaps.

3 Annotation

In the process of compiling the DGT-UD corpus, the TMX encoding and file structure was first changed to consist of files for individual languages, while preserving the document structure and alignment between the translation units. The text of the language files was then annotated with UD-Pipe (Straka and Straková, 2017a), which performs text segmentation, morphosyntactic annotation, lemmatisation and dependency parsing, trained on the data from the Universal Dependencies (UD) project.[5]

For processing this corpus, we used off-the-shelf models (Straka and Straková, 2017b) trained on UD version 2.0 (Nivre et al., 2017). We managed to process 23 out of 24 languages, namely, Bulgarian, Croatian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hungarian, Italian, Irish (Gaelic), Latvian, Lithuanian, Polish, Portuguese, Romanian, Slovak, Slovenian, Spanish, and Swedish on all four annotation levels. The language not covered is Maltese, as it did not have UD 2.0 annotated data. But given that there now exists a (small) Maltese treebank in the UD project, we hope to add this language to the the next version of the corpus.

For languages where multiple models were available, we chose the largest models, assuming that those are based on the largest manually annotated resources, except for Swedish where the larger model (LinES) did not include morphosyntactic features. Given the size of the corpus, annotation is resource intensive, and took 98 hours on a 1.6 GHz, 24 CPU workstation.

4 DGT-UD on-line

We have made DGT-UD available on-line for exploration under the two CLARIN.SI concordancers and for download in the CLARIN.SI repository (Ljubešić and Erjavec, 2018) under the CC-BY licence.

The corpus is stored in the so called vertical format used by the Manatee back-end. This is is a tabular format (one line per token), but also supports XML-like structure tags. The vertical format (and Manatee) also have limitations, in particular not supporting (dependency) links between tokens and 1-n mappings between positional annotations. In CONLL-U the latter encode cases where one token is split into several syntactic words that are then further annotated.[6] In the conversion to vertical format, we therefore only retained the syntactic words. Each language is encoded as one file, e.g. `dgtud-bg.vert` for Bulgarian. The files have three structural attributes:

- *Text* with id, year of publication and release, e.g. `<text id="21993A1220(01)" year="-2003" release="DGT-TM-2007">`[7]
- *Anonymous block*, corresponding to the original JRC DGT aligned "sentence", and is successive number in the text, e.g. `<ab n="0">`
- *Sentence*, as annotated by UD-Pipe, i.e. `<s>`

The positional attributes (i.e. token annotations) of the files are:

- *word*: contains the token, and is by definition the first positional attribute, e.g. `förbindelser`; *lempos*: the lemma of the token with added part-of-speech, e.g. `förbindelse-n`;
- *tag*: the "PoS tag" of the token, which is given for convenience, and is the UD part-of-speech and its attribute values conjoined with an underscore; in cases where several attributes can have the same value, the shortened form of the attribute prefixed to the value, e.g. `ADJ_NumSing_GenMasc_DegPos_Def_Nom`;
- *pos*: the UD part-of-speech of the token, e.g. `ADJ`; *feats*: the UD features of the token conjoined with the pipe character, e.g. `Case=Nom|Definite=Def|Degree=Pos|Gender=Masc|Number=Sing`;

[5] `http://universaldependencies.org`

[6] Syntactic words are output by UD-Pipe for 7 out of the 23 DGT-UD languages, even though for the most languages are used infrequently.

[7] The year "-2003" indicates that the date of the document is 2003 or before, as the exact date is not given.

- *deprel*: the UD dependency relation,[8] e.g. `nmod`;
- *head_word*, *head_lempos*, *head_tag*, *head_pos*, *head_feats*: same as above, but for the token that is the head of the current token[9];
- *id*, *head_id*: the index of the token and its head. These attributes are not present in the on-line searchable corpus, as they are useless in that context, but are included in the source files, so the dependency link is not lost in the vertical format.

Apart from the vertical files, there are two further components of the corpus. First is the alignment file, giving the alignments of the `<ab>` elements,and the second the 23 Manatee registry files, specifying the metadata of the corpora.

As mentioned, the corpus is available both for download and under the CLARIN.SI concordancers, where it is already being used by translators at DGT. It is also mounted under the Czech CLARIN/LINDAT KonText[10] concordancer, thus giving users visiting the particular concordancer a better chance of finding and using the corpus.

5 Capacity for comparative linguistic analysis

The major features of the DGT-UD resource is that it is large and contains parallel texts annotated within the same formalism in all the 23 languages. For these reasons we hypothesised that DGT-UD is much more useful for comparative linguistic analyses than most other available resources, including the UD training data, which does have gold annotations, but is much smaller and not parallel or even comparable.

We tested our hypothesis by (1) representing each of the 21 languages of CLARIN member and observer countries (except for Norwegian which is not present in DGT-UD) as a probability distribution of (a) UPOS (Universal Part-of-Speech) trigrams and (b) UDEP (Universal Dependencies relations) trigrams (2) either on (a) a sample of 5,000 parallel sentences from the UD-DGT corpus or (b) from the UD training data. We cluster these language representations via Ward's hierarchical clustering (Ward, 1963) and present the result through dendrograms.

Figure 1 gives the results of the clustering, which show that (1) the structure obtained from DGT-UD data mostly follows our typological (syntactic) expectations and (2) the structure obtained from UD training data has a series of anomalies (Croatian clustered with Czech, and Slovak with Slovene; Hungarian clustered with English and Danish, and not with Finnish; German and Dutch clustered with Romance and not with other Germanic languages) which can be explained by variation in topic and genre in the UD training data.

6 Conclusions

The paper presented DGT-UD, an openly available, large, parallel and highly multilingual uniformly annotated and encoded parsebank, already available under several concordancers, as well as for download. An analysis of the resource indicates that the corpus is better suited for comparative linguistic analyses than the gold-annotated UD training data.

We would like to share the resource also with other (corpus exploration) service providers. In the first instance, these could be corpus analysis tools that also support treebank querying and display, such as PLM-TQ (Pajas et al., 2009). A very useful platform would be also is Multilingwis[11] (Volk, 2018), which, unlike (no)Sketch Engine, also shows (automatically) aligned words and phrases, further helping DGT and other translators.

[8] If the token is the root of the dependency tree, then its dependency label is '-'. In this case, all the head_ attributes are also hyphens.

[9] This allows searching (possibly with regular expressions) also over the annotations of the dependency head.

[10] `https://lindat.mff.cuni.cz/services/kontext`

[11] `https://pub.cl.uzh.ch/purl/multilingwis2`

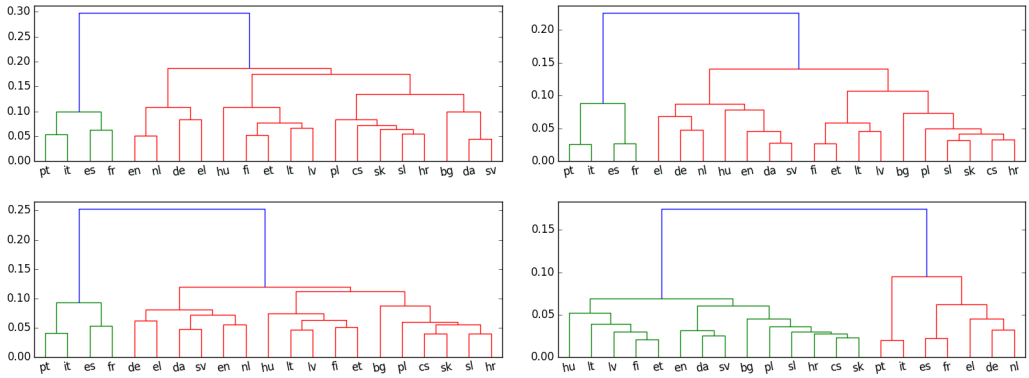


Figure 1: Dendrograms of the language clustering results on UPOS trigrams (upper) and UDEP trigrams (lower), on DGT-UD data (left) and UD training data (right).

References

Oliver Christ. 1994. A Modular and Flexible Architecture for an Integrated Corpus Query System. In *Proceedings of COMPLEX '94: 3rd conference on Computational Lexicography and Text Research*, pages 23–32, Budapest, Hungary.

Nikola Ljubešić and Tomaž Erjavec. 2018. *JRC EU DG Translation Memory Parsebank DGT-UD 1.0*. Slovenian language resource repository CLARIN.SI. `http://hdl.handle.net/11356/1197`.

Joakim Nivre et al. 2017. Universal dependencies 2.0. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. `http://hdl.handle.net/11234/1-1983`.

Petr Pajas, Jan Štěpánek, and Michal Sedlák. 2009. *PML Tree Query*. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. `http://hdl.handle.net/11858/00-097C-0000-0022-C7F6-3`.

Pavel Rychlý. 2007. Manatee/Bonito - A Modular Corpus Manager. In *1st Workshop on Recent Advances in Slavonic Natural Language Processing*, pages 65–70, Brno. Masarykova univerzita.

Ralf Steinberger, Andreas Eisele, Szymon Klocek, Spyridon Pilos, and Patrick Schlüter. 2012. DGT-TM: A freely available Translation Memory in 22 languages. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Ralf Steinberger, Mohamed Ebrahim, Alexandros Poulis, Manuel Carrasco-Benitez, Patrick Schlüter, Marek Przybyszewski, and Signe Gilbro. 2014. An overview of the European Union's highly multilingual parallel corpora. *Language Resources and Evaluation*, 48(4):679–707, Dec.

Milan Straka and Jana Straková. 2017a. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada, August. Association for Computational Linguistics.

Milan Straka and Jana Straková. 2017b. Universal dependencies 2.0 models for UDPipe (2017-08-01). LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. `http://hdl.handle.net/11234/1-2364`.

Martin Volk. 2018. Parallel Corpora, Terminology Extraction and Machine Translation. In *16ÖTT-Symposion. Terminologie und text(e)*, pages 3–14.

Joe H. Ward. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244.

DI-ÖSS - Building a digital infrastructure in South Tyrol

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

Abstract

This paper presents the DI-ÖSS project[1], a local digital infrastructure initiative for South Tyrol, which aims at connecting institutions and organizations that are working with language data. It shall serve to facilitate and increase data exchange, joint efforts in processing and exploiting data and the overall increase of synergies, and thus links to big European infrastructure initiatives. However, while sharing the overall objectives to foster standardization and increase efficiency and sustainability, on the implementation level a local initiative faces a different set of challenges. 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 digital language infrastructure on a local level; lay out the course of action; and depict the targeted mid- and long-term outputs of the project.

1 Introduction

In recent years, Digital Humanities saw the development of multiple infrastructure projects at the European level, among them CLARIN (Krauwer and Hinrichs, 2014) and DARIAH (Edmond et al., 2017) targeting researchers, and Europeana targeting the cultural heritage sector (Europeana Foundation, 2015).

The large field of smaller institutions, both in the public or 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, 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.

2 Motivation

Over the last decade the growing availability and use of digital language data in the humanities and the cultural sector and the need for coordination in standardization, preservation, exchange and reuse has led to a number of large-scale infrastructure initiatives being launched (see Section 1), dedicated to defining standards, best practices and technical solutions in order to foster accessibility, sustainability and shared use of data on the national and European level and even beyond. But small and local actors are prone to be less informed about these existing approaches and the opportunities that come with them and, due to their size, they often lack some of the skills and resources to make use of them. At the same time, it is typically at the local level that valuable resources of local culture are managed, and sometimes these resources can have particularities that can only be accommodated in an effective way at the local level.[2]

The DI-ÖSS project therefore has been devised at the local level of the Northern Italian region of South Tyrol with the two-fold aim to enable smaller actors in the language sector to learn about standards and technologies of the bigger initiatives and to create the conditions for joining efforts across single institutions and organizations, thus growing an interconnected language data ecosystem for South Tyrol. This

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

[2] E.g. the South Tyrolean variety of German is regionally bound and documented and studied first and foremost in the local context.

requires in the first step to raise awareness among local actors of the potential benefits of an infrastructure, to explain the necessary actions to become part of an infrastructure and finally introducing them to concrete standards and workflows that are common in the field of language resource management.

3 Project aim, approach and expected results

The DI-ÖSS project aims at creating a prototypical digital infrastructure for language data and services within the region of South Tyrol. By carrying out the preparatory work of gathering information and by implementing and piloting an infrastructure prototype for selected use cases, it serves to achieve prototypical results and gain an informed long-term perspective on the needs and feasibility of an infrastructure, thus allowing to evaluate the overall need for and potential of a local language infrastructure initiative. In contrast to the top-down approaches of the bigger projects like CLARIN or Europeana, DI-ÖSS is using a bottom-up strategy by *following the actors* (Latour, 2005), focusing on the local institutions in South Tyrol that are producing and collecting linguistic data.

The project is running for three years and is starting its efforts with partners and use cases related to the German variety of South Tyrol, with the perspective to expanding efforts to Italian in a future project. The results of all project phases will be compiled into a well-informed evaluation report and a concrete plan for a follow-up project for building a comprehensive and sustainable digital language infrastructure.

4 Implementation of the project plan

The project is organized along five broad phases. 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 their usage is collected from a wider set of institutions (Section 4.2). Afterwards, building on the previous phases' insights, specific use cases for each partner institution are determined and defined in greater depth (Section 4.3). Subsequently, a prototypical infrastructure is built by implementing the technical setup required to connect partners and their data as needed by each use case (Section 4.4). Finally, the infrastructure is piloted by employing it for each of the individual use cases and improving technical and conceptual aspects within the process (Section 4.5).

4.1 Consortium

The project consortium is made up of four public and private institutions, which represent relevant target groups with regard to the development, distribution and preservation of language data and services. These project partners have been chosen carefully to cover as wide a range of different approaches to language data as possible. The four project partners are:

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 Taking stock of the current state - 'Bestandsaufnahme'

An initial informational phase serves to gain an overview of the partner institutions, their specific workflows and the data they are handling. In order to gather this information a general set of questions is created, concerning four aspects: 1) *Technical data* cover the amount of digitized material, the kind of material (e.g. books, journals, web pages); the format of the material (e.g. PDF, HTML, XML); the

software used internally for working with the material, for archiving it, and for providing access to the public. 2) *Content data* refer to the kind of texts that are digitized, (e.g. scientific texts, literature, newspapers); their amount; and the language. 3) *Workflows* contains all internal processes that are concerned with the language data (especially acquisition, catalogization, and dissemination); user groups and their approximate sizes; and general aims of the institution. 4) *Copyright* is discussed with every institution independently, since this can be a complicated topic and the aim is to adapt to their specific needs. A smaller set of questions is used to also gather information about other institutions in the region that are working with language data, thus creating an overview of actors that can potentially join the infrastructure at a later stage. While collecting this information, it is always taken into consideration how it can be integrated into the larger CLARIN language resource infrastructure at a later stage, especially safeguarding VLO-interoperability.

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. Serving a genuine 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

For example, the use case of the library partner Teßmann addresses the task of researching the digitized content of cultural magazines with regard to topics discussed, personalities mentioned and locations that are referred to. The use case is approached through a close collaboration between computational linguists at Eurac Research and experts in literature and cultural studies at Teßmann library, in which automatic tools for language processing (i.e. detection and Named Entity Recognition) are applied to facilitate the search and information retrieval and analysis process. The use case will get implemented as a generic interface for annotating text and returning it in adaptable formats for the local search system.

4.4 Technical implementation

Each of the specialized use cases comes with specific requirements regarding the digital infrastructure, ranging from compatibility of data formats, over the definition of interfaces to the delivery of specialized services for data processing and annotation. The technical implementation phase is catering to the technical needs posed by each use case and, by extrapolating from the specific use cases, tries to anticipate further challenges to be expected in the long-term. In addition to prototypical infrastructure elements this project phase delivers technical specifications, evaluation reports, and best practice guidelines.

4.5 Piloting the infrastructure prototype

In the piloting phase the project partners are pursuing their work objectives by means of using the newly created DI-ÖSS infrastructure, therefore piloting its functionalities and technical implementation. During this phase insights on technical issues, conceptual shortcomings and possible functional extensions will be gathered and analyzed for improving the infrastructure prototype in the short-term and envisioning comprehensive extensions for development of a fully operable infrastructure in the long-term.

5 Challenges

Because of the special nature of a local infrastructure, the challenges it faces differ quite substantially from those encountered in larger infrastructure initiatives. Especially the fact that a lot of the potential institutions involved are relatively small, have few resources at their disposal and only have limited experiences with large-scale projects proves to be a key factor. The challenges can roughly be divided into conceptual, communicative and technical challenges.

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 advertisement of a local company, DI-ÖSS, which is meant as a pilot project, has to tighten its scope to more obvious "language institutions" like libraries, publishing houses and linguistic research institutions (see also section 4.1). Generally the focus was reduced to institutions that 1) are dealing with language data produced in South Tyrol, 2) are considering working with language data their main activity, and 3) are working with data that is 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.

5.2 Communicative challenges

Communicative challenges arose in the process of getting institutions interested in joining the project as it has proved difficult to properly communicate the scope and purpose of the project. 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 possible 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.

5.3 Technical challenges

This is another point where this small-scale infrastructure differs a lot from its larger counterparts. Many language partners in DI-ÖSS have very limited resources, both on the personnel as 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 more feasible for the bigger institutions involved in infrastructure projects like CLARIN. This means the DI-ÖSS infrastructure has to be built up 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.

6 Conclusions

This article reports on an initiative towards local infrastructure creation. It details the motivation and the plan of action for its implementation, while trying to be as much in sync with CLARIN standards as possible, and finally describes the specific challenges such a small-scale project is facing, starting even with raising awareness about infrastructures and their benefits and needs, and how they could be dealt with. 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 and the goal is to use it as a facilitator to establish a comprehensive and powerful digital language infrastructure in South Tyrol in the mid to long-term.

References

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.

Linked Open Data and the Enrichment of Digital Editions: the Contribution of CLARIN to the Digital Classics

Monica Monachini, Fahad Khan
ILC-CNR
Pisa, Italy
{name.surname}@ilc.cnr.it

Francesca Frontini
University Paul-Valéry
Montpellier, France
francesca.frontini@univ-montp3.fr

Anika Nicolosi
Dip. DUSIC
Parma University, Italy
anika.nicolosi@unipr.it

Abstract

Semantic Web technologies allow scholars in the humanities to make links and connections between the multitude of digitised cultural artifacts which are now available on the World Wide Web, thus facilitating the making of new scientific discoveries and the opening up of new avenues of research. Semantic Web and Linked Data technologies, by their very nature, are complex and require the adoption of a sustainable, long term approach that takes research infrastructures like CLARIN into consideration. In this paper, we present the case-study of a project (DEA) on an augmented digital edition of fragmentary Ancient Greek texts using Linked Data; this will highlight a number of the core issues that working in the Digital Classic brings up. We will discuss these issues as well as touching on the role CLARIN can play in the overall linked data lifecycle and in particular on humanities datasets.

1 Introduction

The Semantic Web and the Linked Data (LD) publishing paradigm allow the World Wide Web to be used as a means for data and knowledge integration wherein both documents/texts and other kinds of data are linked together. Publishing datasets as linked data has become an important way of sharing valuable structured information in a flexible and extensible manner across the Web, this is especially true in the humanities. Indeed the Semantic Web puts an emphasis on interoperability and interlinking between datasets; it can therefore be viewed as playing a similar role as that played in the past by artifacts such as the Rosetta Stone.

Informed by the success of the Semantic Web in other fields, efforts to use LD technologies to enable open access to cultural data are emerging. A crucial example is represented by the Digital Classics. Indeed the "Graph of the Ancient World", as it is often dubbed, is considered to be one of the most well developed parts of the Social Sciences and Humanities linked data cloud. It encompasses resources developed by classicists and archaeologists and is at least in part supported by dedicated tools[1]. The existence of large repositories of freely available digitised texts and other datasets, such as the Perseus Library (Crane, 2012) and gazetteers such as Pleiades, is encouraging a greater uptake of linked data technologies in the classics and especially for use in enriching digital editions. At the same time the advantages offered by the Semantic Web to publishers of data are only fully available to institutions who are able to make the technological investment and to develop tools for the (manual and automatic) linking of editions to datasets, and interfaces for the exploitation of such links for querying and research. Several ongoing projects are contributing towards the provision of the necessary support; for instance, the Recogito annotation system in Pelagios (Simon et al., 2015; Simon et al., 2017) aims to provide a graphic online environment for the manual annotation of digital editions with references to ancient places; automatic linking tools such as REDEN (Brando et al., 2016; Frontini et al., 2016), are built to cater specifically for the needs of the DH community (with TEI support and the possibility to

[1] We refer to (Isaksen et al., 2014) for a thorough description of this ecosystem and of the reason why this area was particularly susceptible to a Semantic Web turn.

customise datasets). At the same time the heavy demands that these approaches make with respect to computational capacity and long term support call for a stronger involvement on the part of existing research infrastructures.

The work described in this paper originates from the DEA (Digital Edition of Archilochus) project, which grew out of a collaboration between the University of Parma and the Istituzione di Linguistica Computazionale "A. Zampolli" in Pisa for the development of digital editions of Ancient Greek texts. In particular, we focus here on the issue of enriching such editions using Linked Data. Crucially we do not only aim to produce an enriched edition, but also to demonstrate the potential of the LD approach in the creation of openly available, scientifically reliable, linked, digital texts and its usefulness to a range of users including scholars, students and the wider public. The perfect environment in which to develop such a project is CLARIN which is able to provide advanced functionalities to store, develop, share, integrate and connect data, instruments and resources that are crucial in the field of the Digital Humanities. The project represents an important contribution of CLARIN-IT (Monachini and Frontini, 2016) towards ensuring that CLARIN has a greater impact on the field of digital classics.

2 Encoding the Fragments of Archilochus

The DEA project concentrates on a particularly complex type of text, namely texts of fragmentary authors, and builds on previous work in this area (Berti et al., 2014)[2]. The goal of the project is the digitisation, with critical and philological control, of the whole corpus of Archilochus' fragments[3]. The digitisation of Archilochus' texts will result in a complete, scientifically reliable, born digital, TEI/XML, edition (annotated and interoperable according to current standards, enriched with information from lexical and geographical knowledge bases) which will pay particular attention to scholarly users' needs and requirements, as well as to usability and portability. This new edition will both integrate already available digital resources as well as incorporating new datasets and resources: we will utilise and (where needed) develop digital methods that allow for the enrichment of the texts with different levels of linguistic and textual annotation (morpho-syntatic, semantic, etc.), as well as exploring the interaction between the text and already existing terminologies, ontologies, and lexicons available as Linked Open Data.

From the CLARIN perspective, this project provides an example of how it is possible to integrate and support the proof-reading, encoding and enrichment of the texts in a CLARIN repository. How can we build a platform which responds to the desiderata of the scholars themselves, facilitating the enrichment and exploration of digital editions with Semantic Web technologies? The investigation of such technological aspects should hopefully serve as a important test bed for future projects in the field of classical studies.

2.1 Enriching the Text with Data from the Semantic Web

A crucial issue for DEA is the identification of standards and best practices for the enrichment of our digital edition with structured knowledge from the Semantic Web[4]. In order to cater for the needs of classicists, both the enrichment and the exploration phases should be supported by easy to use interfaces, requiring little knowledge of the LD technicalities. A good example in this sense is the aforementioned Recogito annotation tool (developed by the Pelagios project), which allows for the tagging and searching of places, people, events with reference to geo-historical knowledge. Recogito handles input and output in TEI of a selection of Archilochus' fragments; and suggests possible referents to be linked to from the text, for instance ancient places from the Pleiades dataset of Greco-Roman antiquity places. Place

[2] See also the LOFT project http://www.dh.uni-leipzig.de/wo/lofts/.

[3] Archilochus of Paros (VII BC) is an important figure in archaic Greek Poetry, closely linked to Homer. However the corpus of Archilochus' surviving texts is currently not available in a user friendly and complete online digital edition. We know of 300 existing fragments, some of which were published in 2005 and so they are not yet included in all paper canonical editions. An updated edition of a selection of fragments in (Nicolosi, 2013).

[4] We shall not dwell here on the technicalities of representing the scholarly edition and its apparatus in TEI, and the representation of such editions via a web interface. Numerous projects currently provide tools to this end. See CDE https://dig-ed-cat.acdh.oeaw.ac.at/

mentions in the text can be later visualised on the map within the Recogito environment, thanks to the geospatial information contained in the dataset.

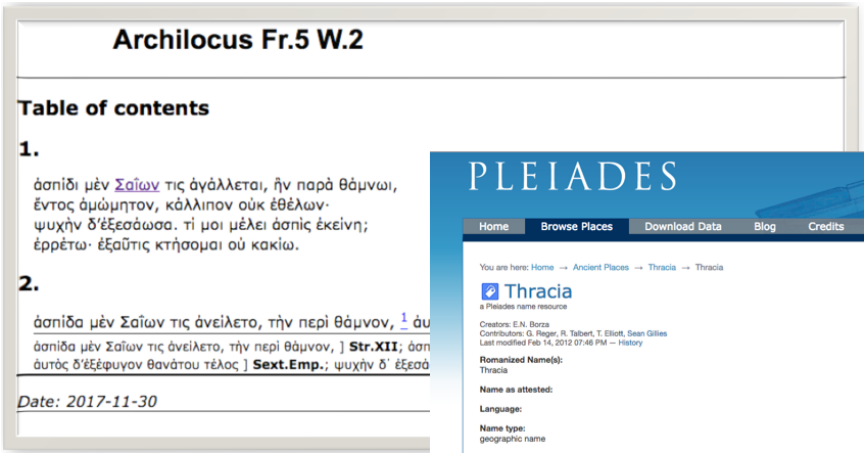


Figure 1: A Fragment from Archilocus linked to Pleiades.

Figure 2.1 (Archil. fr. 3 Nicolosi = fr. 5 West) for example shows how the Pleiades dataset provides a place reference for the Greek ethnonym Σαίων (line 1) which is an ancient clan of Thrace. Other datasets such as Geonames or DBpedia may be used to the same purpose, adding useful information on the ancient region and its inhabitants.

At the same time this type of LOD integration is quite basic, as it only allows us to add links via TEI tags by means of a "ref" attribute. One possibility which we intend to pursue in order to enrich our planned digital edition of Archilochus with Semantic Web datasets is the use of the Resource Description Framework in Attributes (RDFa). RDFa is a W3C recommendation which allows for RDF statements to be added to TEI-XML encoded text [5]. This approach has already been used in the Diachronic Spanish Sonnet Corpus (DISCO) project (Ruiz Fabo et al., 2018). In that instance RDFa attributes were used to link to LD resources such as VIAF[6] which provided biographical data, as well as for literary annotation.

Thanks to RDFa we can go beyond the scope of tools such as Recogito, and enrich our digital edition with more complex information, of a linguistic as well as an encyclopedic type. In particular linguistic resources such as the linked data version of the intermediate Liddell Scott lexicon (ILS) (Liddell and Scott, 1896; Khan et al., 2016) can be used to enrich each word token, immediately identifying the lemma and providing easy access to useful linguistic and philological data. This makes it simpler to add grammatical information, analyse sentences and correctly translate text. In addition, it is possible to identify *loci similes* that allow for the better explanation of the text and its exegesis. The goal is therefore to produce a linguistically rich, interoperable and researchable, text analysis.

2.2 Syntactic annotation

In addition to the enriched critical edition, the project plans to transform the corpus into a Treebank, thereby making an important contribution to the already existing Ancient Greek Treebank[7], this will enable the improvement of currently existing parsers for ancient Greek with regards to their performance on fragmentary texts. This newly created resource will be deposited in CLARIN and integrated into the Tündra exploration system[8]. In particular, in order to allow for the deposit of the resource, all texts will be provided with adequate metadata so as to become findable via the CLARIN metacatalogue (VLO)

[5] https://rdfa.info/
[6] https://viaf.org/
[7] http://perseusdl.github.io/treebank_data/
[8] https://www.clarin-d.net/en/tundra

and searchable online.

3 Conclusion

Classical studies are at the forefront of the adoption of Semantic Web technologies within the humanities. However a deeper investigation into the standards for creating enriched editions is required, in order to identify best practices. The DEA project intends to make a contribution to the interaction between literary and philological studies, the classics in particular, and LD. E-infrastructures appear to be the perfect framework through which to spread knowledge about good practices related to a discipline.

The project will address the need for an integrated and standardised platform for the consultation, updating and searching of textual materials and structured information. In addition to this the textual analysis, with the use of a treebank, and translation of each fragment will be a important development for study and learning ancient Greek. In our presentation, we shall discuss the XML-TEI encoding of the Archilocus fragments in depth as well as the solution proposed for linking them to existing datasets. Finally, we will show how CLARIN can provide significant developments in the field of Digital Classics by offering opportunities to store, develop, share and access datasets of enriched digital editions.

References

Monica Berti, Bridget Almas, David Dubin, Greta Franzini, Simona Stoyanova, and Gregory R. Crane. 2014. The Linked Fragment: TEI and the Encoding of Text Reuses of Lost Authors. *Journal of the Text Encoding Initiative*, (Issue 8), December.

Carmen Brando, Francesca Frontini, and Jean-Gabriel Ganascia. 2016. REDEN: Named Entity Linking in Digital Literary Editions Using Linked Data Sets. *Complex Systems Informatics and Modeling Quarterly*, 0(7):60–80, July.

Gregory Crane. 2012. The Perseus Project. In *Leadership in Science and Technology: A ReferenceHandbook*, pages 644–652. SAGE Publications, Inc., Thousand Oaks.

Francesca Frontini, Carmen Brando, and Jean Gabriel Ganascia. 2016. REDEN ONLINE: Disambiguation, Linking and Visualisation of References in TEI Digital Editions. In *Digital Humanities 2016: Conference Abstracts*, pages 193–197, Kraków. Jagiellonian University & Pedagogical University.

Leif Isaksen, Rainer Simon, Elton T.E. Barker, and Pau de Soto Cañamares. 2014. Pelagios and the Emerging Graph of Ancient World Data. In *Proceedings of the 2014 ACM Conference on Web Science*, WebSci '14, pages 197–201, New York, NY, USA. ACM.

Fahad Khan, Francesca Frontini, Federico Boschetti, and Monica Monachini. 2016. Converting the Liddell Scott Greek-English Lexicon into Linked Open Data using lemon. In *Digital Humanities 2016: Conference Abstracts*, pages 593–596, Kraków. Jagiellonian University & Pedagogical University.

Henry George Liddell and Robert Scott. 1896. *An intermediate Greek-English lexicon: founded upon the seventh edition of Liddell and Scott's Greek-English lexicon*. Harper & Brothers.

Monica Monachini and Francesca Frontini. 2016. CLARIN, l'infrastruttura europea delle risorse linguistiche per le scienze umane e sociali e il suo network italiano CLARIN-IT. *IJCoL - Italian Journal of Computational Linguistics*, 2(2):11–30.

Anika Nicolosi. 2013. *Archiloco: elegie*. Pàtron, Bologna. Google-Books-ID: 9uj5oAEACAAJ.

Pablo Ruiz Fabo, Clara Isabel Martínez Cantón, and José Calvo Tello. 2018. Disco: Diachronic spanish sonnet corpus.

Rainer Simon, Elton Barker, Leif Isaksen, and Pau de Soto Cañamares. 2015. Linking early geospatial documents, one place at a time: annotation of geographic documents with Recogito. *e-Perimetron*, 10(2):49–59.

Rainer Simon, Elton Barker, Leif Isaksen, and Pau De Soto CaÑamares. 2017. Linked Data Annotation Without the Pointy Brackets: Introducing Recogito 2. *Journal of Map & Geography Libraries*, 13(1):111–132, January.

How to use DAMESRL: A framework for deep multilingual semantic role labeling

Quynh Ngoc Thi Do **Artuur Leeuwenberg** **Geert Heyman** **Marie-Francine Moens**
Department of Computer Science, KU Leuven, Belgium
{quynhngocthi.do,tuur.leeuwenberg,geert.heyman,sien.moens}@cs.kuleuven.be

Abstract

This paper presents DAMESRL, a flexible and open source framework for deep multilingual semantic role labeling. It provides flexibility in its model construction in terms of word representation, sequence representation, output modeling, and inference styles and comes with clear output visualization. The framework is available under the Apache 2.0 license[1].

1 Introduction

Semantic role labeling (SRL) is an essential natural language processing (NLP) task which aims at identifying "Who did What to Whom, and How, When and Where" in a sentence (Palmer et al., 2010). In the last few years, a number of neural mechanisms have been used to train end-to-end SRL models that do not require task-specific feature engineering. Zhou and Xu (2015) introduced the first deep end-to-end model for SRL by stacking multiple Bi-LSTM layers along with a conditional random field (CRF) layer. This architecture was then simplified by He et al. (2017), who proposed the use of a highway Bi-LSTM network without CRF layer. More recently, Tan et al. (2018) replaced the common Bi-LSTM architecture by a self-attention network resulting in better results and faster training. Although generally successful in providing accurate scores on English benchmark data, a common criticism of these systems is the lack of considering the influence of multilingual aspects on the optimal model structure. It is still an open issue whether a model behaves similarly on languages which differ in characteristics and the available amount of training data.

DAMESRL facilitates exploration and fair evaluation of new SRL models for different languages by providing flexible neural model construction on different modeling levels, the handling of various input and output formats, and clear output visualization. Beyond the existing state-of-the-art models (Zhou and Xu, 2015; He et al., 2017; Tan et al., 2018), we exploit character-level modeling, beneficial when considering multiple languages. This paper gives practical guidance on how to use our framework and reports its performance evaluation on English, German and Arabic.

2 System Overview

2.1 Task Definition

Given a sentence $(w_1, w_2, \ldots, w_n)$, the SRL task is divided into two sub-tasks: (1) Identifying predicates. (2) Predicting a sequence $(l_1, l_2, \ldots, l_n)$ of semantic labels for each predicate w_p. Each label l_i, which belongs to a discrete set of PropBank BIO tags, is the semantic tag corresponding to the word w_i in the semantic frame evoked by w_p. Here, words outside argument spans have the tag $\mathbf{O}$, and words at the beginning and inside of argument spans with role r have the tags $\mathbf{B_r}$ and $\mathbf{I_r}$, respectively. For example, the sentence "the cat chases the dog ." should be annotated as "the$_{\mathbf{B-A0}}$ cat$_{\mathbf{I-A0}}$ chases$_{\mathbf{B-V}}$ the$_{\mathbf{B-A1}}$ dog$_{\mathbf{I-A1}}$.$_{\mathbf{O}}$".

2.2 Architecture

DAMESRL's architecture (see Fig. 1) facilitates the construction of models that prioritize certain language-dependent linguistic properties, such as the importance of word order and inflection, or that adapt to the amount of available training data. The framework is implemented in Python 3.5 using TensorFlow, and can be used to train new models, or make predictions with the provided trained models.

[1] https://liir.cs.kuleuven.be/software_pages/damesrl.php.

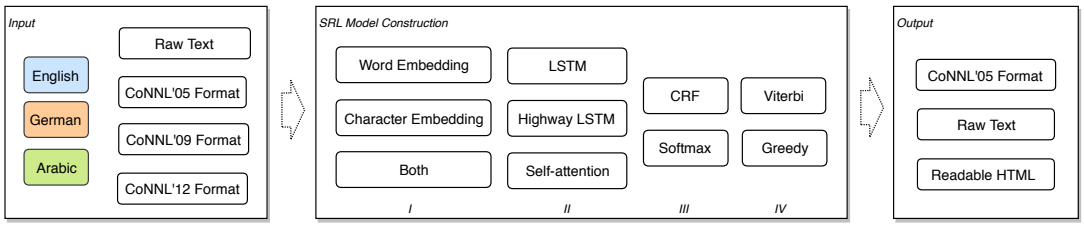


Figure 1: Schematic overview of the DAMESRL architecture from input to output.

2.2.1 Input and Output

The format DAMESRL uses to specify sentences labeled with predicates and semantic roles is a shortened version of the CoNLL'05 format, which only contains the Words, Targets and (possibly) Props columns[2]. This format is used to represent the training input as well as the prediction output. Furthermore, DAMESRL provides an HTML format to visualize the system output directly in a web browser (as in Fig. 2) and allows plain text input for prediction.

2.2.2 Model Construction Modules

As can be seen in Fig. 1, the framework divides model construction for both predicate identification and semantic label prediction, in four phases: (I) word representation, (II) sentence representation, (III) output modeling, and (IV) inference.

Phase I: For predicate identification, the word representation of a word w_i consist of two optional concatenated components: a word-embedding and a character representation. For semantic label prediction, a Boolean indicating if w_i is the predicate of the semantic frame (w_p) is used as an extra component. DAMESRL provides a Bi-LSTM network to learn character-level word representations helping for languages where important SRL cues are given through inflections, such as case markings in German and Arabic. Despite the foreseen importance, character-level embeddings have not been used in previous work (Zhou and Xu, 2015; He et al., 2017; Tan et al., 2018).

Phase II: As core sequence representation component, the user can choose between a self-attention encoding (Tan et al., 2018), a regular Bi-LSTM (Hochreiter and Schmidhuber, 1997) or highway LSTM (Zhang et al., 2016; He et al., 2017).

Phase III: The user can choose between a regular softmax to compute model probabilities, or a linear chain CRF as proposed by (Zhou and Xu, 2015), which can be useful for languages where word order is an important SRL cue, such as English, or when less training data is available (shown in Section 3).

Phase IV: The inference phase provides two inference options from the computed model probabilities including greedy prediction and Viterbi decoding. The outputs are binary labels and PropBank labels for predicate identification and semantic label prediction, respectively.

3 Performance Evaluation

To evaluate our framework, and show the benefits of choosing certain model components, we construct five SRL models[3]: HLstm, Char, CRFm, Att, and CharAtt, whose configurations are shown in Tab. 1.

The selected models are evaluated in three languages: English, German and Arabic using the standard CoNLL'05 metrics[4].

In Tab. 2-3, we compare the five models on English, German and Arabic. The proposed CharAtt outperforms all the other models in almost all the cases except for the English out-of-domain dataset. As

[2]http://www.lsi.upc.edu/srlconll/conll05st-release/README

[3]As is common practice we evaluate the SRL models with groundtruth predicate labels, the predicate identification models are only used for prediction.

[4]In CoNLL 2005 evaluation setting, gold predicates are used. To evaluate predicate identification, we train three CharAtt models for English, German and Arabic and obtain the F1 scores of 95.64%, 75.95%, and 94.80%, respectively. The poor performance of German dues to the data sparsity.

	HLstm	Char	CRFm	Att	CharAtt
Word Emb.	✓		✓	✓	
Word + Character Emb.		✓			✓
Highway LSTM	✓	✓	✓		
Self-Attention				✓	✓
Softmax	✓	✓		✓	✓
CRF			✓		

Table 1: Configurations of experimental models.

Model	Arabic		German		
	Dev	Eval	Dev	Ood	Eval
HLstm	45.7	46.3	67.1	56.4	67.6
Char	50.7	46.7	67.8	54.6	67.6
CRFm	49.2	49.9	67.3	54.6	65.1
Att	49.2	48.3	71.2	55.7	71.7
CharAtt	**56.5**	**55.2**	**74.3**	**57.3**	**73.5**

Table 2: F1 results on CoNLL'12 Arabic[5] and CoNLL'09 German data[6].

Model	5% Data			Full Data		
	Dev	Ood	Eval	Dev	Ood	Eval
Lstm + CRF (Zhou and Xu, 2015)	-	-	-	79.6	69.4	82.8
HLstm (He et al., 2017)	-	-	-	81.6	72.1	83.1
Att (Tan et al., 2018)	-	-	-	83.1	**74.1**	84.8
HLstm-ours	62.8	54.3	64.9	82.0	71.9	83.1
Char	64.8	55.2	65.8	82.2	72.5	83.4
CRFm	**65.8**	**57.5**	**67.0**	81.7	70.9	83.5
Att-ours	57.4	51.7	59.6	83.2	73.7	84.8
CharAtt	58.2	52.4	60.7	**83.5**	73	**84.9**

Table 3: F1 results on CoNLL'05 English data compared to other state-of-the-art deep *single* models.

shown in Tab. 3, our implementation achieves competitive performance to other state-of-the-art systems for English. To the best of our knowledge, we report the first SRL results for German and Arabic without using linguistic features. In general, we find that using character embeddings improves the performance of HLstm and Att, although at a cost of increased processing time. Among the three languages, the gain by using character-level representations is larger when processing German or Arabic compared to English. This may be because of the smaller training set size compared to the vocabulary size and predicate patterns. Moreover, many grammatical cases, which are very strong predictors for semantic roles, are explicitly marked through the use of inflection in German and Arabic. In order to evaluate the influence of the training size on model performance, we train the models on a random sample of 5% of the CoNLL'05 English training data as in Tab. 3. Interestingly, in this case, the attention models suddenly perform worst while the CRFm reaches the top rank. We consider this as evidence that a large training size is crucial for attention models. In contrast, the CRFm model exploits not only the input sequence but also the output dependencies, when it computes the output probabilities. We can see that this is beneficial for a strict word order language such as English especially when less training data is available.

4 Practical Guidance

4.1 End Users

The framework comes with several pre-trained models that work out-of-the-box, which can be easily tested through a web service. To start the web server, run `python www/proxy_server.py` in the terminal from the project root directory. This will instantiate pre-trained models as specified in `default_server_config.yml`. Users that want to work with their own models (see Section 4.2) can start the web server with their own configuration: `python www/proxy_server.py ---config custom_server_config.yml`. By default the server listens to port 8080, but a different port can be configured using the `python www/proxy_server.py --port $number`. A screenshot from the web service when accessed with a web browser is shown in Fig. 2.

[5]To the best of our knowledge, there has not been any reported result using CoNLL'05 metrics on this data. Pereyra et al. (2017) only report the Precision score of the argument classification instead of the standard CoNLL'05 metrics.

[6]To the best of our knowledge, there has not been any reported result using CoNLL'05 metrics on this data.

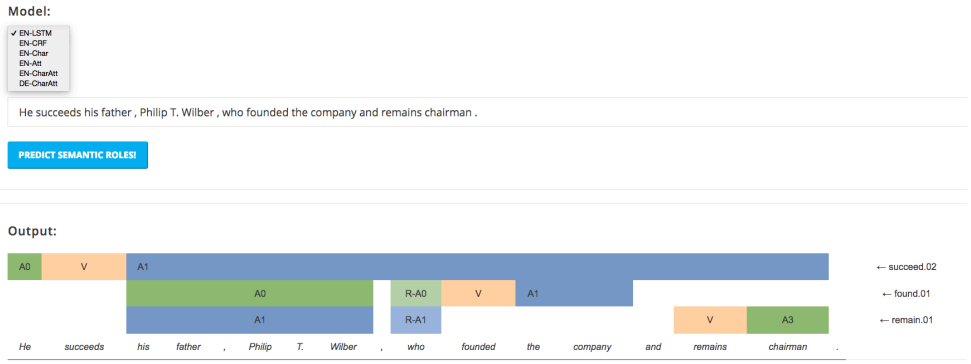


Figure 2: Screenshot of the DameSRL web-service included in the framework.

4.2 Developers

DameSRL is implemented in Python3.5 and Tensorflow, which are both easily accessible and widely used. This makes it easy for developers to flexibly change the code, add new models, or change existing models, allowing them to conduct their own experiments in the same setup. New models, or model modifications can be implemented in `dame/srl/DSRL.py` and by changing the configuration file reader accordingly in `dame/srl/ConfigReader.py` . New neural network components can be added in `dame/core/nn/`. Data pre-processing is handled by `dame/srl/DataProcessor.py`, and reading, writing, and visualization of the various formats by the modules in `dame/core/io`.

5 Conclusions

We introduced DAMESRL, an open source SRL framework which provides flexible model construction using state-of-the-art model components, handles various input and output formats, and which comes with clear output visualization. We have shown that the flexible model construction provided by the framework is crucial for exploring good model structures when considering different languages with different characteristics, especially in the case when training data is limited. DAMESRL is available under the Apache 2.0 license.

References

Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. 2017. Deep semantic role labeling: What works and whats next. In *Proceedings of ACL*, volume 1. ACL.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8).

Martha Palmer, Daniel Gildea, and Nianwen Xue. 2010. *Semantic role labeling*, volume 3. Morgan & Claypool Publishers.

Gabriel Pereyra, George Tucker, Jan Chorowski, Lukasz Kaiser, and Geoffrey E. Hinton. 2017. Regularizing neural networks by penalizing confident output distributions. In *Proceedings of the ICLR Workshop*.

Zhixing Tan, Mingxuan Wang, Jun Xie, Yidong Chen, and Xiaodong Shi. 2018. Deep semantic role labeling with self-attention.

Yu Zhang, Guoguo Chen, Dong Yu, Kaisheng Yaco, Sanjeev Khudanpur, and James Glass. 2016. Highway long short-term memory RNNs for distant speech recognition. In *Proceedings of ICASSP*. IEEE.

Jie Zhou and Wei Xu. 2015. End-to-end learning of semantic role labeling using recurrent neural networks. In *Proceedings of ACL-IJCNLP*, volume 1. ACL.

Speech Recognition and Scholarly Research: Usability and Sustainability

Roeland Ordelman
Research & Development
Netherlands Institute for Sound and Vision
The Netherlands
rordelman@beeldengeluid.nl

Arjan van Hessen
Human Media Interaction
University of Twente
The Netherlands
a.j.vanhessen@utwente.nl

Abstract

For years we have been working on speech recognition (ASR) as a tool for scholarly research. The current state-of-the-art can be useful for many scholarly use cases focusing on audiovisual content, but practically applying ASR is often not so straightforward. In the CLARIAH Media Suite, a secured online portal for scholarly research for audiovisual media, we solved the most important hurdles for the practical deployment of ASR by focusing on usability and sustainability aspects.

1 Introduction

Since many years scholars have been promised the virtues of automatic speech recognition (ASR) to be successfully deployed in their research, relieving them from the burden of manual *transcription* of the spoken word, and increasing the efficiency of *discovery* in large audiovisual collections by combining ASR with Spoken Document Retrieval (SDR) and linking (Linked Data, Recommendation) (Gustman et al., 2002; de Jong et al., 2008; Boves et al., 2009; Ordelman et al., 2015a). Although in the past decade, many research projects have been funded based on the potential of ASR and SDR for scholarly research, automatic speech recognition still does not seem to be a tool that scholars can easily deploy in their research. In our project we focus therefor explicitly on those aspects that are crucial for scholars with respect to actually deploying the technology: (i) *Usability* of the technology, that refers to working with a speech recognition engine itself, and using the output of it, and also (i) *Sustainability*, that refers to the longer-term availability of the technology for scholars, with state-of-the-art performance, maintained, updated, and accessible.

In this paper, we will focus especially on aspects of usability of speech recognition as deployed in the context of the CLARIAH Media Suite, a portal focusing among others on working with large audiovisual collections in scholarly research. At the conference, we will go into more detail on sustainability aspects, the choices we have made, and future plans.

2 Requirements

In the past decade, we have been working with humanities scholars –especially Oral Historians– on topics related to speech recognition in a variety of projects such as CHoral (Heeren et al., 2009), Verteld Verleden (Ordelman and de Jong, 2011), Oral History Today (Kemman et al., 2013), and more recently CLARIAH (Ordelman et al., To appear). These projects led, first of all, to a better understanding and collaboration between humanities scholars and ICT-researchers and developers (De Jong et al., 2011). Moreover, the projects provided a better insight into the variety of uses of digital collections, how a specific tool such as speech recognition could play a role here, and finally also, how speech recognition should be provided as a tool in a research infrastructure.

Below we will list some of the key requirements of scholars with respect to usability of speech recognition in scholarly research. We used these requirements as a starting point for the implementation of ASR in the CLARIAH infrastructure.

2.1 Speech recognition and quality requirements

One requirement that is prominent in the discussion about the usability of automatic speech recognition, is quality. Quality typically translates to word error rate (WER) or its counterpart word accuracy: the number of errors –in terms of word substitutions and words that are deleted or inserted by the speech recognition system– divided by the total number of words spoken. Scholars refer to this quality from the perspective of two scholarly primitives (Unsworth, 2000; Blanke and Hedges, 2013): (i) from the perspective of *annotation*, creating verbatim transcripts of –typically personally collected– spoken data (such as interviews) with the use of speech recognition, and (ii) from the perspective of *discovery*, searching audiovisual collections using indexed speech transcripts. Although these perspectives could coincide, requirements with respect to quality differ.

2.1.1 Quality for annotation

First of all, as the multi-semiotic nature of audiovisual data ads dimensions for inquiry that do not exist in written text (Goldman et al., 2005), scholars working with A/V are used to create manual transcriptions and annotations based on their codebooks. The creation of *highly accurate* manual transcriptions that account for all speech events as well as other metadata (such as speaker identities and changes) or codes from a scholar's codebook (e.g., based on intonation of a speaker's voice), can take up to 50 times real-time, depending on the nature of the data and the level of detail (Barras et al., 2001). Support from a tool such as automatic speech recognition could therefore be very helpful, especially as the technology also provides time-labels that link the transcript directly to the locations in the A/V data.

However, as a support tool for a scholarly primitive, speech recognition has a number of requirements: (i) it should fit into the transcription and annotation workflow and tools that scholars are used to, (ii) it should at least be able to provide *approximately correct* transcriptions[1], and (iii) it should facilitate error correction when the quality is (partly) insufficient (e.g., with respect to proper names or locations that are 'out-of-vocabulary[2]').

There is an obvious trade-off between the time needed to correct an error-full transcript and generating a transcript from scratch. In studies (Gaur et al., 2016), a threshold around WER of 30% is observed: if the WER measure of the speech recognition system is much greater than 30%, humans are better off writing everything from scratch rather than getting help from automatic speech recognition. The design of the transcription and annotation tools should reflect this trade-off and provide means to increase productivity, for example by implementing auto-complete functions (e.g., based on the lattices or domain-specific vocabularies), or by detecting that a transcript is above the 30% threshold and stop providing it in such cases. As one needs ground-truth data to compute WER, the latter approach is however difficult to implement. The same holds for an alternative option, using confidence scores as an estimate of the WER.

2.1.2 Quality for discovery

The other perspective on speech recognition quality is related to discovery using a search engine and does not only concern personally collected data but also large A/V collections that are sitting in cultural heritage institutions, archives, libraries, and knowledge institutions. The metadata that are available for these collections are typically sparse and limited to summaries. The goal of using speech recognition is "bridging" the semantic gap (Smeulders et al., 2000): to decode the information in the audio (speech) to semantic representations (words) and to increase the chances for discovery.

Quality of the speech recognition here relates to measures in the domain of information retrieval such as precision and especially, *recall*: discovering the interesting bits that were hidden at first. We know from SDR experiments that there is a near-linear relationship between WER and search performance (Garofolo et al., 2000), but also, that even if transcriptions are *approximately correct* –say with a threshold of 30% WER again– such transcriptions provide a useful basis for searching.

[1] Some scholars even require transcripts enhanced with appropriate punctuation and capitalization

[2] Out-of-vocabulary (OOV) refers to the fact that speech recognition systems need to "know" a word (have a word in its vocabulary) before it can be recognized.

However, scholarly discovery may in cases depend on a speech recognition system *not* making a notorious type of error: substituting a spoken word for another word as the spoken word is *out-of-vocabulary* (OOV). Typically this happens when applying speech recognition trained on a general domain (e.g., broadcast news) to a specific topic, for instance, a topic discussed during an Oral History interview. Especially names and locations, but also content words related to a special topic, will be OOV and will consequently not be recognized, and in turn, never be found during searching.

When deploying speech recognition in the context of large archives with multiple collections and an intrinsically large variation in topics, hence semantic variation, the OOV problem (or language model mismatch), or from a discovery point of view, the recall problem, needs to be taken into account. However, there is not a straightforward solution for this. When scholars can control the speech recognition vocabulary –for example when running speech recognition on their personally collected data (see section 2.2.1)– adding domain-specific words could be an option, assuming that the required provision of the pronunciation[3] of the added words can also be organized. However, when speech processing audiovisual collections on a very large scale (see section 2.2.2), dynamic adaptation of the speech recognition vocabulary (and language model) to the contents of the individual items or collections, is hard to implement efficiently.

2.2 Deployment of speech recognition

Discussing quality requirements assumes that speech recognition can be made available for scholars, either as a tool or service in itself or as output that scholars can use for distant reading or close reading purposes. But what is "available" exactly?

2.2.1 Speech recognition as a service

Especially for scholars that have their personally collected audiovisual data such as interviews, having high-quality speech recognition available as a tool they can use, has been the holy grail for years. Although speech recognition is becoming mainstream, for the purpose of the transcription of audiovisual data it is not yet a commodity that a scholar can just download, install, and use. Or, from a speech recognition as a service point of view: using cloud based APIs. There are various repositories and open-source speech recognition toolkits available, but the installation of such kits requires expert skills. Cloud solutions and APIs are not available for every language, typically do not have options for adaptation, are not always suitable for (large) batch processing, and, last but not least, impose serious issues in terms of data privacy and IPR.

Providing scholars with an online service within the closed environment using a federated authentication mechanism (SURFConext[4]), to interact with a speech recognition system may be the best approach, but there are many questions that need to be addressed before such a service can be deployed, such as: (i) Who is/are responsible for maintaining the speech recognition engine that is behind the service. The responsible(s) should have expert knowledge in speech recognition; (ii) Who is responsible for maintaining the service that connects the scholar with the underlying speech recognition engine. The responsible may also need to manage a computer cluster to accommodate for either many users or large volumes of data; (iii) Which speech recognition toolkit/engine to choose, depending on state-of-the-art performance, language support, but also available local expertise with the toolkit; (iv) What functionality should the service accommodate for, ranging from various types of input/output procedures (e.g., file upload/download, bulk processing, status information), and input and output formats (e.g., both audio an video), to special options such as adapting the speech recognition vocabulary to the contents of the data (see also section 2.1.2).

An important additional requirement for a speech recognition service for scholarly use is that it can assure that it is designed in such a way that it incorporates privacy requirements in a way that makes privacy violations unlikely to occur (van den Hoven et al., 2016).

[3] A pronunciation lexicon would have an entry that states that "CLARIN" is pronounced as /k l E r I n/.

[4] `https://www.surf.nl/en/services-and-products/surfconext/index.html`

2.2.2 Large scale speech recognition

When the focus of scholars is on access to very large (+500K hours) A/V collections that are sitting in cultural heritage institutions, archives, libraries, and knowledge institutions, a speech recognition service for individual scholars is not feasible. First of all, (bulk) access to the audiovisual sources of institutional collections is something that needs to be organized on an infrastructural level, taking care of workflows that are capable of processing large amounts of data but also taking care of security and privacy aspects. In addition, to process these amounts of data efficiently, dedicated machinery (e.g., local or cloud-based computer clusters) is needed. Finally, 'big data' speech processing is a task that has its specific dynamics with respect to robustness, latencies, process management, and storage of intermediate data. Implementing adaptive workflows to address the OOV problem as discussed in section 2.1.2, is an endeavour in this context.

Additional issues with large-scale speech processing are keeping track of provenance information (which version of a speech recognition system was used) and decisions about what information to keep (e.g., 1-best versus lattices).

A hybrid approach that is in between speech recognition as a service and bulk processing, is to make an estimation of the most important collections in the infrastructure (e.g., for media scholars interested in radio and television broadcasts, news and actualities are most interesting), process these collections in bulk, and allow individual scholars to make requests for the processing of specific collections or batches.

2.3 Using the output of speech recognition

Given requirements on quality of speech recognition transcripts and on the methods to obtain these transcripts, the next question is what scholars actually want to do with the transcripts. The main purpose of deploying speech recognition in scholarly research is discovery, especially when deployed in the context of large audiovisual archives. We distinguish two types of discovery: one that uses standard search methods, and one that uses exploratory type of approaches, such as content-based recommendation (Yang and Meinel, 2014), video hyperlinking (Ordelman et al., 2015b), and linked data technology. For the exploratory type of discovery, various additional processing steps can be thought of that use the speech transcripts as input, such as word cloud generation, named entity extraction or topic modelling.

Next to discovery, time-labelled speech transcripts can also be useful for the scholarly primitive "comparing" as it allows scholars to go back and forth between segments within and across items.

3 Speech recognition in a research infrastructure

At the conference, we will show how speech recognition is implemented in the CLARIAH infrastructure and MediaSuite (see screenshots in Figures 1 and 2) given the requirements of scholars discussed above. Compared to a decade ago, the quality of speech recognition systems has improved substantially thanks to the revival of the use of neural networks (Graves et al., 2013). Especially in less optimal conditions, the accuracy of systems increases with large steps (Hinton et al., 2012).

We are using the open-source KALDI[5] (Povey et al., 2011) speech recognition toolkit that supports deep neural nets, together with the LIUM speech diarization toolkit (Rouvier et al., 2013). Dutch models have been developed at University of Twente using the Spoken Dutch Corpus (Oostdijk, 2000) and a large corpus of text data from a variety of sources (Ordelman et al., 2007). The resulting "Kaldi_NL" instance, has a lexicon of around 250K words and with NNET3-TDNN-LSTM models (Peddinti et al., 2018). Its performance is state-of-the-art with around 10% WER ($<$ 0.5xRT) tested on the NBest BN-NL benchmark set (Kessens and Leeuwen, 2007).

KALDI_NL is running as a sustainable service in the CLARIAH infrastructure at the Netherlands Institute for Sound and Vision, one of the CLARIAH Centers. Until now, we have been processing 350K hours of audiovisual content via the High Performance Computing data infrastructure for science and industry, SURFSara.

[5] `http://kaldi-asr.org/`

Figure 1: Browsing transcripts

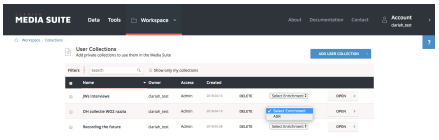


Figure 2: Activating ASR

References

Claude Barras, Edouard Geoffrois, Zhibiao Wu, and Mark Liberman. 2001. Transcriber: development and use of a tool for assisting speech corpora production. *Speech Communication*, 33(1-2):5–22.

Tobias Blanke and Mark Hedges. 2013. Scholarly primitives: Building institutional infrastructure for humanities e-science. *Future Generation Computer Systems*, 29(2):654–661.

Lou Boves, Rolf Carlson, Erhard Hinrichs, David House, Steven Krauwer, Lothar Lemnitzer, Martti Vainio, and Peter Wittenburg. 2009. Resources for speech research: Present and future infrastructure needs. In *10th Annual Conference of the International Speech Communication Association [Interspeech 2009]*, pages 1803–1806.

F M G de Jong, D W Oard, W F L Heeren, and R J F Ordelman. 2008. Access to recorded interviews: A research agenda. *ACM Journal on Computing and Cultural Heritage (JOCCH)*, 1(1):3:1–3:27, jun.

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.

John S. Garofolo, Cedric G. P. Auzanne, and Ellen M. Voorhees. 2000. The trec spoken document retrieval track: A success story. In *Content-Based Multimedia Information Access - Volume 1*, RIAO '00, pages 1–20, Paris, France, France. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE.

Yashesh Gaur, Walter S Lasecki, Florian Metze, and Jeffrey P Bigham. 2016. The effects of automatic speech recognition quality on human transcription latency. In *Proceedings of the 13th Web for All Conference*, page 23. ACM.

Jerry Goldman, Steve Renals, Steven Bird, Franciska De Jong, Marcello Federico, Carl Fleischhauer, Mark Kornbluh, Lori Lamel, Douglas W Oard, Claire Stewart, et al. 2005. Accessing the spoken word. *International Journal on Digital Libraries*, 5(4):287–298.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*, pages 6645–6649. IEEE.

Samuel Gustman, Dagobert Soergel, Douglas Oard, William Byrne, Michael Picheny, Bhuvana Ramabhadran, and Douglas Greenberg. 2002. Supporting access to large digital oral history archives. In *Proceedings of the second ACM/IEEE-CS joint conference on Digital libraries - JCDL '02*, page 18, New York, New York, USA. ACM Press.

W.F.L. Heeren, Laurens Bastiaan van der Werff, Franciska M.G. de Jong, Roeland J.F. Ordelman, T. Verschoor, Adrianus J. van Hessen, and Mies Langelaar. 2009. Easy listening: Spoken document retrieval in choral. *Interdisciplinary science reviews*, 34(2-3):236–252, 9. 10.1179/174327909X441135.

G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury. 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, Nov.

Max Kemman, Stef Scagliola, Franciska de Jong, and Roeland Ordelman. 2013. Talking with scholars: Developing a research environment for oral history collections. In *International Conference on Theory and Practice of Digital Libraries*, pages 197–201. Springer.

Judith Kessens and David A van Leeuwen. 2007. N-best: The northern-and southern-dutch benchmark evaluation of speech recognition technology. In *Eighth Annual Conference of the International Speech Communication Association*.

Nelleke Oostdijk. 2000. The spoken dutch corpus. overview and first evaluation. In *LREC*.

Roeland JF Ordelman and Franciska MG 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*. Stanford University Library.

Roeland JF Ordelman, Franciska MG de Jong, Adrianus J van Hessen, and GHW Hondorp. 2007. Twnc: a multifaceted dutch news corpus. *ELRA Newsletter*, 12(3-4).

Roeland Ordelman, Robin Aly, Maria Eskevich, Benoit Huet, and Gareth J.F. Jones. 2015a. Convenient discovery of archived video using audiovisual hyperlinking. In *Proceedings of the Third Edition Workshop on Speech, Language & Audio in Multimedia*, SLAM '15, pages 23–26, New York, NY, USA. ACM.

Roeland JF Ordelman, Maria Eskevich, Robin Aly, Benoit Huet, and Gareth Jones. 2015b. Defining and evaluating video hyperlinking for navigating multimedia archives. In *Proceedings of the 24th International Conference on World Wide Web*, pages 727–732. ACM.

Roeland Ordelman, Carlos Martínez Ortíz, Liliana Melgar Estrada, Marijn Koolen, Jaap Blom, Willem Melder, Jasmijn Van Gorp, Victor De Boer, Themistoklis Karavellas, Lora Aroyo, Thomas Poell, Norah Karrouche, Eva Baaren, Johannes Wassenaar, Oana Inel, and Julia Noordegraaf. To appear. Challenges in enabling mixed media scholarly research with multi-media data in a sustainable infrastructure. In *Digital Humanities 2018 (DH2018)*, Mexico City, Mexico.

Vijayaditya Peddinti, Yiming Wang, Daniel Povey, and Sanjeev Khudanpur. 2018. Low latency acoustic modeling using temporal convolution and lstms. *IEEE Signal Processing Letters*, 25(3):373–377.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. 2011. The kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, December. IEEE Catalog No.: CFP11SRW-USB.

Mickael Rouvier, Grégor Dupuy, Paul Gay, Elie Khoury, Teva Merlin, and Sylvain Meignier. 2013. An open-source state-of-the-art toolbox for broadcast news diarization. In *Interspeech*.

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.

Jeroen van den Hoven, Martijn Blaauw, Wolter Pieters, and Martijn Warnier. 2016. Privacy and information technology. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2016 edition.

H. Yang and C. Meinel. 2014. Content based lecture video retrieval using speech and video text information. *IEEE Transactions on Learning Technologies*, 7(2):142–154, April.

Towards TICCLAT, the next level in Text-Induced Corpus Correction

Martin Reynaert[31] **Maarten van Gompel**[2] **Ko van der Sloot**[2] **Antal van den Bosch**[12]
KNAW Meertens Institute, Amsterdam[1] CLS / Radboud University Nijmegen[2] TSHD / Tilburg University[3]
The Netherlands
martin.reynaert|antal.van.den.bosch@meertens.knaw.nl
m.vangompel|k.vandersloot@let.ru.nl

Abstract

We give an update of the state-of-affairs of the tools we have gradually been developing for the Dutch CLARIN infrastructure over the past 10 years. We first focus on our OCR post-correction system TICCL, next describe its wider environment, the corpus building work flow PICCL, and then sketch the various guises in which these are made available to the broad research community.

1 Introduction

Having worked in Dutch CLARIN projects for going on for a decade, we want to give a brief overview of what we have achieved. Our work centers around facilitating the building of text corpora, around improving the lexical quality of text corpora, around enhancing search and retrieval and around providing necessary infrastructure for researchers to actually achieve all these steps on their own. We first, in Section 2, introduce major extensions to our OCR post-processing system TICCL and next in Section 3 outline how it is embedded in a larger corpus building system called PICCL. Section 4 lists the various ways in which the systems are made available to the larger community. Section 5 provides a glimpse of where we go next, onwards with TICCLAT, the Text-Induced Corpus Correction and Lexical Assessment Tool.

2 Recent developments in Text-Induced Corpus Clean-up

We have steadily worked on improving our post-OCR corrrection system Text-Induced Corpus Clean-up or TICCL. We here provide a technical overview of major extensions that either help to optimize both the processing of large corpora and the end result to be obtained, i.e. a greater overall accuracy of the texts processed.

In our revised evaluation of TICCL's performance on about 10,000 books mainly from the late 18th century in Reynaert (2016) we stated we need to solve run-on and split words to enhance recall and to perform language recognition in order to avoid precision errors. We also stated we urgently need to find a way to boost TICCL's recall by fully exploiting its potential. We next list the steps taken to remedy the shortcomings identified.

2.1 Language recognition

We can now deploy FoLiA-langcat[1] to separately perform language recognition on each XML paragraph and label it accordingly. Downline TICCL tools being aware of the language labels, we prevent scores of false positives and boost the system's precision.

2.2 Focus list

In Reynaert (2010) we presented two approaches to post-OCR correction, i.e. the focus-word approach and the character confusion approach. Both obtain exactly the same result, but we argued the focus word approach was indicated for smaller corpora, while the character confusion approach seemed better suited

[1] One of the C++ FoLiA tools available from: https://github.com/LanguageMachines/foliautils. The TICCL tools are at https://github.com/LanguageMachines/ticcltools

to larger corpora. We have now achieved a major optimisation in the processing time required for both approaches. This is the result of providing them with a list of word forms only present in the corpus to be corrected in order for the system to focus solely on these, rather than evaluating all the word forms present in the corpus as well as in the lexicon and possible background corpus, as used to be the case.

2.3 Harnessing various OCR-versions of the same text

Due to the major digitisation programmes around the globe, these days it is quite likely one may find online various OCR versions for the same book, especially if it is an out-of-copyright work. These versions will have been obtained by means of different OCR engines, from different hard copy versions, etc. The idea being that these independent processes will have delivered varying results, the combination of the frequency lists obtained from these different versions should help to enforce the post-OCR correction process. The system now allows for easy incorporation of various OCR-versions into the process.

2.4 Bi/trigram correction

We are currently extending the C++ code base of TICCL with the capabilities necessary to properly process word bi- and trigrams. This is a major extension which we have first explored in a Perl prototype and tested on about 200 years of text from the Dutch Acts of Parliament[2]. We regard word bi- and trigrams as local evidence for the global decisions regarding word forms in a corpus TICCL makes. These decisions are based on the frequency list obtained from the corpus to be corrected, possibly complemented with a validated word form lexicon one may have for the language of the actual time period and/or a possible background corpus possibly comprising thousands of texts from the same era (Reynaert, 2014).

2.5 Ranking

We have added two major new ranking features to the ten features we described previously in Reynaert (2014), a combination of symbolic (i.e. character identity based) and corpus or lexicon derived statistical features. Pairs of variant and Correction Candidate (CC) are always ranked per feature, but it is the ensemble of features that determines their overall ranking in relation to the other pairs that share the same variant. The first new feature is a natural result of the bi/trigram correction extension. For every bi- or trigram that testifies to a particular variant and CC pair, a point is awarded. The pair that obtains the highest number of ngram points, is ranked best on this feature. The second new feature is based on the cosine distance between variant and CC, the CC with the closest distance naturally being ranked first. The drawback is that apart from the regular anagram hashing of the corpus, a separate e.g. word2vec[3] word vector space needs to be built. The feature so far is experimental only, its implementation being slow.

2.6 Chaining

The last step in the TICCL pipeline recently developed is that of chaining. Chaining is meant to greatly expand the reach of TICCL in terms of Levenshtein Distance (LD), even if the system as such is still run with an imposed limit of say LD = 2. Chaining consists of gathering the best-first ranked variants that in a very specific way, due to the LD limit imposed, are in fact interconnected. A criticism of limiting TICCL's reach to LD = 2, is that it is then unable to correct e.g. the allegedly likely OCR-error 'iii' for an 'm'. While this particular character confusion might very easily be built in to TICCL and exempted from the LD 2 limit imposed on all the other character confusions we search for, we currently explore another solution. As far as we can tell, the OCR process is even more likely to produce the character confusions 'in' or 'ni' for an 'm'. For the word 'Amsterdam', TICCL in correcting a large corpus of OCR-ed texts may well encounter 'Ainsterdam' or 'Amsterdani' (anagrams of one another) and correctly propose 'Amsterdam' as the best-first ranked correction candidate. LD = 2 being imposed, it cannot reach 'Amsterdam' as the CC for the OCR-errors 'Aiiisterdam' or 'Amsterdaiii', or indeed 'Ainsterdani', but it will propose 'Ainsterdam' or 'Amsterdani' instead. From 'Aiiisterdam' or 'Amsterdaiii' it may then reach the even further (in real Levenshtein distance terms) variants: 'Aiiisterdani' or 'Ainsterdaiii' (and all other variants within LD =2 present). In this way, CCs and their variants, serving as CCs for further variants, form a

[2] `http://www.statengeneraaldigitaal.nl`

[3] `https://code.google.com/archive/p/word2vec/`

'chain' and led to the idea of the new 'TICCL-chain' module. Between 'Amsterdam' and 'Aiiisterdaiii' the LD is already six characters, so this new module should greatly extend TICCL's reach and therefore its recall.

In the actual implementation, we start off with the output file of TICCL-rank, when it was run to produce only the best-first correction candidates for fully-automatic spelling or post-OCR correction. This file is sorted numerically descending on the CCs' frequency. In chaining, we start off by checking whether the variant(s) retrieved for the most frequent or topmost CC are in their turn reported to be CCs for other variants. If they are, each variant as CC is replaced by its own CC. This is continued down the list until new variants for the particular topmost CC run out. We then restart the process by taking the next to topmost CC, i.e. the second most frequent CC, and chain up all its variants. And so on down the list, until this is exhausted. Of course, at any rank, as there may be more than just one variant linked to any particular CC, care has to be taken to ensure all the possible 'chains' are collected. In this way, chaining may even correct for small chinks in the chain where variants have not in fact been properly ranked to the CC closest to the topmost CC. It must be clear that this process relies on high-precision variants having been reported for any CC. Otherwise totally unrelated word strings may very quickly be linked up and chained into constituting variant-CC pairs, while ostensibly they are not. As stated before, this process, when successful, greatly enhances TICCL's recall beyond LD = 2. The above should also help to explain why TICCL works better with larger corpora. If one tries to correct only a single text with perhaps only three or four occurrences of a particular word most of which have been badly recognised by the OCR, it is unlikely all the small intermediate steps towards the topmost CC are present. In that case, the chances of being able to reconstitute the complete chain are small.

3 TICCL in its broader environment, PICCL

In Reynaert et al. (2015) we have previewed the corpus building work flow 'Philosophical Integrator of Computational and Corpus Libraries'. PICCL[4] has now grown into a more mature production system. It is available to anyone with an academic logon as a single sign-on system offering the user her own private work space[5]. The user may now freely upload her own texts in a wide range of formats and have them converted to PICCL's pivot format FoLiA XML (van Gompel et al., 2017). Alternatively she may have them digitized by means of Optical Character Recognition through Tesseract[6] in case these are just images of printed pages. The electronic text may next at will be corrected for OCR-errors, normalized or modernized into today's standard language by TICCL. The text can then be tokenized by Ucto[7] or, if Dutch, further linguistically enriched by Frog[8].

Now being implemented in the Nextflow[9] work flow environment, PICCL has become fully distributable, ready for any parallelization framework and capable of seamlessly running on a grid platform, or in the cloud, or indeed on your own servers. On our development server, OCR engine Tesseract may well simultaneously OCR 60 pages of a book, depending on availability of processors, cores or threads.

The whole work flow as defined in Nextflow is then wrapped into CLAM[10] and thereby turned into a web application and RESTful web service[11] which allows the user to (de)select the tools required for her job and even to import say a particular book straight through the API of the Dutch National Library or any other repository world-wide that offers this facility.

4 TICCL's availability in the CLARIN Infrastructure

TICCL is now an integral part of the production system PICCL and part of the CLARIN research infrastructure. It is available as a web application or service running on dedicated hardware at CLARIN

[4] https://github.com/LanguageMachines/PICCL
[5] portal.clarin.ivdnt.org/piccl To be available for production work from June, 1st 2018
[6] https://github.com/tesseract-ocr/tesseract
[7] https://languagemachines.github.io/ucto/
[8] https://languagemachines.github.io/frog/
[9] https://www.nextflow.io/
[10] http://proycon.github.io/clam/
[11] https://webservices-lst.science.ru.nl

Center INT for less data-intensive purposes, e.g. processing a single book. Furthermore, the system is freely available as part of the LaMachine distribution[12], a meta-distribution system for software designed to allow any research team to set up its own local installation in a hassle-free manner.

5 Onwards to project TICCLAT

CLARIAH and the Netherlands eScience Center have in the meantime awarded us a new project, TICCLAT. In this TICCL is to be yet further expanded and will become the 'Text-Induced Corpus Correction and Lexical Assessment Tool'. With the help of an eScience engineer, we are to extend TICCL's data structures in such a way that it becomes accumulative and retains what it has learned correcting one corpus, for further use in the correction of other corpora. TICCLAT will be trained again and again on the ever noisier diachronic Dutch corpora contained in the Nederlab corpus (Brugman et al., 2016), approximately 20 billion word tokens, which we have been collecting. As soon as TICCL as we have described in the previous sections is fully operational, it will serve as the baseline for these further enhancements. In parallel with current TICCL development, we are collaborating with the KB, the Dutch National Library, on building more extensive evaluation sets than we currently have for Dutch. We choose to defer new TICCL evaluations until these become available.

6 Conclusion

We have described the latest additions to TICCL, our post-OCR correction system, and situated it in its wider environment, the corpus building tool PICCL. We have further discussed their wide availability options within the CLARIN community and all too briefly given a glimpse of TICCL's future.

Acknowledgements

The authors acknowledge having been funded by CLARIN-NL, CLARIAH, KNAW and especially NWO in numerous projects over the past decade.

References

[Brugman et al.2016] Hennie Brugman, Martin Reynaert, Nicoline van der Sijs, René van Stipriaan, Erik Tjong Kim Sang, and Aantal van den Bosch. 2016. Nederlab: Towards a Single Portal and Research Environment for Diachronic Dutch Text Corpora. In Nicoletta Calzolari et al., editor, ***Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC-2016)***, pages 1277–1281, Portorož, Slovenia. ELRA.

[Reynaert et al.2015] Martin Reynaert, Maarten van Gompel, Ko van der Sloot, and Antal van den Bosch. 2015. PICCL: Philosophical Integrator of Computational and Corpus Libraries. In ***Proceedings of CLARIN Annual Conference 2015 – Book of Abstracts***, pages 75–79, Wrocław, Poland. CLARIN ERIC.

[Reynaert2010] Martin Reynaert. 2010. Character confusion versus focus word-based correction of spelling and OCR variants in corpora. ***International Journal on Document Analysis and Recognition***, pages 1–15. 10.1007/s10032-010-0133-5.

[Reynaert2014] Martin Reynaert. 2014. Synergy of Nederlab and @PhilosTEI: diachronic and multilingual Text-Induced Corpus Clean-up. In Nicoletta Calzolari et al., editor, ***Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)***, pages 1224–1230, Reykjavik, Iceland. European Language Resources Association (ELRA). ACL Anthology Identifier: L14-1624.

[Reynaert2016] Martin Reynaert. 2016. OCR post-correction evaluation of Early Dutch Books Online – revisited. In Nicoletta Calzolari et al., editor, ***Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC-2016)***, pages 967–974, Portorož, Slovenia. ELRA.

[van Gompel et al.2017] Maarten van Gompel, Ko van der Sloot, Martin Reynaert, and Antal van Den Bosch. 2017. FoLiA in practice: The infrastructure of a linguistic annotation format. In J. Odijk and A. van Hessen, editors, ***CLARIN-NL in the Low Countries***, chapter 6, pages 71–81. Ubiquity.

[12] `https://proycon.github.io/LaMachine/`

SenSALDO: a Swedish Sentiment Lexicon for the SWE-CLARIN Toolbox

Jacobo Rouces Lars Borin
Nina Tahmasebi Stian Rødven Eide
Språkbanken
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 describe the creation of SenSALDO, a comprehensive sentiment lexicon for Swedish, which is now 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 (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: (1) creation of a gold standard; (2) implementation and evaluation of different automatic methods for creating the sentiment lexicon (see section 2; and (3) manual curation of the results of the best performing method. 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 $[-1, 1]$.

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). 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 usage. Any number of secondary descriptor relations provide additional semantic properties of the entry that are not conveyed by the primary descriptor, like an inversion or negation or some important semantic argument in an hypothetical definition. The primary descriptor structure forms a tree and the secondary descriptors define a directed acyclic graph.[1]

2 Methods

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

[1] For a detailed description of the organization of SALDO and a discussion of the underlying theoretical and methodological principles, see Borin et al. (2013).

word embeddings as features for different classifiers.

We model the sentiment associated to a word sense using a real 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., 2018a).

We experiment with different approaches, which we describe below, extending the methods of Rosell and Kann (2010) and Nusko et al. (2016) and also trying a corpus-oriented approach similar to the one described by Hamilton et al. (2016). For all methods, we produce continuous scores and discrete labels (positive, neutral, negative). What is relevant about the 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.

2.1 Inheritance over Graph

Our first method is a modified and extended version of the tree traversal method presented by Nusko et al. (2016), where the sentiment of a word sense is inherited from the primary descriptor (which defines a tree structure). We extended it in a way that 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.

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. The other third is used for testing. This is the only learning needed by this algorithm.

2.2 Random Paths over Graphs

For our second experimental setup, we develop an adaptation of the method by Rosell and Kann (2010), which uses random paths over a graph of synonyms built using the Synlex/People's Dictionary of Synonyms (Kann and Rosell, 2005). Synlex uses Swedish-English lemma pairs concatenated with their inverse relation to generate pairs of candidate synonym pairs. The pairs were filtered by first asking users to grade pairs and then averaging the grades, resulting in 16006 words with 18920 weighted pairs. Our modification consists of adapting Synlex to use SALDO word senses instead of Swedish sense-ambiguous lemmas (the adaptation is performed 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 close to synonymy.

The discrete labels are obtained using the same thresholding method as in the inheritance-based method.

2.3 Classification over word2vec

As opposed to the previous methods, which are purely lexicon-driven, the third approach is partly corpus-based. We use 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. As source for the vector representations we have used the *Swedish Culturomics Gigaword Corpus* (Eide et al., 2016).[2] The corpus size is 1 billion words, and the vector space dimensionality is 512. We train a logistic regression classifier (word2vec-logit) and a support vector classifier with a radial basis function

[2]The corpus can be downloaded from Språkbanken under a CC-BY license: `https://spraakbanken.gu.se/eng/resource/gigaword`

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 (pos,neu,neg) classes. For each fold, the SVM/RBF meta-parameters (C, γ) were estimated using 5-fold cross-validation over the training set. Although not equivalent, the linear nature of the logit classifier makes it comparable to the method of Rothe et al. (2016).

The classifiers' final output are discrete labels (positive, neutral, negative), 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).

2.4 Results

For training and testing the different methods, we use the direct annotation gold standard developed by Rouces et al. (2018a), which contains of 1997 entries from SALDO entries labeled as negative (-1), neutral (0), or positive ($+1$). For reasons of space, we do not present the detailed results her, but refer the reader to Rouces et al. (2018b).

The method word2vec-svc-rbf performed consistently better than the rest, and therefore we use it for the input to the manual curation step. Table 1 shows some examples of sentiment scores obtained using this method.

word sense	gloss	value	label	word sense	gloss	value	label
ond..4	'bad'	-0.9959	neg	förhållande..1	'relationship'	-0.0345	neu
farlig..1	'dangerous'	-0.9919	neg	radio..1	'radio'	-0.0264	neu
kriminalitet..1	'criminality'	-0.9838	neg	sälja..1	'sell'	-0.0223	neu
skrämma..1	'frighten'	-0.9797	neg	surdeg..1	'sourdough'	0.0426	neu
problem..1	'problem'	-0.9716	neg	god..2	'tasty'	0.9675	pos
angrepp..1	'attack (n)'	-0.9594	neg	riktig..2	'genuine'	0.9716	pos
risk..1	'risk (n)'	-0.9473	neg	hjälpa..1	'help (v)'	0.9797	pos

Table 1: Examples of sentiment values and labels obtained with the word2vec-svc-rbf method

3 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 assigment 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 by corpus frequency in the Gigaword Corpus (described in section 2.3 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: 2640 neutral, 1584 positive, and 3394 negative items. A detailed analysis of the differences is still pending.

4 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.[3]

[3]The first version of this resource – SenSALDO v. 0.1 – is freely available for downloading under a CC-BY license from Språkbanken: `https://spraakbanken.gu.se/eng/resource/sensaldo`

Merely providing the downloadable lexicon is generally not sufficient for the user community targeted by CLARIN. For this reason, 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 corpora, which can cover cases that the lexicon-based approach cannot cover either due to limited coverage or non-compositional expressions. We have also included sentiment annotation based on SenSALDO in Språkbanken's online annotation tool *Sparv*[4] and the new document-oriented infrastructure component *Strix*, with the aim to provide document filtering based on sentiment.[5]

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, Markus Forsberg, and Lennart Lönngren. 2013. SALDO: A touch of yin to WordNet's yang. *Language Resources and Evaluation*, 47(4):1191–1211.

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.

William L Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. 2016. Inducing domain-specific sentiment lexicons from unlabeled corpora. *arXiv preprint arXiv:1606.02820*.

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.

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, Nina Tahmasebi, Lars Borin, and Stian Rødven Eide. 2018a. Generating a gold standard for a Swedish sentiment lexicon. In *LREC 2018*, Miyazaki. ELRA.

Jacobo Rouces, Nina Tahmasebi, Lars Borin, and Stian Rødven Eide. 2018b. SenSALDO: Creating a sentiment lexicon for Swedish. In *LREC 2018*, Miyazaki. ELRA.

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.

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.

[4] `https://spraakbanken.gu.se/sparv`
[5] `https://spraakbanken.gu.se/strix/`

Error Coding of Second-Language Learner Texts Based on Mostly Automatic Alignment of Parallel Corpora

Dan Rosén
Språkbanken, Department of Swedish
University of Gothenburg, Sweden
dan.rosen@svenska.gu.se

Mats Wirén
Department of Linguistics
Stockholm University, Sweden
mats.wiren@ling.su.se

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

Abstract

Error coding of second-language learner text, that is, detecting, correcting and annotating errors, is a cumbersome task which in turn requires interpretation of the text to decide what the errors are. This paper describes a system with which the annotator corrects the learner text by editing it prior to the actual error annotation. During the editing, the system automatically generates a parallel corpus of the learner and corrected texts. Based on this, the work of the annotator consists of three independent tasks that are otherwise often conflated: correcting the learner text, repairing inconsistent alignments, and performing the actual error annotation.

1 Introduction

Error coding is a highly useful way of increasing the value of second-language learner data, but it is also ”beset with difficulties” (Granger, 2008, page 266). Like all manual annotation, it is very time-consuming, but the problem is exacerbated by the fact that the text must be interpreted to decide what the correct forms (the target hypotheses) are. Our approach is based on regarding the learner source text and the corrected text as a parallel corpus. The annotator makes the corrections of the learner text using a text editor upon which the words of the two resulting texts are automatically aligned. The annotator may re-align inconsistent links, and can then perform the actual error annotation.

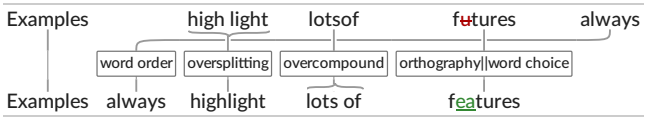
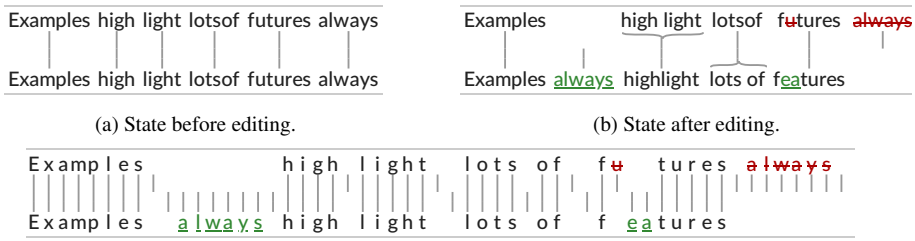


Figure 1: Example learner sentence (above), corrected target hypotheses (below) and error annotation.

Figure 1 shows a hypothetical learner sentence, its correction and the automatic alignments generated by the system. The misspelled or wrongly chosen word *futures* is aligned with the corrected word *features*. Movements are kept track of by alignments of the words involved, for example, *always* at the end of the learner sentence with *always* at the second position in the corrected sentence. Compound oversplitting, as in the linking of *high light* and *highlight*, is represented by a many-to-one alignment. Conversely, overcompounding, as in *lotsof* and *lots of*, is represented by a one-to-many alignment.

2 Related work

Error coding in second-language learner texts is typically made with a purpose-built editing tool, using a hierarchical set of error categories (Granger, 2008), for example, the XML editor Oxygen in the ASK



(a) State before editing.

(b) State after editing.

(c) Character based diff: this is lifted to the token level in Figure 2b. (not shown to the user)

Figure 2: Before and after editing. The automatic aligner gets most words correct and the user will later manually connect the unlinked words.

project (Tenfjord et al., 2006), the TEITOK editor in the COPLE2 project (Mendes et al., 2016), and QAWI in the QALB project (Obeid et al., 2013). In many systems, however, the conceptually separate stages of error coding — error detection, correction and annotation (Ellis, 1994; Granger, 2008, page 266) — are conflated such that correction is carried out as a subtask of error annotation, for example, Tenfjord et al. (2006) and Mendes et al. (2016). In contrast, two projects that factor out correction as an independent step are MERLIN (Boyd et al., 2014) and CzeSL (Hana et al., 2012). The latter of these, with its parallel corpus editor *feat*, is the one most closely related to our system. The main difference is that our editor makes word alignments automatically. On the other hand, *feat* includes two tiers of target hypotheses, roughly corresponding to correction of spelling and grammatical errors, respectively, whereas our system currently only has one level of correction.

In translation and related fields, several tools for manual or semi-automatic word alignment of parallel corpora have been developed, for example, *Interactive Linker* (Merkel et al., 2003), *Yawat* (Germann, 2008), 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 been used for alignment of parallel corpora, but not for alignment of learner corpora as far as we are aware of.

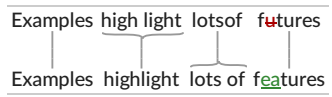
3 Correction and alignment

3.1 Basic workings

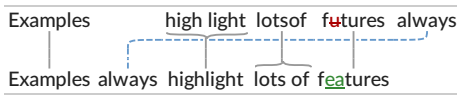
The system interface includes two main panes: one editable text pane displaying the text being corrected, and a graphic rendering of the links between the source text and the corrected text. An optional third pane displays the static learner text. Initially the corrected text is identical to the source text, and each word is aligned to its copy. In the example "*Examples high light lotsof futures always*", the graph initially looks as in Figure 2a. To correct a text, the annotator uses the editable text pane, which works like a standard display editor. Editing operations can be made in any order, and upon each change, the system immediately updates the alignments between the two texts. Changing the text to "*Examples always highlight lots of features*" results in the alignment in Figure 2b.

Our system builds a corrected version aligned with the learner text using an underlying JSON representation of the tokens, the groups of links between these, and any labels attached to the groups. How is this calculated? We start with a standard diff edit script on the *character level*. Internally, the representation in Figure 2c is generated, which is calculated using Myers' diff algorithm (Myers, 1986) provided by the diff-match-patch[1] 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 a link to their corresponding tokens. For example,

[1] `https://github.com/google/diff-match-patch`

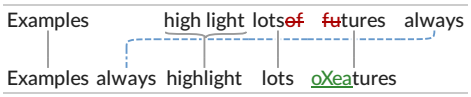


(a) Diff without *always*.



(b) Diff with *always* correctly inserted from Figure 3a.

(c) Editing from Figure 3b between *always* and *highlight*.



(d) Editing from Figure 3b between *of* and *features*.

Figure 3: Aligning and editing in the presence of a manually aligned word (here *always*).

since there is a link between the *h* in *high* and *highlight*, these two words are aligned. Furthermore, since there is a link between the *l* in *light* to this target word, all these three words should also be aligned. 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.

3.2 Manual alignments: word order changes and inconsistent links

In Figure 2b, 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 3a), and then *insert* the manually aligned words again in their correct position (Figure 3b). Here the correct position is where they were removed from the respective texts. The editor indicates that this link is manually aligned by colouring it blue and (for the purposes of this article) making it dotted. These links interact differently with automatically aligned (grey) links. How this works is explained in the next section.

3.3 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 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 3b and replace it with *alwayXighligt*, the state becomes as shown in Figure 3c. However, if we cross the boundary of *of features* in the same starting state of Figure 3b to *oXeatures*, we get the state of Figure 3d. Here the edit is not contagious: the automatic alignment decided not to make a big component and instead chose to split to align the words independently. In case the manual aligner has absorbed too much material, our editor provides a way of removing the manual alignment tag. The editor will then fall back to the automatic aligner for those tokens.

4 Error annotation

Error categories can be seen as relations between words in the learner text and corrected text, and are therefore associated with the alignments, as in Figure 1. Our error taxonomy is inspired by that of ASK (Tenfjord et al., 2006), with two hierarchical levels (main categories and subcategories).[2] Error annotation is carried out by selecting one or several words in the learner and corrected texts, or by selecting the corresponding link(s). A pop-up menu is displayed and the annotator selects one or more error categories, whereupon the system attaches these categories to the corresponding links as shown in Figure 1. For meaningful error annotation to take place, some correction must have been made, but the order between the tasks is otherwise arbitrary.

[2]The names of the categories differ from the example labels used in Figure 1.

5 Concluding remarks

CLARIN currently lacks an established infrastructure for research in second-language learning. The work reported here is part of the SweLL project aiming at filling this gap for Swedish, but our intent is to make the methods as generally applicable as possible. To this end, the system described here is free software under the MIT license.[3] Also, the error taxonomy is customisable in the system. We expect to soon be able to provide a more complete description of the system and the rationale for our methodology, including the error taxonomy, practical experiences with annotators, use of the system for anonymisation, and functionalities for management of annotators.

Acknowledgements

This work has been supported by Riksbankens Jubileumsfond through the SweLL project (grant IN16-0464:1, `https://spraakbanken.gu.se/eng/swell_infra`). Some of the basic ideas were first tested in a pilot system by Hultin (2017), supervised by Robert Östling and Mats Wirén. We are grateful for valuable comments and feedback on the system from Markus Forsberg, Lars Borin and our co-workers in the SweLL project.

References

[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).

[Ellis1994] Rod Ellis. 1994. *The Study of Second Language Acquisition*. Oxford University Press, Oxford.

[Germann2008] Ulrich Germann. 2008. Yawat: Yet another word alignment tool. In *Proc. ACL: HLT Demo Session*, pages 20–23. Association for Computational Linguistics.

[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.

[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, Budapest, Hungary.

[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.

[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.

[3] `https://github.com/spraakbanken/swell-editor`.

Using Apache Spark on Hadoop Clusters as Backend for WebLicht Processing Pipelines

Soheila Sahami
Natural Language Processing Group
University of Leipzig, Germany
sahami@informatik.uni-leipzig.de

Thomas Eckart
Natural Language Processing Group
University of Leipzig, Germany
teckart@informatik.uni-leipzig.de

Gerhard Heyer
Natural Language Processing Group
University of Leipzig, 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 SOA-based platforms - like the CLARIN-D 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. Those systems and their corresponding software architecture already support many of the features relevant for processing big data for large user groups. This submission gives an example of a specific 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 can be 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*[1] 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 already mentioned benefits of a service-oriented architecture.

[1] https://cordis.europa.eu/project/rcn/103684_en.html

- Efficient use of parallelization, including the parallel processing of large document collections and the support of large user groups.
- 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[2] framework on Hadoop clusters as a 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. Related issues - like potential consequences for the design of user interfaces and workflow structuring - will be discussed afterwards.

2 Technical Approach

Distributed data processing systems to improve performance, response times, and modularity are research topics in the computer sciences for many decades (Enslow, 1978). With an increasing availability of text material, NLP researchers and developers are encouraged to analyze massive amounts of data, and to benefit from their huge quantities and a potential higher quality and significance of gained results. However, required hardware resources and runtimes to process "big data" input material are still major challenges. Apache Hadoop, an open-source software for reliable, scalable and distributed computing, provides a framework for the distributed processing of large data sets and is widely used in many scientific fields, including processing of natural language. It also provides with the Hadoop Distributed File System (HDFS) a distributed storage solution for processing large data sets with eminently fault-tolerant and high-throughput access to application data (Apache Hadoop, 2018). Apache Spark is a cluster computing platform which can be used as execution engine to process huge data sets on Hadoop-based clusters. Apache Spark uses a multi-threaded model where splitting tasks on several executors improves processing times and fault tolerance. In the MapReduce approach, input data, intermediate results, and outputs are stored on disk which requires additional I/O operations for each processing unit. In contrast, In-Memory Databases (IMDB) are designed to run completely in RAM. IMDB is one of the salient features of Apache Spark which has the potential to reduce processing times significantly (Karau et al., 2015).

The service-oriented WebLicht architecture allows all kind of platforms as processing backends for its endpoints by hiding implementation details behind public and standardized interfaces. As a consequence, the utilization of a Hadoop/Spark-based system fits into this architecture and can provide enhanced processing capabilities to the infrastructure.

3 Implementation and Results

In the context of a first implementation, a variety of typical NLP tools - including sentence segmentation, pattern-based text cleaning, tokenizing, and language identification - were implemented[3]. During the execution, input text files are loaded in Spark-specific data representations called "Resilient Distributed Datasets" (RDD). Those RDDs are distributed over the allocated cluster hardware to be processed by several executors and cores in parallel. For every job, the hardware configuration 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

[2] `https://spark.apache.org/`

[3] `http://hdl.handle.net/11022/0000-0007-CA50-B`

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 TCF format (TCF, 2018) (i.e. tokenization and sentence segmentation) converters between TCF and the RDDs were written. As result, the endpoint is structured as presented in Figure 1.

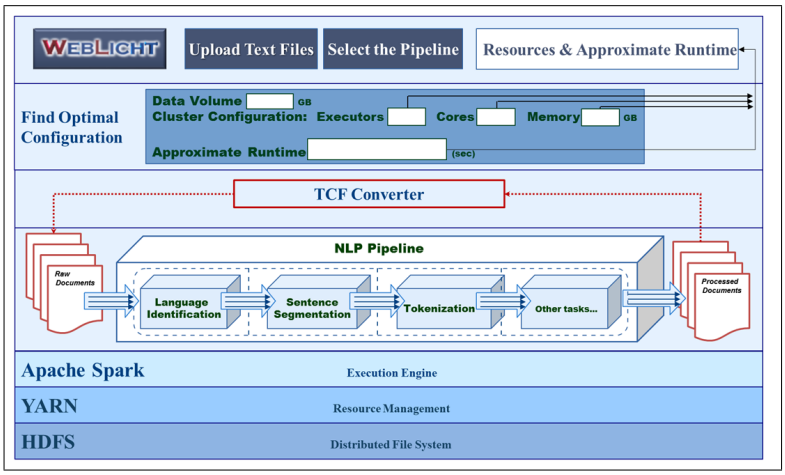


Figure 1: Service-oriented Architecture with a Spark-based Backend.

For evaluation of the established solution, benchmarks were executed to show the impact of parallelization on every task. Table 1 illustrates the hardware and configuration of the cluster (Lars-Peter Meyer, 2018) which was used in the context of the Dresden-Leipzig Big Data center of competence *ScaDS*[4].

Number of nodes	CPUs	Hard drives	RAM	Network
90	6 cores per node	>2 PB in total	128 GB per node	10 GB/s Ethernet

Table 1: Cluster Characteristics

The following diagrams 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.

Resulting execution times for a specific configuration (i.e. number of used workers, 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 based on their general characteristics (i.e. size of input data and used tool configuration). This estimate can be provided to the users, which may help to increase their willingness to wait for results. It also is valuable information to find an optimal balance between number of parallel user tasks, available hardware configuration, and waiting times that are still acceptable for the users.

4 Potential Consequences and Further Work

The presented approach can be integrated in the current WebLicht architecture and helps solving - or at least mitigating - the aforementioned problems. However, for a systematic support of big data processing in the context of WebLicht pipelines, changes in the default workflows and

[4]Rahm and Nagel (2014)

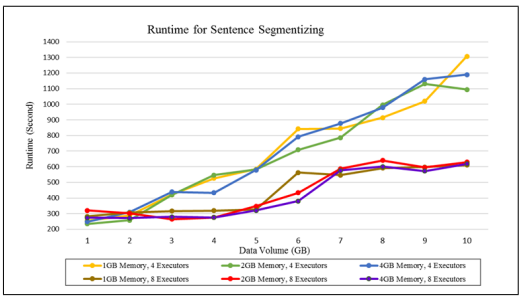


Figure 2: Segmenting 1 to 10 GB Text Data using 4 or 8 Executors and 1, 2 or 4 GB RAM.

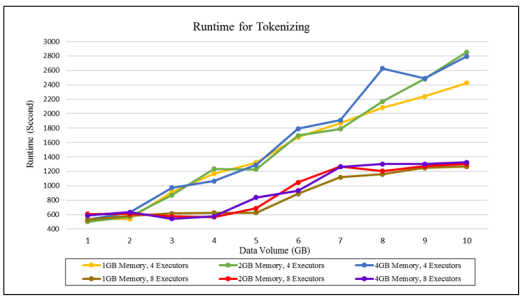


Figure 3: Tokenizing 1 to 10 GB Text Data using 4 or 8 Executors and 1, 2 or 4 GB RAM.

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 shift from the current *push* communication behavior to *pull* communication patterns.

The latter is especially important as synchronous communication is hardly feasible for the handling of large data resources in a SOA. An alternative might be a stronger focus on data storage platforms that support workspaces for individual users like B2DROP. User information about status and outcome of scheduled processing jobs can be transfered via Email or job-specific status pages. Those status reports should be seen as an import means to inform end 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.

As a next step, it is planed to extend the amount of supported annotation tools and increase the number of potential hardware backends. Attempts to port the current pipeline to a high-performance computing (HPC) cluster are currently carried out and might lead to contact points with other established research infrastructures in this field[5].

References

[Apache Hadoop2018] Apache Hadoop. 2018. Apache Hadoop Documentation. Online. Date Accessed: 11 Apr 2018. URL http://hadoop.apache.org/.

[Enslow1978] Philip H. Enslow. 1978. What is a "distributed" data processing system? *Computer*, 11(1):13–21.

[Gate Cloud2018] Gate Cloud. 2018. GATE Cloud: Text Analytics in the Cloud. Online. Date Accessed: 11 Apr 2018. URL https://cloud.gate.ac.uk/.

[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.

[Karau et al.2015] Holden Karau, Andy Konwinski, Patrick Wendell, and Matei Zaharia. 2015. *Learning spark: lightning-fast big data analysis.* O'Reilly Media, Inc.

[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-nothing-architecture-in-leipzig.

[PRACE2018] PRACE. 2018. PRACE Research Infrastructure. Online. Date Accessed: 27 Apr 2018. URL http://www.prace-ri-eu.

[5]Like the project "Partnership for Advanced Computing in Europe" (PRACE, 2018).

[Rahm and Nagel2014] Erhard Rahm and Wolfgang E. Nagel. 2014. ScaDS Dresden/Leipzig: Ein service-orientiertes Kompetenzzentrum für Big Data. In E. Plödereder, L. Grunske, E. Schneider, and D. Ull, editors, *Informatik 2014*, pages 717–717, Bonn. Gesellschaft für Informatik e.V.

[TCF2018] 2018. The TCF Format. Online. Data Accessed: 27 Apr 2018. URL https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/The_TCF_Format.

UWebASR – Web-based ASR engine for Czech and Slovak

Jan Švec
Department of Cybernetics
University of West Bohemia
Plzeň, Czech Republic
honzas@kky.zcu.cz

Martin Bulín
Department of Cybernetics
University of West Bohemia
Plzeň, Czech Republic
bulinm@kky.zcu.cz

Aleš Pražák
Department of Cybernetics
University of West Bohemia
Plzeň, Czech Republic
aprazak@kky.zcu.cz

Pavel Ircing
Department of Cybernetics
University of West Bohemia
Plzeň, Czech Republic
ircing@kky.zcu.cz

Abstract

The paper introduces a beta-version of a user-friendly Web-based ASR engine for Czech and Slovak that enables users without a background in speech technology to have their audio recordings automatically transcribed. The transcripts are stored in a structured XML format that allows efficient manual post-processing.

1 Introduction

The recent deployment of the automatic speech recognition (ASR) technology in the applications for both the personal computers and the mobile devices increased the awareness of the people outside the ASR community about the possibility to have the spoken content transcribed automatically. As a results, many researchers including (but not limited to) the scholars working in the field of digital humanities started to explore the potential of the ASR technology for their own research agenda. However, the development of the ASR system is still a very complex task even for people who are otherwise experienced in machine learning in general, not to mention people with only the humanities background.

The main factor the sparked the work on the cloud-based speech recognition engine presented in this paper was a growing number of requests from researchers that collect and curate the speech recordings in various archives (TV, radio and – most frequently – oral history collections). The reason for their interest in ASR processing stems from the fact that they have quickly found out that without the textual representation of the speech content, it is very difficult to access the relevant passages of the archive. Consequently, they usually resorted to manual transcription and/or metadata annotation of the individual recording which is both time-intensive and costly. Unfortunately, especially the organizations working with limited budget and dealing with relatively smaller languages such as Czech and Slovak could hardly find a cheap and easy-to-use of-the-shelf solution for automatic transcription of spoken content. Recently, they could use the Google Speech API service (`https://cloud.google.com/speech/`) but it is neither free nor easy-to-use for a person without a solid IT background. The usage of the open source KALDI toolkit (Povey et al., 2011) is completely out of the question for anybody who has not done at least some research work in the ASR area previously.

We are therefore introducing a first version of a user-friendly Web-based ASR engine for Czech and Slovak that will be free to use for research purposes and does not require any background knowledge about the inner workings of the ASR engine or the API usage. It was inspired by the webASR service (`www.webasr.org` – see (Hain et al., 2016) for details) that provides (among other things) the automatic transcription functionality for spoken data in English.

2 Technical Description of the System

2.1 Underlying Machine Learning Models

To get the best results from general-purpose ASR system (i.e. not domain-oriented), the key components - an acoustic and a language model - should be trained from a large corpus of varied data. Our acoustic training data consists of 1 990 hours from 15 000 different speakers - both clear read speech and real speech data with different levels of noise are included. For acoustic modeling, we use common three-state Hidden Markov Models (HMMs) with output probabilities modeled by a Deep Neural Network (DNN) (Veselý et al., 2013). Language model training corpus contains texts from newspapers (480 million tokens), web news (535 million tokens), subtitles (225 million tokens) and transcriptions of some TV programs (210 million tokens). Resulted trigram language model incorporates over 1.2 million words with 1.4 million different baseforms (phonetic transcriptions).

2.2 Back End

The back-end uses the *SpeechCloud* platform developed at the Department of Cybernetics of University of West Bohemia. The platform provides a remote real-time interface for speech technologies, including speech recognition and speech synthesis. Currently, it uses the real-time optimized speech engines supplied by the university spin-off company SpeechTech s.r.o. The speech recognition engine allows to recognize the speech with a recognition vocabulary of the size exceeding 1 million words in a real-time. In addition, it outputs the word confidence scores and lattices of multiple hypotheses. One of the unique features of the recognition engine is the ability to dynamically modify the recognition vocabulary by adding new words during the recognition. The platform also allows the use of plug-ins written in Python language to post-process the recognition results depending on the required application (for example to detect named entities or perform statistical-based classification of utterances).

Due to the real-time nature of the SpeechCloud, the allocation of recognition engines is based on sessions, not on requests, i.e. the client first needs to connect to the platform to create a new session with allocated recognition engine. The audio data could be sent only during the session. When the client disconnects, the session is destroyed and the allocated recognition engine could be used by another session (and client). Due to the limited resources of the computational platform, the number of simultaneously running engines is also limited.

In the background, the SpeechCloud client opens two connections to the SpeechCloud platform - the first connection uses JSON messages carried over WebSocket protocol and it is used for interaction with the speech engines (start/stop the recognition, send the recognition result, synthesize text), the second connection transfers the audio data over standard SIP/RTP (Session Initiation Protocol/Real-time Transport Protocol).

The deployment of UWebASR required modification of the SpeechCloud platform to allow the processing of uploaded speech files in various audio formats (like WAV, MP3 etc.). The current implementation uses the FFmpeg transcoder to extract the raw PCM data from the streamed audio file on-the-fly. The PCM data are then fed into the recognition engine in the same way like the data received over SIP/RTP. This allows to display the state of the recognition process in the real-time including the partial hypothesis and confidence scores.

2.3 Graphical User Interface

Services of the powerful ASR engine are available through a web-based graphical interface placed at `https://uwebasr.zcu.cz` (the secure mode is required). The purpose is to make it both user-friendly and able to utilize the wide scale of back-end features. The web page interactivity is enabled by Javascript exploiting event listeners aptly provided by the SpeechCloud platform and the graphics is designed in HTML/CSS. Out of the planned-to-offer services, only Recognition from file is enabled at the moment.

Once the page is loaded and the mic is enabled, a new session is created by connecting to the platform with the default (Czech) vocabulary selected. A successful connection is indicated by a green dot in the

upper right corner (a red dot means not connected). Each time the language is changed, the active session is destroyed and a new one is created. The service life of a single session is limited to 10 minutes.

Besides the initial (*waiting-for-file*) state, there are three more states the GUI can take during the overall procedure of Recognition from file.

1. *Processing a file*. The processed audio file is cut into little chunks and sent to the engine sequentially, overcoming the limitations in the WebSocket frame size and so allowing to process even files of the order of 100 MB. The progress of sending the file to the engine as well as the signal energy is graphically illustrated (see Fig. 1). Utilizing the unique back-end ability to dynamically modify the recognition result on-the-fly, a partial result is shown and updated each time a new prediction arrives from the engine. Moreover, a confidence score of each single word is indicated by color.

2. *File processed - got a recognition result*. When the final (not partial) recognition result comes from the engine, a file with transcription (in the XML-based Transcriber format (Barras et al., 2001)) is generated and offered for download.

3. *File processing error*. An error occurs when *1)* the uploaded file does not include any audio track; *2)* the connection to the engine is lost.

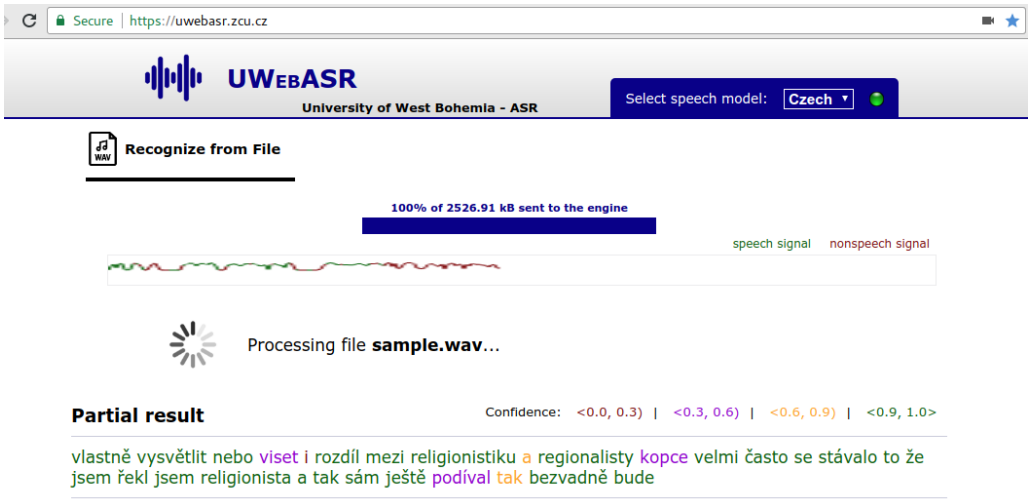


Figure 1: Screenshot of the web-based GUI - processing a file.

More experienced users appreciate a log console that can be shown/hidden by a button in the lower right corner. Besides other notifications it shows the URI of the current session containing raw engine messages, useful for debugging. The layout of the GUI is customized for adding new services (TTS, Real-time ASR) as well as for switching among unlimited number of language models in the future.

3 Connection to CLARIN

The UWebASR engine is currently being thoroughly tested at the University of West Bohemia as well as in our partner Czech labs connected via LINDAT/CLARIN. Unfortunately, the evaluation of the system performance (in terms of Word-Error-Rate - WER) and the actual user experience is still underway and thus we will be able to report it only later. The plan is to offer the tool to the general public through the LINDAT/CLARIN website by the end of 2018.

4 Conclusion and Future Work

The first version of the a user-friendly Web-based ASR engine for Czech was successfully implemented and it is currently being tested. Recently, we have also added the ASR engine for Slovak to the web

interface described in this paper. The next step will be the thorough evaluation of the system, both from the recognition accuracy and the user experience point of view. Implementation of the REST API – as required by the LINDAT/CLARIN coordinator – is also planned in the in the near future. Should the user testing reveal the need for any modifications of the interface, those will of course be done as well.

References

Claude Barras, Edouard Geoffrois, Zhibiao Wu, and Mark Liberman. 2001. 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):5–22.

Thomas Hain, Jeremy Christian, Oscar Saz, Salil Deena, Madina Hasan, Raymond W. M. Ng, Rosanna Milner, Mortaza Doulaty, and Yulan Liu. 2016. webASR 2 - improved cloud based speech technology. In *Proceedings of Interspeech 2016*, pages 1613–1617.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukáš Burget, Ondřej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlíček, Yanmin Qian, Petr Schwarz, Jan Silovský, Georg Stemmer, and Karel Veselý. 2011. The Kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, December. IEEE Catalog No.: CFP11SRW-USB.

Karel Veselý, Arnab Ghoshal, Lukáš Burget, and Daniel Povey. 2013. Sequence-discriminative training of deep neural networks. In *Proceedings of Interspeech 2013*, pages 2345–2349.

Pictograph Translation Technologies for People with Limited Literacy

Vincent Vandeghinste
Dutch Language Institute
Leiden, Netherlands
vincent.vandeghinste@ivdnt.org

Leen Sevens
Centre for Computational Linguistics, KU Leuven
Leuven, Belgium
leen@ccl.kuleuven.be

Ineke Schuurman
Centre for Computational Linguistics, KU Leuven
Leuven, Belgium
ineke@ccl.kuleuven.be

Abstract

We present a set of Pictograph Translation Technologies, which automatically translates natural language text into pictographs, as well as pictograph sequences into natural language text. These translation technologies are combined with sentence simplification and an advanced spelling correction mechanism. The goal of these technologies is to enable people with a low level of literacy in a certain language to have access to information available in that language, and to allow these people to participate in online social life by writing natural language messages through pictographic input. The technologies and demonstration system will be added to the CLARIN infrastructure at the Dutch Language Institute in the course of this year, and have been presented on Tour De CLARIN.

1 Introduction

The set of Pictograph Translation Technologies we present consists of Text2Picto, which automatically converts natural language text (Dutch, English, Spanish) into a sequence of Sclera or Beta pictographs,[1] and of Picto2Text, which converts pictograph sequences into regular text (Dutch).

The use of these technologies was instigated by WAI-NOT,[2] a safe internet environment for users with cognitive disabilities, which often also have trouble reading or writing. It was further developed in the EU-funded Able-to-Include project, which built an accessibility layer, allowing software and app developers to build tools that can easily use a number of language technologies, such as the pictograph translation technologies, but also text-to-speech and text simplification. An example of such an application is the e-mail client developed by Saggion et al. (2017).

The Pictograph Translation Technologies for Dutch are further extended in a PhD project in which the tools are not only refined, but also evaluated by a group of targeted users. The initial version of Text2Picto is described in Vandeghinste et al. (2017).

The initial version of Picto2Text is described in Sevens et al. (2015). Refinements consist of the development of a dedicated pictograph selection interface, and of improved translation of pictograph sequences into natural language text through the use of machine translation techniques.

While the current version of the Pictograph Translation Technologies is running on the servers of the Centre for Computational Linguistics at KU Leuven, we are transferring these services to the *Instituut voor de Nederlandse Taal* (Dutch Language Institute), the CLARIN-B centre for Flanders, a region of Belgium which is a member of CLARIN through the flag of the Dutch Language Union (DLU). This transfer will ensure the longevity of the web service, and hence facilitate the ease of communication for people with reading and writing difficulties through the use of this web service beyond the end of the current research projects.

Furthermore, through the extra exposure the service receives as part of CLARIN, we hope to facilitate development of other language technology applications that can use the links between the pictograph sets and the WordNet (Miller, 1995) or Cornetto (Vossen et al., 2008) synsets, as described in Vandeghinste and Schuurman (2014). A demo of the system and its components can be found at its original location at http://picto.ccl.kuleuven.be/DemoShowcase.html

[1] http://www.sclera.be and http://www.betasymbols.com for more information about the pictograph sets.
[2] http://www.wai-not.be

In what follows we give a brief overview of related work, the system description and the evaluation by the target groups, before we conclude.

2 Related Work

We found only few works related to translating texts for pictograph-supported communication in the literature. Mihalcea and Leong (2009) describe a system for the automatic construction of pictorial representations of the nouns and some verbs for simple sentences and show that the understanding, which can be achieved using visual descriptions, is similar to those of target-language texts obtained by means of machine translation.

Goldberg et al. (2008) show how to improve understanding of a sequence of pictographs by conveniently structuring its representation after identifying the different roles which the phrases in the original sentence play with respect to the verb (structured semantic role labelling is used for this).

Joshi et al. (2006) describe an unsupervised approach for automatically adding pictures to a story. They extract semantic keywords from a story and search an annotated image database. They do not try to translate the entire story.

Vandeghinste and Schuurman (2014) describe the linking of Sclera pictographs with synonym sets in the Cornetto lexical-semantic database. Similar resources are PicNet (Borman et al., 2005) and ImageNet (Deng et al., 2009), both large-scale repositories of images linked to WordNet (Miller 1995), aiming to populate the majority of the WordNet synsets. These often contain photographs which might be less suitable for communication aids for the cognitively challenged, as they may lack clarity and contrast. The Sclera and Beta pictograph sets are specifically designed to facilitate communication with this user group.

There exist a number of systems that translate pictographs into natural language text (Vaillant 1998; Bhattacharya and Basu, 2009; Ding et al., 2015).[3] Most of these language generation tools expect grammatically or semantically complete pictograph input and they are not able to generate natural language text if not all the required grammatical or semantic roles are provided by the user.

3 System Architecture

Both translation directions make use of the hand-made links between pictographs and Cornetto synsets. Pictographs are linked to one or more Cornetto synsets, indicating the meaning they represent. This has been done for the Sclera and for the Beta set.

3.1 Text2Picto

The first version of this system is described in Vandeghinste et al. (2017). The input text goes through shallow syntactic analysis (sentence detection, tokenization, PoS-tagging, lemmatization, for Dutch: separable verb detection) and each input word is looked up, either in a dictionary (e.g. for pronouns, greetings and other word categories which are not contained in Cornetto) or in Cornetto.

Once the synsets that indicate the meaning of the words in the sentence are identified, the system retrieves the pictographs attached to these synsets. If no pictographs are attached to these synsets, the system uses the relations between synsets (such as hyperonymy, antonymy, and xpos-synonymy) in order to retrieve nearby pictographs. An A* algorithm retrieves the best matching pictograph sequence.

The system was further refined, integrating sentence simplification (Sevens et al., 2017b), as long sequences of pictographs are hard to interpret, temporal detection, as pictograph sequences are usually undefined for morpho-syntacic features and conjugation, spelling correction tuned to the specific user group (Sevens et al., 2016b), which has its own spelling error profile, and proper word sense disambiguation (Sevens et al. 2016a), which identifies the correct sense of polysemous words and retrieves the correct pictograph for that sense.

[3]We do not consider systems that generate the pictographs' labels/lemmas instead of natural language text, or systems that require users (with a motor disability) to choose the correct inflected forms themselves.

3.2 Picto2Text

In the Picto2Text application we have to distinguish the pictograph selection interface from the actual Picto2Text translation engine.

The pictograph selection interface (Sevens et al., 2017a) is a three-level category system. For both Beta and Sclera, there are 12 top categories, which consist of 3 to 12 subcategories each. A total of 1,660 Beta pictographs and 2,181 Sclera pictographs are included, meaning that an average of 21 (for Beta) and 28 (for Sclera) pictographs can be found within each subcategory. The choice for the top-level categories is motivated by the results of a Latent Dirichlet Allocation analysis applied to the WAI-NOT corpus of nearly 70,000 e-mails sent within the WAI-NOT environment.The following categories were created: *conversation, feelings and behaviour, temporal and spatial dimensions, people, animals, leisure, locations, clothing, nature, food and drinks, objects,* and *traffic and vehicles*. The subcategories were largely formed by exploring Cornetto's hyperonymy relations between concepts. Pictographs occurring within each subcategory are assigned manually. They are ordered in accordance with their frequency of use in the WAI-NOT email corpus, with the exception of logical ordering of numbers (1, 2, 3, ...) and months (January, February, March,...), pairs of antonyms (small and big), or concepts that are closely related. Note that some pictographs can appear in different subcategories and that only pictographs are available that match words that occur more than 50 times in the WAI-NOT corpus.

The Picto2Text translation engine is still under development. The initial system (Sevens et al., 2015) takes a sequence of pictograph names as input, retrieves the synsets to which these pictographs are linked, and generates the full morphological paradigm for each of the lemmas that form that synset. A trigram language model trained on a large corpus of Dutch is used to determine the most likely sequences. In later versions, we are using a fivegram language model trained on a more specific data set, and we are comparing these models with long short-term memory models (LSTM) recurrent neural language models, but have not found improvements yet. A different and promising approach we are pursuing is the use of machine translation tools, such as Moses (Koehn et al., 2007) and OpenNMT (Klein et al., 2017), trained on an artificial parallel corpus, for which the source side (the pictograph side) was automatically generated through the use of the Text2Picto tool, described in section 3.1.[4]

4 Evaluation

Each of the components of the Pictograph Translation Technologies have been evaluated by the users, in two iterations. The first systems have been evaluated through user observations and focus groups. The conclusions of these evaluations were used to make improvements for the second versions, which are currently being re-evaluated. A detailed description of the evaluations is given in Sevens et al. (in press). In general these technologies enable the use of textual communication via the internet for people with limited literacy.

5 Conclusions

The Pictograph Translation Technologies, which allow people with reading and/or writing difficulties to participate in the written society are becoming available as a CLARIN tool. These technologies have been developed in such a way that they are easily extendible to other languages and other pictograph sets. They have been developed specifically for users with reading and writing difficulties in mind, but can also be useful for other user groups, in order to resolve communication difficulties, such as migrants that have not learned the language of their host country (yet).

References

[Bhattacharya and Basu 2009] Bhattacharya, S. and Basu, A. (2009). Design of an Iconic Communication Aid for Individuals in India With Speech and Motion Impairments. In *Assistive Technology, n° 21(4)*: 173–187.

[Borman et al. 2005] Borman, A., Mihalcea, R., and Tarau, P. (2005). PicNet: augmenting semantic resources with pictorial representations. In: *Proceedings of the AAAI Spring Symposium on Knowledge Collection from Volunteer Contributors*, pp. 1–7. Menlo Park, California.

[4] This parallel corpus will be made available through the CLARIN infrastructure.

[Deng et al. 2009] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: a large-scale hierarchical image database. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Miami, FL, pp. 248–255.

[Ding et al. 2015] Ding C., Halabi N., Alzaben L., Li Y., Draffan E.A. and Wald M. (2015). A Web based Multi-Linguists Symbol-to-Text AAC Application. In *Proceedings of the 12th Web for All Conference.*

[Goldberg et al. 2008] Goldberg, A., Zhu, X., Dyer, C. R., Eldawy, N., and Heng, L. (2008). Easy as ABC? Facilitating pictorial communication via semantically enhanced layout. In *Proceedings of the 12th Conference on Computational Natural Language Learning (CoNLL)*, Manchester, England, pp. 119–126.

[Klein et al. 2017] Klein, G., Kim, Y., Deng, Y., Senellart, and Rush, A.M. (2017). OpenNMT: Open-Source Toolkit for Neural Machine Translation. *ArXiv e-prints* 1701.02810.

[Koehn et al. 2007] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A. and Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. *Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session.* Prague, Czech Republic.

[Joshi et al. 2006] Joshi, D., Wang, J., and Li, J. (2006). The story picturing engine — a system for automatic text illustration. *ACM Transactions on Multimedia Computing, Communications and Applications 2(1):* 1–22.

[Mihalcea and Leong 2009] Mihalcea, R., and Leong, C. W. (2009). Toward communicating simple sentences using pictorial representations. *Machine Translation 22(3)*: 153–173.

[Miller 1995] Miller, G. A. (1995). Wordnet: A lexical database for English. *Communications of the ACM 38(11)*: 39–41.

[Saggion et al. 2017] Saggion, H., Ferrés, D., Sevens, L. and Schuurman, I. (2017). Able to Read my Mail: An Accessible E-mail Client with Assistive Technology. In: *Proceedings of the 14th International Web for All Conference (W4A'17)*. Perth, Australia.

[Sevens et al. 2015] Sevens, L., Vandeghinste, V., Schuurman, I. and Van Eynde, F. (2015). Natural Language Generation from Pictographs. In: *Proceedings of 15th European Workshop on Natural Language Generation (ENLG 2015)*. Brighton, UK. pp. 71-75.

[Sevens et al. 2016a] Sevens, L., Jacobs, G., Vandeghinste, V., Schuurman, I. and Van Eynde, F. (2016a). Improving Text-to-Pictograph Translation Through Word Sense Disambiguation. In: *Proceedings of the 5th Joint Conference on Lexical and Computational Semantics*. Berlin, Germany.

[Sevens et al. 2016b] Sevens, L., Vanallemeersch, T., Schuurman, I., Vandeghinste, V. and Van Eynde F. (2016b). Automated Spelling Correction for Dutch Internet Users with Intellectual Disabilities. In: *Proceedings of 1st Workshop on Improving Social Inclusion using NLP: Tools and Resources (ISI-NLP, LREC workshop).* Portorož, Slovenia, pp. 11-19.

[Sevens et al. 2017a] Sevens, L., Daems, J., De Vliegher, A., Schuurman, I., Vandeghinste, V. and Van Eynde, F. (2017a). Building an Accessible Pictograph Interface for Users with Intellectual Disabilities. In: *Proceedings of the 2017 AAATE Congress*. Sheffield, UK.

[Sevens et al. 2017b] Sevens, L., Vandeghinste, V., Schuurman, I. and Van Eynde, F. (2017b). Simplified Text-to-Pictograph Translation for People with Intellectual Disabilities. In: *Proceedings of the 22nd International Conference on Natural Language & Information Systems (NLDB 2017)*. Liège, Belgium.

[Sevens et al. in press] Sevens, L., Vandeghinste, V., Schuurman, I. and Van Eynde F. (in press). Involving People with an Intellectual Disability in the Development of Pictograph Translation Technologies for Social Media Use. In: *Cahiers du CENTAL, Volume 8*. Louvain-La-Neuve, Belgium.

[Vaillant 1998] Vaillant P. (1998). Interpretation of iconic utterances based on contents representation: Semantic analysis in the PVI system. *Natural Language Engineering, n 4(1)*: 17–40.

[Vandeghinste and Schuurman 2014] Vandeghinste, V. and Schuurman, I. (2014). Linking Pictographs to Synsets: Sclera2Cornetto. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland. pp. 3404-3410.

[Vandeghinste et al. 2017] Vandeghinste, V., Schuurman, I., Sevens, L. and Van Eynde, F. (2017). Translating Text into Pictographs. *Natural Language Engineering 23 (2):* 217-244.

[Vossen et al. 2008] Vossen, P., Maks, I., Segers, R., and van der Vliet, H. (2008). Integrating lexical units, synsets, and ontology in the Cornetto Database. In: *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC'08)*, pp. 1006–13, Marrakech, Morocco.