Selected papers from the

CLARIN Annual Conference 2016

Aix-en-Provence, 26–28 October 2016

Linköping Electronic Conference Proceedings 136
eISSN 1650-3740 (Online) • ISSN 1650-3686 (Print) 2017
ISBN 978-91-7685-499-0

Selected papers from the

CLARIN Annual Conference 2016

Aix-en-Provence, 26–28 October 2016

edited by Lars Borin

Front cover illustration:
Group photo by Quirijn Backx, picture composition by Karolina Badzmierowska • CLARIN ERIC

Linköping Electronic Conference Proceedings 136
eISSN 1650-3740 (Online) • ISSN 1650-3686 (Print) 2017
ISBN 978-91-7685-499-0

Preface

These proceedings present the highlights of the CLARIN Annual Conference 2016 that took place in Aix-en-Provence, France. As the third volume in the proceedings series it illustrates that CLARIN has developed into a community with an ambition that goes beyond the realization of a research infrastructure for language resources. The multiannual record of CLARIN's progression also demonstrates the ambition of coupling a data infrastructure to a knowledge sharing infrastructure: a common platform that will guide researchers from the Humanities and Social Sciences in making optimal use of the infrastructure and benefitting from the experience of other researchers and the best practices applied across the case studies collected.

The papers selected for this volume present the results of a number of projects conducted within and between CLARIN's national consortia, but the proceedings of 2016 also contain papers with contributions from authors outside the CLARIN consortium. Furthermore, the fact that France generously hosted the conference already in the year before it actually joined CLARIN as an observer highlights the potential for growth and for pan-European collaboration.

CLARIN provides sustainable access to language resources in all forms, analysis services for the processing of language materials, and a platform that can stimulate the use, reuse and repurposing of the available data. This contributes to realizing the vision associated with the Open Science agenda and to strengthening Europe's capacity to lower the barriers for researchers to entry digital scholarship and cutting edge research. As shown by the range of topics addressed in the proceedings, language resources can play a multitude of roles, including carrier of information, record of the past, means of literary expression, social signal, or object of linguistic study.

Due to the diversity of the data types supported, the communities of use to be served by CLARIN are also diverse. Combined with the multitude of languages covered, CLARIN can help to realize a multilingual European Research Area for digital research in the Humanities and Social Sciences, to turn Europe's multilingualism into a basis for the comparative investigation of a wide range of intellectual and societal phenomena and to ensure that the multidisciplinary research agendas addressing societal challenges will have impact.

The CLARIN Annual Conference is one of the communication instruments between those who build and maintain the infrastructure, those who provide data and tools, and those who use the CLARIN infrastructure in their scholarly projects. A similar role is played by the workshops focused on specific data types organized in 2016 and 2017. All these CLARIN events have demonstrated the importance of coordination in sharing the insights into problems, solutions, failures and successes across national and linguistic borders. Hopefully this volume will help to attract new categories of scholars, with ideas and requirements for use cases that can help us identify the directions and next steps to take in the further development of the CLARIN infrastructure as a pillar of Europe's Open Science policies.

Utrecht, 14 May 2017

Franciska de Jong
Executive Director CLARIN ERIC

Franciska de Jong 2017. Preface. *Selected papers from the CLARIN Annual Conference 2016.* Linköping Electronic Conference Proceedings 136: i–i.

Introduction

This volume contains a selection of papers presented at the CLARIN Annual Conference 2016 which was held in Aix-en-Provence, France, on 26–28 October 2016.

This was the fifth edition of the conference. It started in 2012 as an internal event, where members of the national CLARIN consortia came together to share their experiences of and thoughts on the development of the CLARIN ERIC infrastructure.

In 2014, it was felt that the time was ripe to change the format of the conference into an event with an open call for contributions, in order to include also the humanities and social-science research communities – the intended users of the infrastructure – in the exchange of ideas and experiences on the CLARIN infrastructure. This includes its design, construction and operation, the data and services that it contains or should contain, its actual use by researchers, its relation to other infrastructures and projects, and the CLARIN Knowledge Sharing Infrastructure.

As a result of the 2016 call for papers we received 34 anonymous extended abstracts, each of which was anonymously reviewed by at least three members of the program committee, which as always consisted of the members of the CLARIN ERIC National Coordinators' Forum, i.e., one member from each participating country or NGO. In order to avoid conflicts of interest, no PC member reviewed submissions from their own country. As a result of the reviewing process, a total of 25 submissions were accepted for presentation at the conference, 14 as oral presentations and 11 as posters.

In addition to the submitted presentations, the conference featured two invited speakers. The keynote on the first day was presented by professor Ian Gregory from Lancaster University, under the title *Texts, language and geography: Understanding literature using geographical text analysis*, and on the second day, professor Sally Wyatt, Maastricht University talked about *Why technologies are not neutral, and why it matters for linguists*.

As a new feature, the CLARIN 2016 call for papers included a call for submissions to a thematic session, focusing on *Language resources and historical sources*. The general area of interest for the thematic session was stated in the call for papers as CLARIN-related research in the historical sciences, understood in a wide sense to encompass fields such as History, "History of ..."/"... history" (e.g., History of science, Rhetorical history), as well as the various historically oriented subfields of linguistics (e.g., Historical linguistics, Historical pragmatics, etc.), and philology. We invited submissions on two separate but overlapping aspects that we construed this theme to encompass:

(1) The historical aspect in a narrower sense: Processing historical language stages in the form of text or speech, with the concomitant issues of digitization, non-standardized language, etc.

(2) The diachronic aspect: Discovering, characterizing and tracking change through time, both linguistic changes and changes in the world as reflected in the content of text.

Two of the oral presentations and several posters addressed this theme. Ian Gregory's keynote speech together with the two oral presentations were organized into a thematic session scheduled at the very beginning of the conference program.

The conference was video recorded; see the YouTube playlist:
https://www.youtube.com/playlist?list=PLlKmS5dTMgw2pP-uvhKNVSgOuuZjvmLwy

Following the conference, authors of the accepted papers were invited to submit full versions of their papers to be considered for the conference proceedings volume,

Lars Borin 2017. Introduction. *Selected papers from the CLARIN Annual Conference 2016*. Linköping Electronic Conference Proceedings 136: iii–iv.

although this time the submissions were not anonymous. Again the papers were reviewed (anonymously) by two to four PC members, at least one of which had not reviewed the original abstract submitted for the conference. We received 14 full-length submissions, out of which 10 were accepted for this volume. Most of these address core CLARIN issues dealing with the construction, maintenance and use of the European infrastructure coordinated in the framework of the CLARIN ERIC, such as search engine design, resource discovery, metadata quality, researcher training in infrastructure use, and design of specific tools and resources. There is one "pure" research paper in this volume – by Hinrichs, Erdmann and Joseph – but many of the contributions refer to research conducted using the CLARIN infrastructure. In two cases – the papers by Beißwenger et al. and by MacWhinney – the focus is on resource-building with specific research questions or a specific research field in mind, where the research and infrastructure-building activities feed into each other and actually become hard to disentangle.

I would like to thank the reviewers for the dedicated efforts they put down in evaluating the submissions, and also Peter Berkesand at Linköping University Electronic Press, who (as usual) has ensured that the digital publication of this volume went smoothly and painlessly.

Lars Borin
University of Gothenburg
Program committee chair

Program committee for the CLARIN Annual Conference 2016

Members of the CLARIN ERIC National Coordinators' Forum

Jan Theo Bakker
Lars Borin
António Branco
Koenraad De Smedt
Tomaž Erjavec
Eva Hajičová
Erhard Hinrichs
Krister Lindén
Bente Maegaard
Monica Monachini
Karlheinz Mörth
Jan Odijk
Maciej Piasecki
Stelios Piperidis
Kiril Simov
Inguna Skadiņa
Jurgita Vaičenonienė
Kadri Vider
Martin Wynne

Additional reviewer

Paul Meurer

Contents

Closing a Gap in the Language Resources Landscape: Groundwork and Best Practices from Projects on Computer-mediated Communication in four European Countries

Michael Beißwenger
University of Duisburg-Essen, Germany
michael.beisswenger@uni-due.de

Thierry Chanier
Université Clermont Auvergne, France
thierry.chanier@univ-bpclermont.fr

Tomaž Erjavec
Jožef Stefan Institute
Ljubljana, Slovenia
tomaz.erjavec@ijs.si

Darja Fišer
University of Ljubljana
Ljubljana, Slovenia
darja.fiser@ff.uni-lj.si

Axel Herold
Berlin-Brandenburg Academy of Sciences, Berlin, Germany
herold@bbaw.de

Nikola Ljubešić
Jožef Stefan Institute
Ljubljana, Slovenia
nikola.ljubesic@ffzg.hr

Harald Lüngen
Institute for the German Language, Mannheim, Germany
luengen@ids-mannheim.de

Céline Poudat
Université de Nice Sophia Antipolis
France
poudat@unice.fr

Egon Stemle
Eurac Research
Bolzano
Italy
egon.stemle@eurac.edu

Angelika Storrer
University of Mannheim, Mannheim, Germany
astorrer@mail.uni-mannheim.de

Ciara Wigham
Université Clermont Auvergne, France
ciara.wigham@uca.fr

Abstract

The paper presents best practices and results from projects dedicated to the creation of corpora of computer-mediated communication and social media interactions (CMC) from four different countries. Even though there are still many open issues related to building and annotating corpora of this type, there already exists a range of tested solutions which may serve as a starting point for a comprehensive discussion on how future standards for CMC corpora could (and should) be shaped like.

1 Introduction

The paper presents best practices and results from projects dedicated to the creation of corpora of computer-mediated communication and social media interactions (henceforth referred to as *CMC*) from four European countries. The projects are inter-related via a bottom-up network of researchers interested in fostering the transfer of expertise and solutions for handling this relatively new type of language resources and for modeling the structural and linguistic peculiarities of (written and multimodal) discourse found in chat, forum, sms and whatsapp interactions, in weblogs and wikis, on social network sites and in multimodal CMC environments. This new type of discourse exhibits features that cannot be adequately handled by the schemas and tools which have been developed for

Michael Beißwenger, Thierry Chanier, Tomaž Erjavec, Darja Fišer, Axel Herold, Nikola Ljubešić, Harald Lüngen, Céline Poudat, Egon Stemle, Angelika Storrer and Ciara Wigham 2017. Closing a gap in the language resources landscape: Groundwork and best practices from projects on computer-mediated communication in four European countries. *Selected papers from the CLARIN Annual Conference 2016.* Linköping Electronic Conference Proceedings 136: 1–18.

the representation, annotation and processing of discourse which conforms to the written standard and the structural conventions of established text types (e.g., newspaper articles, prose, scientific articles). In addition with the collection and redistribution of CMC data in linguistic corpora, legal and ethical issues arise which are not yet sufficiently covered by existing laws and ethical standards. What is more, there are no established standards for metadata and for the documentation of the (technological, hypermedial and social) context in which CMC data are typically embedded, produced and used.

Corpus-linguistic approaches to CMC have so far not found answers to all of these challenges. Nevertheless, existing projects in the field have proposed and tested an encouraging range of solutions and best practices. The joint goal of the projects and initiatives described in this paper is to pave the ground for standards which will allow CMC corpora to be interoperable (a) with each other and (b) with language resources for other types of discourse (text and speech corpora).

The paper is structured as follows: Section 2 gives an overview of existing CMC corpora and corpus projects. Section 3 describes two initiatives dedicated to the development of standards and to the exchange of knowledge related to the collection, annotation, representation and provision of CMC corpora. Section 4 gives an overview of the results and best practices from CMC corpus projects in four countries which may be useful for other projects in the field and which may serve as a starting point for a more comprehensive discussion on how future standards for CMC corpora could (and should) be shaped like.

2 Overview of CMC corpora and corpus projects

Even though research on CMC in linguistics and social sciences from its very beginning had a strong empirical focus, only few corpora or datasets have been made available to the scientific public. An overview of CMC corpora is given in Beißwenger and Storrer (2008). Examples of ‘early-bird’ CMC corpora are:

- the *NPS Chat Corpus* for English (Forsyth and Martell, 2007) with 45.000 tokens from age-specific chat rooms which have been annotated with part-of-speech information and a dialog-act classification. The corpus is available via the Linguistic Data Consortium (LDC).
- The *Dortmund Chat Corpus* for German (Beißwenger, 2013) which comprises 1 million tokens of chat discourse with annotations of selected CMC-specific phenomena. The corpus is available for free download since 2005[1] and will be released in an enhanced version as part of CLARIN-D in spring 2017 (cf. Sect. 4.2).

More recently, a range of projects has created (or is currently creating) resources which have been or will be made available to the public – for example (in alphabetical order):

- *CoMeRe*: a collection of 14 French corpora for 9 different CMC genres represented in TEI, available for download via ORTOLANG (cf. Sect. 4.1).[2]
- *CorCenCC-CMC*: The “e-language” component in the project “National Corpus of Contemporary Welsh” (CorCenCC, since 2016).[3]
- *DEREKO-News*: Corpus of German Newsgroups in DEREKO, since 2013, 98 million tokens, available for online querying via COSMAS II (Schröck and Lüngen, 2015).[4]
- *DEREKO-Wikipedia*: Wikipedia corpora in DEREKO: German language article talk and user talk (cf. Margaretha and Lüngen, 2014), 581 million tokens, available for online querying via COSMAS II; also downloadable.
- *DiDi corpus*: The CMC corpus from the DiDi project with 570.000 tokens of German, Italian and South Tyrolean Facebook posts and interactions, available for online querying via ANNIS (Frey et al., 2016; cf. Sect. 4.3).[5]
- *DWDS blog corpus*: The blog corpus in the corpus collection of the DWDS project: 103 million tokens from CC-licensed, mainly German blog entries, available for online querying.[6]

[1] http://chatkorpus.tu-dortmund.de/
[2] http://hdl.handle.net/11403/comere
[3] Project page: http://sites.cardiff.ac.uk/corcencc/
[4] https://cosmas2.ids-mannheim.de/
[5] http://www.eurac.edu/didi
[6] https://www.dwds.de

- *Janes*: The Corpus of Nonstandard Slovene comprising >200 million tokens from tweets, forum posts, blogs, comments on news articles and Wikipedia discussions (Fišer et al., 2016; cf. Sect. 4.4).[7]
- *sms4science.ch*: a donation-based corpus of 650.000 tokens of SMS messages collected in Switzerland and comprising discourse in non-dialectal German, French, Swiss German, Italian and Romansh (Dürscheid and Stark, 2011), available for online querying in a full text version (SMS Navigator) and as a partially annotated version represented in ANNIS.[8]
- *SoNaR-CMC*: the CMC component (chats, tweets and sms messages) in the Reference Corpus of Contemporary Dutch (SoNaR, Oostdijk et al., 2013) which is available for online querying via CLARIN-NL (OpenSoNaR).[9]
- *Suomi24*: a collection of 2.38 billion tokens of discourse from Finnish discussion forums with morpho-syntactic annotations, available for download.[10]
- *whatsup-switzerland.ch*: Corpus of the project "Whats's up, Switzerland?": a collection of 5 million tokens from 650 whatsapp chats donated by Swiss smart phone users.[11]
- *Web2Corpus_it*: a balanced CMC corpus for Italian (in preparation) including discourse from forums, blogs, newsgroups, social networks and chats (Chiari and Canzonetti, 2014) created in the context of a project on negotiation strategies.[12]

Even though the sheer availability of CMC corpora is already a big step ahead towards closing the "CMC gap" in the corpus landscape, the existing corpora, in their current state, are represented and provided using heterogeneous technologies, representation formats and annotation schemas. The availability of a flexible standard for the representation and exchange of CMC resources would allow researchers and corpus providers to combine, merge and connect their resources (*interoperability*), and facilitate corpus-based research across languages and CMC genres and beyond the limitations of single corpora. The creation of such a standard in compliance with the existing standards in the field of digital humanties would, additionally, allow to combine CMC corpora with corpora of other type (text corpora, speech corpora) and thus open up new perspectives also for corpus-based research on commonalities and differences between CMC discourse and monologic written language and spoken conversations. Moreover, compliance with existing standards would increase the *sustainability* and *reusability* of resources.

3 *cmc-corpora.org*: a European network of CMC corpus projects

Since 2013 a loose network of projects with a joint interest in building, annotating and analyzing CMC corpora has set up two initiatives in order to (1) strengthen the exchange of expertise and best practices between projects and (2) lead the discussion of a representation standard for CMC genres in the context of a well-acknowledged standardization initiative in the Digital Humanities:

3.1 Conference series on CMC corpora

The network has established a series of international workshops and conferences dedicated to the creation of CMC corpora with previous events held in Dortmund/DE (2013, 2014), Rennes/F (2015) and Ljubljana/SI (2016), and a next event (the *5th Conference on CMC and Social Media Corpora for the Humanities*) scheduled to be held in October 2017 at Eurac Research in Bolzano/IT. These conferences are defined as peer-reviewed events with a coordinating and a scientific committee.[13] Since 2016 the conferences are accompanied by peer-reviewed proceedings which are published online (cf. Fišer and Beißwenger, 2016).

[7] Project page: http://nl.ijs.si/janes/
[8] http://www.sms4science.ch
[9] https://portal.clarin.nl/node/4195
[10] http://urn.fi/urn:nbn:fi:lb-201412171
[11] http://www.whatsup-switzerland.ch/
[12] Project page: http://www.glottoweb.org/web2corpus/
[13] http://www.cmc-corpora.org

3.2 TEI special interest group on CMC

The network succeeded with a proposal for the creation of a special interest group (SIG) on Computer-Mediated Communication in the *Text Encoding Initiative* (*TEI*, http://tei-c.org) in 2013. The goal of this SIG is to extend the TEI framework with additions dedicated to the representation of the structural and linguistic peculiarities of CMC genres. Starting from a discussion of a first schema draft defined by Beißwenger et al. (2012) the SIG created two advanced schema drafts ('CoMeRe schema', 2014, 'CLARIN-D schema', 2015) which have been tested with French and German corpora and which are currently being adopted by other further projects. The schemas developed by the SIG are defined following the rules for *customization* described in the TEI guidelines[14]. The basic structure and CMC-specific models of the schemas have been discussed with the TEI community in several panels at the annual TEI conferences and members' meetings and will be presented to the TEI Technical Council in the form of feature requests, i.e. suggestions for the extension of the "official" TEI standard. The latest version of the schema which builds on its predecessors is described in Sect. 4.2.2.

4 Groundwork and best practices from projects in Germany, France, Italy and Slovenia

4.1 The CLARIN-D curation project *ChatCorpus2CLARIN* (Germany)

4.1.1 Project description

In the project *ChatCorpus2CLARIN*, an existing chat corpus for German (the *Dortmund Chat Corpus*, Beißwenger, 2013) served as a use case to demonstrate how an integration of CMC and social media resources into the CLARIN-D corpus infrastructures could be accomplished in a way that the target resource (1) conforms to established standards for the representation and linguistic annotation of corpora in the Digital Humanities context and (2) can be a useful resource for doing comparative analyses of CMC discourse with other types of corpus resources in CLARIN-D (text and speech corpora). The original resource has been compiled in 2002–2005 and comprises 1 million tokens of German chat discourse from various domains (social chat, chat in the context of learning and teaching, advisory chats, chats in the media context). The data is represented using a 'homegrown' XML format which describes (different types of) individual user posts, selected linguistic phenomena (such as emoticons, addressing terms, action words and acronyms) and selected metadata about the chats and their participants. The corpus has been available online for download since 2005.[15] It has been used as a resource in a broad range of research and teaching contexts in linguistics and language technology.

In the project, the original resource was remodeled building on schema drafts from the TEI CMC-SIG (Sect. 3) to increase its interoperability with other types of corpora provided via CLARIN-D. To extend the research and query options for the target resource the corpus, in addition, was enhanced with a layer of linguistic annotations (tokens, parts of speech, lemmas).

The project was headed by Michael Beißwenger (U Dortmund) and Angelika Storrer (U Mannheim). Researchers from the CLARIN-D hubs at the Institute for the German Language (IDS), Mannheim (Harald Lüngen), and from the Berlin-Brandenburg Academy of Sciences (Axel Herold) were closely involved into all work packages of the project. A visualization of the workflow and resources used in the integration process is given in Figure 1 and described in detail in Lüngen et al. (2016).

[14] http://www.tei-c.org/release/doc/tei-p5-doc/en/html/USE.html

[15] http://chatkorpus.tu-dortmund.de/

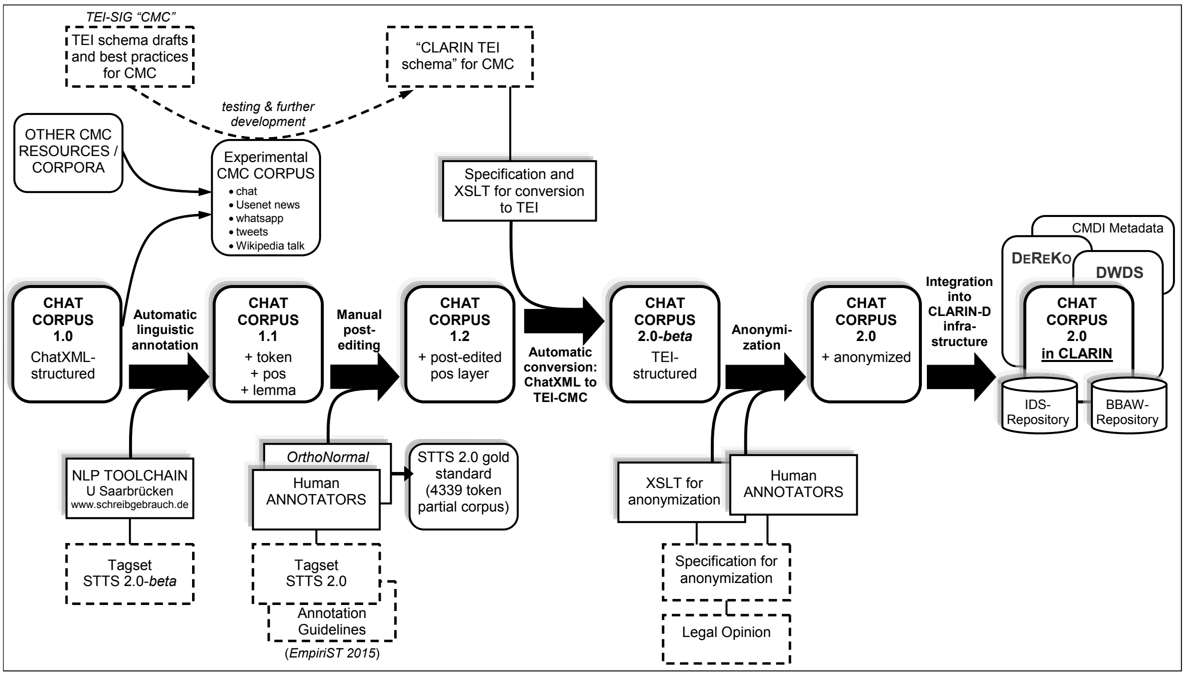


Figure 1: Integration of the Dortmund Chat Corpus into the CLARIN-D corpus infrastructure.

Main resources and work packages in the project workflow were:

- An **experimental CMC corpus:** For developing and testing the solutions developed for representing and annotating the corpus, we compiled a small experimental corpus with data from several CMC and social media genres (chat, news messages, Wikipedia talk pages, tweets, whatsapp interactions). This was done to guarantee that the annotation schema and tagset are useful not only for chat but also for a range of other types of (mainly) written CMC genres.
- **Linguistic annotation:** Tokenization, part-of-speech (PoS) tagging and lemmatization were done in two stages: (1) an automatic tagging process done at Saarland University applying the NLP toolchain described in Horbach et al. (2014) and (2) a manual post-editing phase with two trained annotators for a part of the resource (to demonstrate how a 'gold' annotation for chat data could look like).
- **The 'STTS 2.0' Part-of-Speech Tagset:** As target standard for the PoS layer, we adopted the tag set ('STTS 2.0'; Beißwenger et al., 2015) developed in the GSCL shared task on automatic linguistic annotation of CMC and social media (EmpiriST2015; Beißwenger et al., 2016)[16]. 'STTS 2.0' builds on the categories of the "Stuttgart-Tübingen Tagset" (*STTS*, Schiller et al., 1999) and introduces two types of new tags: (1) tags for phenomena which are specific for CMC and social media discourse, (2) tags for phenomena which are typical of spontaneous spoken language in colloquial registers and which can also be found in corpora of transcribed speech (e.g., in the FOLK corpus of spoken language at the IDS which uses an STTS extension which is compatible with 'STTS 2.0', Westpfahl and Schmidt, 2016). The resulting tag set is still downwardly compatible with STTS (1999) and therefore allows for interoperability with other corpora that have been tagged with STTS.
- **Legal clearance and anonymization:** Prior to the integration of the curated resource in CLARIN infrastructures, we sought a legal opinion to get a better picture of the legal conditions for republishing the material as a whole or in parts. The legal opinion (iRights.Law, 2016) carefully checked for possible restrictions arising from individual property rights, copyrights and other legal statutes. One result was that the possibility to identify individuals from their utterances (with the exception of public figures) needed to be circumvented by means of an anonymization. Only parts of the anonymization task could be done automatically; occurrences of names that had not been annotated in the original source,

[16] http://sites.google.com/site/empirist2015/

or that could not be matched to entries in the participant list automatically, had to be anonymized manually which was a very time-consuming process so that the date for the release of the integrated resource had to be postponed to spring 2017.

- **Development of a schema for remodeling the chat corpus in TEI:** To achieve interoperability with a broad range of other language resources in the digital humanities, the original resource was converted into a TEI format using customizations. Main features of the TEI schema developed in the project are outlined in Sect. 4.2.2.

The TEI schema, PoS tagset and anonymization guidelines will be reused and refined in follow-up CMC corpus projects within CLARIN-D – e.g. for representing and annotating data in the project *MoCoDa* (*Mobile Communication Database*) in which a database and web frontend for the repeated collection of data donations from whatsapp, sms and similar CMC 'apps' for mobile use will be created. The project which started in January 2017 is funded by the Ministry for Innovation, Science, Research and Technology of the German federal state North Rhine-Westfalia and by Michael Beißwenger (U Duisburg-Essen), Wolfgang Imo (U Halle-Wittenberg) and Evelyn Ziegler (U Duisburg-Essen).

4.1.2 Results beyond the resource: The 'CLARIN-D TEI schema' for CMC

The TEI schema developed in the project (the 'CLARIN-D TEI schema') is the result of a continued development based on two previous schemas for the representation of CMC discourse: the schema suggested by Beißwenger et al. (2012) and the schema developed in the CoMeRe project (Chanier et al., 2014; cf. Sect. 4.1.2). The development of the schema was fostered by extensive discussions within the TEI CMC-SIG (Sect. 3) and by discussions within the DFG scientific network *Empirikom*[17]. While aiming at providing a generic model for CMC discourse within the framework of the TEI, the CLARIN-D schema focuses on the representation of discourse captured in chat logfiles, whatsapp interactions, tweets and Wikipedia discussions. To amend the TEI guidelines (TEI-P5) for this CMC genre and reflect properties specific to logfiles, different types of customizations of the TEI guidelines had to be implemented:

(1) **New elements:** Three new elements were introduced to cater for building-blocks of computer-mediated interactions not yet covered by the TEI guidelines, namely <post> (which has first been described in the 'DeRiK TEI schema', Beißwenger et al., 2012 and see figure 2) and <prod> (originally introduced in the CoMeRe schema, Chanier et al., 2014). <post> is used to represent any written contribution to an ongoing CMC interaction which (1) has been composed by its author in its entirety as part of a private activity and, subsequently, (2) has been sent to the server *en bloc*. In contrast to <post>, <prod> represents *non-verbal* acts within a CMC environment (for details cf. Sect. 4.2.2). As another new element, we introduced <signatureContent> to allow for the unified representation of (most often automatically created) user signatures. This element may occur in meta-data descriptions of the discourse participants.

(2) **New attributes:** A new binary attribute @auto ('automatically generated') was introduced to better reflect the influence of the communication system on the discourse. In many CMC systems, non-verbal actions of participants may result in automatically generated verbal messages, e.g. the insertion of quoted material when hitting a "reply" button, the insertion of signatures into posts, or the generation of status messages. In combination with TEI's @who attribute, fine grained modeling of a message's creation context becomes possible. Because computer-assisted writing or collaborative writing in CMC may lead to parts of messages being produced by different participants, the exploitation of @who was allowed within a wider range of elements than is accepted in TEI proper (see figure 2).

[17] http://www.empirikom.net

```
<post xml:id="m645" who="#A02" synch="#t058" type="standard" auto="false">
   <note auto="true" who="#A02">for all</note>
   <anchor type="sentence_start"/>
   <ref type="addressingTerm" corresp="#A27">
      <w xml:id="m645.t1" type="ADV" lemma="nun">nun</w>
      <w xml:id="m645.t2" type="VVFIN" lemma="bitten">bitte</w>
      <w xml:id="m645.t3" type="NE" lemma="[_FEMALE-STUDENT-A27_]">[_FEMALE-STUDENT-A27_]</w>
      <w xml:id="m645.t4" type="$." lemma="!">!</w>
   </ref>
   <time> 16:48 </time>
</post>
```

Figure 2: CLARIN-D TEI snippet encoding a chat message, demonstrating the use of <post> and custom attributes. The attributes @who, @corresp, and @synch point to the list of participants and the timeline, respectively, in the TEI header.

(3) **Adaptation and extension of content models:** The content model of the generic <s> (sentence), <p> (paragraph), and <quote> elements was extended to allow for sub-elements such as <closer>, <signed>, or <postscript> to occur in a wider range of contexts than envisioned by the TEI. In CMC discourse, these types of text structure tend to be used without the rigid positional constraints found e.g. in traditional books and letters. The content model of some of the elements containing e.g. TEI's <p> and <s> elements was adapted to allow for combining these elements with the newly introduced elements as well as less restricted use of these elements in their typical contexts.

In addition to these customizations, we have defined best practices for using the TEI-P5 models <w>, <phr>, <signed>, <time>, <div>, <name> and other elements for annotating CMC phenomena and for adding part-of-speech information for every word token. Best practices have also been proposed for metadata modeling the discourse level as well as on the level of individual posts.

4.2 The *CoMeRe* project (France)

4.2.1 Project description

The *CoMeRe* project ('Communication Médiée par les Réseaux', supported in 2013-2015 by the National Written Corpora Consortium *IRCE*[18]) brought together researchers who had previously collected different types of CMC corpora in their local research teams or in previous research projects, and had structured these in a variety of formats (different XML schemas for text chat corpora, SMS corpora, and for LEarning and TEaching Corpora (LETEC)).

The primary aim of CoMeRe was to design a common model for CMC discourse that would fit the pre-existing CMC corpora, as well as new corpora collected both during the project or post-project. The secondary aim was to release these corpora in a common repository as open data, in order to provide access to a dataset with significant coverage to researchers interested in the linguistic study of CMC genres.

To address the project's primary goal, it was first necessary to develop a common document model that would fit different types of multimodal CMC data, the TEI CoMeRe schema (2014). All the 14 corpora stored in the CoMeRe repository (2016) have been structured according to this schema. To ensure open access to everyone, the data were collected from sources with appropriate licenses, anonymized, and the corpora were released under Creative Common licenses with the least possible constraints for reuse.

4.2.2 Results beyond the resource: models for representing multimodal CMC in TEI

The opportunity to collect various types of CMC corpora in different formats led us to develop a uniform format complying with the TEI-CMC SIG (Sect. 3). The CoMeRe schema had to be compatible with various genres, including sms, wiki discussions, tweets, weblogs, emails, discussion forums, text chats, oral and multimodal interactions, and multimodal interactions in 3D environments.

For a part of these genres (such as text chat or sms interactions) the users' interactions may be encoded in a way similar to the encoding suggested by Beißwenger et al. (2012), directly relying on the new <post> element (Sect. 4.1.2). For other corpora based on textual interactions, it has been

[18] http://corpusecrits.huma-num.fr/

necessary to enrich the <post> element with extra attributes such as explicit references to previous posts (email, discussion forum, weblog, wiki discussions), or to add sub-elements which describe specific structures encountered within the message contents (tweets). Since the LETEC (LEarning & TEaching) corpora had the most complex structure (Chanier and Wigham, 2016), they served as a basis to develop the CoMeRe schema. A LETEC corpus is a structured entity containing all the elements resulting from an online learning situation whose context is described by an educational scenario and a research protocol. The core data collection includes all the CMC interaction data, the course participants' productions, and the tracks, resulting from the participants' actions in the learning environment.

Indeed, LETEC participants in a course generally used several CMC tools to communicate over a period of 8 to 10 weeks. Participants resorted to various written synchronous and asynchronous communication tools, including emails, text chats, and discussion forums. The challenge was to organize the various interactions in a coherent way within the corpus structure which was the reason to develop the notion of *Interaction Space* (*IS*), with participants interacting on similar subjects using different tools within a time frame. The CoMeRe project members all agreed to adopt this concept, detailed in Chanier et al. (2014), which was generic enough to encompass the CMC genres they were dealing with.

Briefly, an Interaction Space is located within a timeframe, during which interactions occur between a set of participants within an online location. This location is defined by the properties of the set of environments used by the participants who may be either individual members or groups. The environments may be synchronous or asynchronous, mono- or multimodal, simple or complex. The traces of actions within an environment and one particular modality of a CMC tool are termed 'acts'. Working with this concept, the TEI *CoMeRe* schema was proposed. The various components of the Interaction Space are defined in the <teiHeader> of the TEI file, while the actual use of the environments by the participants interacting is described in the <body> part of the TEI file.

IS relates to the intrinsic dialogic nature of such corpora and interactions and all CoMeRe corpora were structured this way. In most cases, the structure of the dialogues could be automatically detected. Only Wikipedia discussions (and particularly the *Wikiconflits* corpus described in Poudat et al., 2017) needed further checking – because of the particularities of Wiki editing and of the fact that wikipedians do not necessarily follow Wikipedia editing recommendations.

Another best practice we worked on concerns the encoding of information on participants: this information is of course crucial for the researcher. Here again, LETEC was the type of corpus in which we had the more detailed information about participants, including information on their role (teacher, learner, domain experts), their sex, age, linguistic competence (languages studied, mastered at different levels), the institutions they belong to, etc. Another part of the information relates to the characteristics and the composition of the groups which circumscribe the space of participants' interactions: the classroom, the subgroups belonging to one or different institutions, the roles played by the participants within each group (tutor, facilitator, learner, etc.). This detailed encoding about participants was also applied to the other CoMeRe genres which did not concern learning situations. For instance, in SMS corpora, questionnaires helped researchers to collect information about participants' habits and usage of SMS, the types of phones and the writing tools they use (Panckhurst et al., 2016). All information on participants have to be encoded in a standard way, and the schema we developed will also be used in the working groups of the new national consortium *CORLI* (*Corpus, Langues and Interactions*).

Lastly, and this will be further developed within CORLI, LETEC situations not only concern environments where participants interact simultaneously within different CMC textual tools, but also CMC oral tools, and tools based on non-verbal interactions (such as collaborative word processors, concept maps, whiteboards, even interactions generated through avatars which move in 3D worlds (Wigham and Chanier, 2013)). Thanks to the speech component of the TEI, data from CMC oral tools could be encoded with the <u> element. However, a new element, currently entitled <prod>, had to be created in order to encapsulate the transcription of non-verbal acts. In the IS model, all the three elements <post>, <u>, and <prod> appear at the same level in the hierarchy. This equality reflects the fact that participants can interact at the same time through textual, oral, or nonverbal acts, each of them associated to an author, a specific duration, and a content which may provoke another participant's reaction. Studying multimodal dialogues requires an encoding of the cross references of the different acts through their head characteristic or their contents.

All the *CoMeRe* corpora were encoded according to these principles, and were deposited into ORTOLANG[19]. This infrastructure represents the most important linguistic data service at the national level. It takes care of curation and long-term archiving. It plays a role similar to other CLARIN national structures, and should in the near future become a part of the European network.

Finally, we are currently working on best practices regarding the PoS tagging of CMC corpora. Only one corpus has been processed so far (a text chat corpus, see Riou and Sagot, 2016), thanks to the MElt tagger. The CoMeRe project has a special interest in further advancing that agenda in line with the European partnership.

4.3 The *DiDi* project (Italy)

4.3.1 Project description

The goal of the regionally funded 2-year DiDi project was to build a South Tyrolean CMC corpus and document the current language use. For this purpose, we collected language data from a social networking site (SNS) and combined it with socio-demographic data about the writers, obtained from a questionnaire (Frey et al., 2016). We chose to collect data from Facebook because this SNS is well known in South Tyrol, offers a wide variety of different communication methods, and is used throughout the territory by many social groups and people of different age.

The autonomous Italian province of South Tyrol is characterized by a multilingual environment with three official languages (Italian, German, and Ladin), and an institutional bi- or trilingualism (depending on the percentage of the Ladin population). Although the project focused on the German-speaking language group, all information regarding the project, for example, the invitation to participate, the privacy agreement, the project web site, and the questionnaire for collection socio-demographic data was published in German and Italian. Consequently, speakers of both Italian and German participated in the text collection campaign.

The multilingual CMC corpus combines Facebook status updates, comments, and private messages with socio-demographic data of the writers. The corpus was enriched with linguistic annotations on thread, text and token level, and provides the following socio-demographic information about the participants: gender, education, employment, internet communication habits, communication devices in use, internet experience, first language(s) (L1), and usage of a South Tyrolean German or Italian dialect and its particular origin. On text level, the corpus was semi-automatically annotated with language code(s) and a political vs. non-political topic label. On token level, the corpus was automatically annotated with part-of-speech, lemma, and CMC phenomenon (e.g. emoticons, emojis, and iteration of graphemes and punctuation) information, and manually normalised, anonymised and annotated with information about the use of German variety.

Another focus of the project was on the users' age and on the question whether a person's age influences language use on SNS; where age is understood in two ways: as a numerical value that reflects the life span of an individual and as digital age that reflects a person's experience with the new media.

Overall, the DiDi corpus comprises public and non-public language data of 136 South Tyrolean Facebook users. The users could choose to provide either their Facebook wall communication (status updates and comments), their chat (i.e. private messages) communication or both. In the end, 50 people provided access to both types of data. 80 users only provided access to their Facebook wall and 6 users gave their chat communication. In total, the corpus consists of around 600,000 tokens that are distributed over the text categories status updates (172,66 tokens), comments (94,512 tokens) and chat messages (328,796 tokens). German language content comprises 58% of the corpus. 13% are written in Italian and 4% in English (the remainder of the messages was either classified as unidentifiable language, non-language or other language). The distribution of the languages is in line with the language backgrounds of the participants and is comparable to the multilingual community of South Tyrol.

[19] https://www.ortolang.fr/

4.3.2 Results beyond the resource: A strategy for collecting private, non-public CMC data

Although the creation and analysis of CMC corpora is currently an active research area, projects exploring private conversations have been rare (but see Dürscheid and Stark (2011) and other sms4science[20] projects, Verheijen and Stoop (2016), and also, for example, the "What's up Switzerland?" project[21]). Instead, projects often explore publicly available data from SNS (like Facebook or Twitter), or data from Wikipedia or discussion boards, where data are relatively easy to obtain. Compared to publicly available data, the acquisition of private data is considerably more difficult in terms of privacy issues, technical implementation and sampled data retrieval. Obtaining private CMC data is time-consuming for both the researchers and the participants because direct interaction between the two is needed. Additionally, the data acquisition process might involve various media discontinuities; this, in turn, causes problems in terms of consistency during data transfer and increases the risk of possible data loss.

Bolander and Locher (2014) and Beißwenger and Storrer (2008) discuss general issues and challenges for corpora of publicly available CMC data. When dealing with non-public data, the issues of data acquisition for CMC corpora become even more demanding: *legal concerns* add to *ethical issues*, and *technical demands* related to *authentic* data retrieval and the linking of *mixed resources* (for example, linking language data and socio-linguistic meta information) get more challenging. Also, for technical and legal reasons of data acquisition an interaction between the user and the researcher becomes an inevitable necessity.

The *legal* situation of using publicly available user-generated language data for research is still under debate, but the trend leans towards seeking explicit user consent. Also, the data will be bound to copyright restrictions, making every modification, (re)publication or citation, potentially problematic (Baron et al., 2012). Furthermore, ethical considerations demand that researchers acquiring private personal data should seek the user's consent in advance and that the data is anonymised (Beißwenger and Storrer, 2008). For non-public data, this legal and ethical issues are even more critical. But also *technical constraints* make it necessary to interact with the users: most media platforms offer interfaces for third parties to explicitly request permission from the users to use their data. Finding a *representative sample of participants* for the corpus is another problem that, in fact, many corpus creation projects face. Often expensive public relation campaigns and incentives are necessary to get users to participate in projects where the requested data is personal, private and potentially intimate. Different approaches exist to gather the otherwise non-accessible private data, most of them asking for individual submissions of language data by the users.

Frey et al. (2014) considers 'submission by the user' to be too tedious for users and researchers, and also troublesome because of privacy concerns on the user side and authenticity doubts on the research side (the users might feel that their writing does not reflect "proper" language use, and brush it up before donating it). Instead, they suggest automatic data collection via a web appilcation: In this way, it is possible to gain user consent and socio-linguistic metadata with the highest privacy for participants (without personal interaction, no backtracking via mail addresses, etc.) and also to collect authentic language data. Additionally, it makes participating more attractive by simplifying the procedure of sharing language and metadata in an integrated, easy and time-saving way, that is also genuine in that media setting (i.e. the participation stays within the same platform, using the platform's interface and methods that are already familiar to the user). The data collection process consists of the following steps: (1) inform potential participants about the research project, the privacy policy and the data usage declaration; (2) provide options for the user to choose which content to share (private inbox and/or personal wall) and thereby increase the transparency for the user; (3) authenticate the user via the Facebook login dialogue (by using the Facebook API); (4) obtain the consent to use, save and republish the user's data (via the web application as well as via the Facebook infrastructure for privacy policies); (5) manage the registered user and the granted permissions via the Facebook login dialogue and the Facebook API; (6) request an anonymous and individual user identifier for the survey client, save permission flags, and enlist the user into an internal database; (7) redirect the user to the survey

[20] http://www.sms4science.org

[21] http://www.whatsup-switzerland.ch/

for the acquisition of the user's meta information; (8) provide dynamic feedback to the user about the current progress of the project (for example, about the amount of participants); (9) provide the possibility to share the application with Facebook friends to attract more users.

The web application and these steps keep the participation process as slim and simple as possible, and it takes users two clicks to donate their language data. There is no one-to-one interaction between an authenticated person and a researcher. Furthermore, legal and ethical constraints are met without additional effort: meta information of the questionnaire and actual language data are automatically linked with an individualised anonymous user identifier, provided by Facebook for every registered user of the web application; so, these identifiers can be used with third-party survey services without privacy problems. Moreover, the procedure facilitates the isolation of user acquisition and interaction with the actual crawling of language data. After logging in, the application grants access to the user's account for a period of 60 days, and the web application only manages registered users. Thus, using such a web application enables efficient data crawling: users do not have to wait for the language data download to complete, and the risk of data loss and other loading and saving issues decreases, as data can be retrieved in independent processes whenever capacities allow it best. Furthermore, server or system failures do not result in data loss since the data can be requested repeatedly. And finally, there are various possibilities to support the attractiveness of the research project: Dynamic feedback can be given via the application interface allowing participants to be part of a collective community project. The application can be easily shared as Facebook post, blog comment, twitter status, e-mail or any other media content, and after having finished the survey, participants can directly share the application with their friends. This workflow is genuine to social media contexts and addresses interested users wherever they happen to be. In addition, participants can be reached by Facebook via targeted advertising campaigns that address a specific user subset and are usually paid by conversions or actual reach of the advertisement.

For more details about the procedure and a discussion of problems and weaknesses see Frey et al. (2014). The anonymized corpus without the private messages is freely available for researchers, and the complete anonymized corpus is available after signing an agreement.[22]

4.4 The Janes project (Slovenia)

4.4.1 Project description

The *Janes* project[23] is compiling a corpus of Slovene user-generated content (Fišer et al., 2016) that contains five different text types of public user-generated content of varying lengths and communicative purposes: tweets, forum posts, user comments on on-line news portals (and, for completeness as well as for enabling comparative analyses, also the news articles themselves, even though they are not user-generated and will therefore not be further discussed in this paper), talk and user pages from Wikipedia, and blog posts along with user comments on these blogs. The collection of tweets and Wikipedia talk pages is comprehensive in the sense that the corpus includes all the Slovene users and their posts that we could identify at the time of harvesting. For the other text types, due to time and financial constraints, we selected only a small set of the most popular sources that at the same time offer the most textual content.

The most recent version of the corpus is v0.4 and it contains around 9 million texts comprising roughly 200 million tokens, 107 of which come from tweets, 47 from forum posts, 34 from blogs and their comments, 15 from news comments and 5 from Wikipedia. The texts in the corpus are structured according to the text types they belong to (e.g. conversation threads in forums) and contain rich metadata, which have been harvested directly during crawling and further enriched within the Janes project. The directly harvested metadata include date and time of posting, username, URL of the text, the discussion thread a text belongs to, the number of likes and retweets, etc. The enriched metadata, which have been added at either user- or text-level, are of two types: those that were added manually (account type, author's gender) and those that were added automatically (user's region, text sentiment, text standardness) (Čibej and Ljubešić, 2015, Fišer et al., 2016, Ljubešić et al., 2015). Figures 3 and 4

[22] http://www.eurac.edu/didi
[23] http://nl.ijs.si/janes/english/

show the distribution of sentiments and of levels of standardness, respectively, by account type and gender.

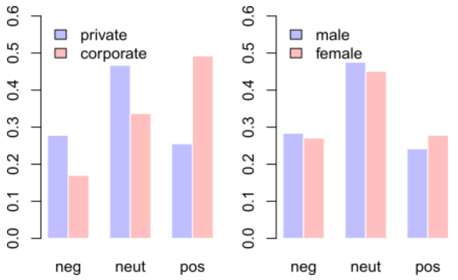


Figure 3: Sentiment of tweets by account type (left) and gender (right).

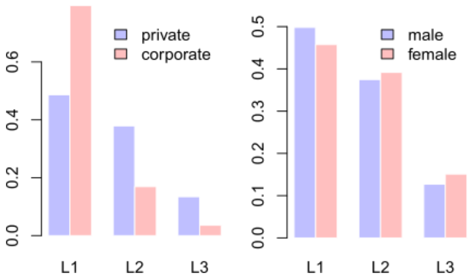


Figure 4: Standardness of tweets by account type (left) and gender (right) (L1 - completely standard, L2 - slightly non-standard, L3 - very non-standard).

For linguistic research on, as well as processing of non-standard language, the most relevant part of the metadata is the assignment of standardness scores to each text. We developed a method (Ljubešić et al., 2015) to automatically classify each text into three levels of technical and linguistic standardness. Technical standardness (T1, quite standard – T3, very non-standard) takes into account the use of spaces, punctuation, capitalisation and similar features, while linguistic standardness (L1, quite standard – L3, very non-standard) takes into account the level of adherence to the written norm and more or less conscious decisions to use non-standard language, involving spelling, lexis, morphology, and word order. On the basis of a manually labelled test set, the method has a mean error rate of 0.45 for technical and 0.54 for linguistic standardness prediction.

As further described in Section 4.5.1, the standard linguistic annotation workflow has been adapted to better tackle CMC-specific features and comprises five steps: tokenization, sentence segmentation, rediacritisation, normalization, morphosyntactic tagging, and lemmatization (Ljubešić and Erjavec, 2016, Ljubešić et al., 2016a, Ljubešić et al., 2016b) and the Janes corpus v0.4 is annotated with these levels of linguistic description.

```
<ab xml:id="janes.blog.publishwall.4264.3" type="blog" subtype="T1L3">
   <s>
      <w lemma="kaj" ana="#Rgp">Kaj</w><c> </c>
      <w lemma="biti" ana="#Va-r3s-y">ni</w><c> </c>
      <w lemma="ta" ana="#Pd-nsn">to</w><c> </c>
      <choice>
         <orig><w>tazadnje</w></orig>
         <reg>
            <w lemma="ta" ana="#Q">ta</w><c> </c>
            <w lemma="zadnji" ana="#Agpnsn">zadnje</w>
         </reg>
      </choice><c> </c>
      <choice>
         <orig><w>AAjevska</w></orig>
         <reg><w lemma="aa-jevski" ana="#Agpfsn">AA-jevska</w></reg>
      </choice><c> </c>
         <w lemma="molitev" ana="#Ncfsn">molitev</w>
         <pc ana="#Z">?</pc>
   </s>
</ab>
```

Figure 5: TEI encoding of a text in the JANES corpus

The corpus is encoded according to a bespoke XML schema that compactly reflects the structure of the corpus and its metadata. Version 1 will be encoded in a CMC-aware TEI (Beißwenger et al., 2012), cf. Figure 5. Apart from the XML source files, the corpus is also made available to linguists on the local installation of the noSketchEngine and SketchEngine concordancers (Kilgarriff et al., 2014), both as the entire Janes v0.4 corpus with the metadata that all the subcorpora have in common and as separate subcorpora with all the metadata available for the given subcorpus. Access to the corpus is currently restricted to project members, but steps are being taken to comply with the copyright, terms of use and privacy issues in order to make an anonymised, sampled and shuffled corpus available to other researchers as well by the end of the project (Erjavec et al., 2016a).

4.4.2 Results beyond the resource: Adaptation of NLP tools for processing Slovenian CMC language

In this section, we present the toolchain for automatic linguistic annotation of CMC we have mostly developed within the Janes project, as well as the datasets to enable its further improvements. Since most of the developed tools rely on supervised machine learning, we briefly report on the training data used and, where available, the estimated accuracy of each tool.

Tokenisation and sentence segmentation. For tokenisation and sentence segmentation, we used a new Python tool that covers Slovene, Croatian and Serbian (Ljubešić and Erjavec, 2016). Like most tokenisers, it is based on manually defined rules in the form of regular expressions and uses language-specific lexicons with, e.g. lists of abbreviations. In addition to standard rules, the tokeniser has an additional non-standard mode in which it uses less strict rules. For example, a full stop can here end a sentence even though the following word does not begin with a capital letter or is even not separated from the full stop by a space. Nevertheless, tokens that end with a full stop and are on the list of abbreviations (e.g. *prof.*) will not end a sentence. The non-standard tokeniser mode also has several additional rules, such as additional regular expressions devoted to recognising emoticons, e.g. *:-], :-PPPP,* ^_^ etc. A preliminary evaluation of the tool on tweets showed that sentence segmentation could still be significantly improved (86.3% accuracy), while tokenisation is relatively good (99.2%), taking into account that both tasks are very difficult for non-standard language.

Normalisation. Normalising non-standard word tokens to their standard form has two advantages. First, it becomes possible to search for a word without having to consider or be aware of all its spelling variants and second, normalisation makes it possible to use downstream tools for standard language processing, such as part-of-speech taggers. In the Janes corpus, the word tokens have been normalised when necessary by using a sequence of two steps. First, we use a dedicated tool (Ljubešić et al., 2016a) to restore diacritics (e.g. *krizisce → križišče*). The tool learns the rediacritisation model on a large collection of texts with diacritics paired with the same texts with the diacritics removed. The evaluation showed that the tool achieves a token accuracy of 99.62% on standard texts (Wikipedia) and 99.12% on partially non-standard texts (tweets). Second, the rediacriticised word tokens are normalised with a method that is based on character-level statistical machine translation (Ljubešić et al., 2016b). The goal of the normalisation is to translate words written in a non-standard form (e.g. *jest, jst, jas, js*) to their standard equivalent (*jaz*). The current translation model for Slovene was trained on a preliminary version of the manually normalised dataset Janes-Norm (cf. below), while the target (i.e. standard) language model was trained on the Kres balanced corpus of Slovene (Logar Berginc et al., 2012) and the tweets from the Janes corpus that were labelled as linguistically standard. It should be noted that normalisation will sometimes also span word-boundaries, i.e. there are cases where one non-standard word corresponds to two or more standard words or vice versa (e.g. *ne malo → nemalo; tamau → ta mali*).

Tagging and lemmatization. As the final step in the text annotation pipeline, the normalised tokens were annotated with their morphosyntactic description (MSD) and lemma. For this, we used a newly developed CRF-based tagger-lemmatiser that was trained for Slovene, Croatian and Serbian (Ljubešić and Erjavec, 2016). The main innovation of the tool is that it does not use its lexicon directly, as a constraint on possible MSDs of a word, but rather indirectly, as a source of features; it thus makes no distinction between known and unknown words. For Slovene, the tool was trained on the ssj500k 1.3 corpus (Krek et al., 2013) and the Sloleks 1.2 lexicon (Dobrovoljc et al., 2015). Compared to the previous best result for Slovene with the Obeliks tagger (Grčar et al., 2012), the CRF tagger reduces the relative error by almost 25% achieving a 94.3% accuracy on the test set comprising the last tenth of the ssj500k corpus. The MSD tagset used within the Janes project follows the MULTEXT Version 4 specifications (Erjavec, 2012), except that we, following Bartz et al. (2014), introduce new MSDs for the annotation of CMC-specific content, in particular Xw (e-mails, URLs), Xe (emoticons and emoji), Xh (hashtags, e.g. *#kvadogaja*) and Xa (mentions, e.g. *@dfiser3*). The lemmatisation, which is also part of the tool, takes into account the posited MSD. For pairs word-form:MSD which are already in the training lexicon, it simply retrieves the lemma, while for the rest it uses its lemmatisation model to guess the lemma.

Manually annotated datasets. To further improve our annotation tool chain, we have manually annotated two gold-standard datasets (Erjavec et al., 2016b): Janes-Norm (Erjavec et al., 2016c), which contains 7,816 texts or 184,755 tokens, is a gold-standard dataset for tokenisation, sentence segmentation and word normalisation, while Janes-Tag, (Erjavec et al., 2016d), a subset of Janes-Norm, comprises 2,958 texts or 75,276 tokens, and is a gold-standard dataset for training and evaluating morphosyntactic tagging and lemmatisation.

The annotation guidelines which were produced to guide the annotation of these two corpora to a large extent follow the guidelines for annotating standard (Holozan et al., 2008) and historical (Erjavec, 2015) Slovene, with some medium-specific modifications (e.g. the annotation of emoticons, URLs, hashtags, and mentions). At the normalisation level, special attention was paid to non-standard words with multiple spelling variants and those without a standard form (e.g. *orng, ornk, oreng, orenk* for *'very'*), foreign language elements (e.g. *updateati, updajtati, updejtati, apdejtati* for *'to update'*) and linguistic features that are not normalised (e.g. hashtags, non-standard syntax and stylistic issues). At the morphosyntactic description (MSD) and lemmatisation levels, the guidelines were designed to deal with foreign language elements, proper names and abbreviations as well as non-standard use of case and particles. All the texts were first automatically annotated, then checked and corrected manually by a team of students, with two students annotating each text and the divergent annotation checked by an experience curator. The platform used for manual annotation was WebAnno (Yimam et al., 2013).

Janes-Norm and Janes-Tag are deposited on the CLARIN.SI repository and freely available for research under the CC BY licence.

5 Outlook

In this paper, we gave an overview of results and best practices from projects in four countries dedicated to the creation of corpora of computer-mediated communication and social media interactions (CMC). The joint goal of the projects is to establish standards for the collection and representation of CMC corpora and for their integration into common resources infrastructures.

Up to now, the network has brought forward two main initiatives: a conference series dedicated to all issues related to building, annotating and analyzing CMC corpora, and a TEI-SIG focused on the integration of standards for CMC resources into the TEI framework. Both initiatives are "bottom up" with the goal to connect researchers all over Europe and to work on solutions driven by practices that have proven useful in ongoing projects. The latest edition of the conference included 22 contributions by 40 authors from 24 research institutions in 11 countries (Fišer and Beißwenger, 2016).

Nevertheless, there's still a lot of open, non-trivial issues in the field. One example is the lack of legal standards for collecting and republishing CMC data as part of language resources. Corpus builders are typically laymen when it comes to legal issues. A general legal opinion on these issues commissioned and disseminated by and via an acknowledged language resources initiative (e.g., CLARIN or its national consortia) would therefore be an important prerequisite for the further development of the CMC corpora landscape and community.

In view of the importance of CMC in everyday communication, in business, public administration, science and education, efforts in the field of establishing state-of-the-art research and resource infrastructures for the analysis of CMC phenomena are an investment in our future knowledge about how the adoption of CMC technologies affects society and how communicative practices reflect the presence of CMC as an innovative means for the organization of social interaction.

References

[Baron et al.2012] Alistair Baron, Paul Rayson, Phil Greenwood, James Walkerdine, and Awais Rashid. 2012. Children Online: A Survey of Child Language and CMC Corpora. *International Journal of Corpus Linguistics,* 17(4):443–81.

[Bartz et al.2014] Thomas Bartz, Michael Beißwenger, and Angelika Storrer. 2014. Optimierung des Stuttgart-Tübingen-Tagset für die linguistische Annotation von Korpora zur internetbasierten Kommunikation: Phänomene, Herausforderungen, Erweiterungsvorschläge. *Journal for Language Technology and Computational Linguistics,* 28(1):157–198.

[Beißwenger and Storrer2008] Michael Beißwenger and Angelika Storrer. 2008. Corpora of computer-mediated communication. In: Lüdeling, Anke; Kytö, Merja (eds.). *Corpus Linguistics HSK*, vol. 29.1. Walter de Gruyter, Berlin, Germany, pp. 292–309.

[Beißwenger et al.2012] Michael Beißwenger, Maria Ermakova, Alexander Geyken, Lothar Lemnitzer, and Angelika Storrer. 2012. A TEI Schema for the Representation of Computer-mediated Communication. *Journal of the Text Encoding Initiative (Online),* (3) (doi: 10.4000/jtei.476). http://jtei.revues.org/476.

[Beißwenger2013] Michael Beißwenger. 2013. Das Dortmunder Chat-Korpus. *Zeitschrift für germanistische Linguistik,* 41(1):161–164.

[Beißwenger et al.2015] Michael Beißwenger, Thomas Bartz, Angelika Storrer, and Swantje Westpfahl. 2015. *Tagset and Guidelines for the PoS Tagging of Language Data from Genres of Computer-mediated Communication / Social Media*. http://sites.google.com/site/empirist2015/home/annotation-guidelines.

[Beißwenger et al.2016] Michael Beißwenger, Sabine Bartsch, Stefan Evert, and Kay-Michael Würzner. 2016. EmpiriST 2015: A shared task on the automatic linguistic annotation of computer-mediated communication and web corpora. In*: Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task.* Berlin, Germany, pp. 44–56. http://aclweb.org/anthology/W/W16/W16-2606.pdf

[Bolander and Locher2014] Brook Bolander and Miriam A. Locher. 2014. Doing Sociolinguistic Research on Computer-Mediated Data: A Review of Four Methodological Issues. *Discourse, Context & Media,* (3):14–26.

[Chanier et al.2014] Thierry Chanier, Celine Poudat, Benoit Sagot, Georges Antoniadis, Ciara Wigham, Linda Hriba, Julien Longhi, and Djamé Seddah. 2014. The CoMeRe corpus for French: structuring and annotating heterogeneous CMC genres. *Journal of language Technology and Computational Linguistics,* 29(2):1–30. http://www.jlcl.org/2014_Heft2/1Chanier-et-al.pdf.

[Chanier and Wigham2016] Thierry Chanier and Ciara Wigham. 2016. Standardizing Multimodal Teaching and Learning Corpora. In: Marie-Jo, Hamel; Caws, Catherine (eds.). *Language-Learner Computer Interactions: Theory, Methodology and CALL Applications.* John Benjamins, Amsterdam, Netherlands, pp. 215-240. DOI:10.1075/lsse.2.10cha.

[Chiari and Canzonetti2014] Isabella Chiari and Alessio Canzonetti. 2014. Le forme della comunicazione mediata dal computer: generi, tipi e standard di annotazione. In: Garavelli, Enrico; Suomela-Härmä, Elina (eds.). *Dal manoscritto al web: canali e modalità di trasmissione dell'italiano. Tecniche, materiali e usi nella storia della lingua.* Atti del XII Convegno della Società Internazionale di Linguistica e Filologia Italiana (SILFI), Helsinki, 18-19 June 2012. Franco Cesati Editore, Firenze, Italy, pp. 595-606.

[Čibej and Ljubešić2015] Jaka Čibej and Nikola Ljubešić. 2015. *"S kje pa si?" – Metapodatki o regionalni pripadnosti uporabnikov družbenega omrežja Twitter.* Zbornik konference Slovenščina na spletu in v novih medijih, Ljubljana, Slovenia, pp. 10-14.

[CLARIN-D schema2015] CLARIN-D TEI schema for CMC corpora. 2015. http://wiki.tei-c.org/index.php?title=SIG:CMC/clarindschema.

[CoMeRe repository2016] CoMeRe repository. 2016. *Corpora of Computer-Mediated Communication in French.* Ortolang.fr, Nancy, France. http://hdl.handle.net/11403/comere.

[CoMeRe schema2014] CoMeRe TEI schema for CMC corpora, version 2. 2014. https://repository.ortolang.fr/api/content/comere/v2/tei_cmr.rng and http://wiki.tei-c.org/index.php/SIG:CMC/CoMeRe_schema_draft_for_representing_CMC_in_TEI_(2014).

[Dobrovoljc et al.2015] Kaja Dobrovoljc, Simon Krek, Peter Holozan, Tomaž Erjavec, and Miro Romih. 2015. *Morphological Lexicon Sloleks 1.2.*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1039.

[Dürscheid and Stark2011] Christa Dürscheid and Elisabeth Stark. 2011. sms4science: An international corpus-based texting project and the specific challenges for multilingual Switzerland. In: Thurlow, Crispin; Mroczek, Kristine (eds.): *Digital Discourse. Language in the New Media*. Oxford University Press, Oxford, UK, pp. 299-320.

[Erjavec2012] Tomaž Erjavec. 2012. MULTEXT-East: morphosyntactic resources for Central and Eastern European languages. *Language Resources and Evaluation*, 46(1):131–142.

[Erjavec2015] Tomaž Erjavec. 2015. The IMP historical Slovene language resources. *Language Resources and Evaluation*, 49(3):753–775.

[Erjavec et al.2016a] Tomaž Erjavec, Jaka Čibej, and Darja Fišer. 2016. Omogočanje dostopa do korpusov slovenskih spletnih besedil v luči pravnih omejitev. *Slovenščina 2.0*, 4(2):189–219.

[Erjavec et al.2016b] Tomaž Erjavec, Jaka Čibej, Špela Arhar Holdt, Nikola Ljubešić, and Darja Fišer. 2016. Gold-Standard Datasets for Annotation of Slovene Computer-Mediated Communication. In*: Proceedings of the Tenth Workshop on Recent Advances in Slavonic Natural Languages Processings*, Brno, the Czech Republic, pp. 29–40.

[Erjavec et al.2016c] Tomaž Erjavec, Darja Fišer, Jaka Čibej, Špela Arhar Holdt, and Nikola Ljubešić. 2016. *CMC Training Corpus Janes-Norm 1.2*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1084.

[Erjavec et al.2016d] Tomaž Erjavec, Darja Fišer, Jaka Čibej, Špela Arhar Holdt, and Nikola Ljubešić. 2016. *CMC Training Corpus Janes-Tag 1.2*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1085.

[Fišer and Beißwenger2016] Darja Fišer and Michael Beißwenger (eds.). 2016. *Proceedings of the 4th Conference on CMC and Social Media Corpora for the Humanities (cmc-corpora2016).* University of Ljubljana, Slovenia. http://nl.ijs.si/janes/cmc-corpora2016/proceedings/

[Fišer et al.2016] Darja Fišer, Tomaž Erjavec, and Nikola Ljubešić. 2016. JANES v0.4: Korpus slovenskih spletnih uporabniških vsebin. *Slovenščina 2.0*, 4(2):67–99.

[Forsyth an Martell2007] Eric N. Forsyth and Craig H. Martell. 2007. Lexical and Discourse Analysis of Online Chat Dialog. In: *Proceedings of the First IEEE International Conference on Semantic Computing* (ICSC 2007), Irvine, USA, pp. 19-26.

[Frey et al.2014] Jennifer-Carmen Frey, Egon W. Stemle, and Aivars Glaznieks. 2014. Collecting Language Data of Non-Public Social Media Profiles. In: *Workshop Proceedings of the 12th Edition of the KONVENS Conference*, edited by Gertrud Faaß and Josef Ruppenhofer. Universitätsverlag Hildesheim, Hildesheim, Germany, pp. 11-15.

[Frey et. al.2016] Jennifer-Carmen Frey, Aivars Glaznieks, and Egon W. Stemle. 2016. The DiDi Corpus of South Tyrolean CMC Data: A Multilingual Corpus of Facebook Texts. Accepted at CLIC-it 2016.

[Grčar et al.2012] Miha Grčar, Simon Krek, and Kaja Dobrovoljc. 2012. *Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik (Obeliks: a statistical morphosyntactic tagger and lemmatiser for Slovene).* Zbornik Osme konference Jezikovne tehnologije, Ljubljana, Slovenia.

[Holozan et al.2008] Peter Holozan, Simon Krek, Matej Pivec, Simon Rigač, Simon Rozman, and Aleš Velušček. 2008. *Specifikacije za učni korpus. Projekt "Sporazumevanje v slovenskem jeziku" (Specifications for the Training Corpus. The "Communication in Slovene" project).* http://www.slovenscina.eu/Vsebine/Sl/Kazalniki/K2.aspx.

[Horbach et al.2014] Andrea Horbach, Diana Steffen, Steffen Thater, and Manfred Pinkal. 2014. Improving the Performance of Standard Part-of-Speech Taggers for Computer-Mediated Communication. In: *Proceedings of KONVENS 2014*, pp. 171–177. https://hildok.bsz-bw.de/frontdoor/index/index/docId/241.

[iRights.Law2016] iRights.Law Rechtsanwälte. 2016. *Rechtsgutachten zur Integration mehrerer Text-Korpora in die CLARIN-D-Infrastrukturen.* (Legal opinion for the ChatCorpus2CLARIN project, 46 pages).

[Kilgarriff et al.2014] Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The Sketch Engine: ten years on. *Lexicography*, 1(1):7–36.

[Krek et al.2013] Simon Krek, Tomaž Erjavec, Kaja Dobrovoljc, Sara Može, Nina Ledinek, and Nanika Holz. 2013. *Training Corpus ssj500k 1.3*. Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1029.

[Ljubešić and Erjavec2016] Nikola Ljubešić and Tomaž Erjavec. 2016. Corpus vs. lexicon supervision in morphosyntactic tagging: the case of Slovene. In: *Proceedings of the 10th Language Resources and Evaluation Conference*, Portorož, Slovenia, pp. 1527–1531.

[Ljubešić et al.2016a] Nikola Ljubešić, Tomaž Erjavec, and Darja Fišer. 2016. Corpus-based diacritic restoration for South Slavic languages. In: *Proceedings of the 10th Language Resources and Evaluation Conference.* Portorož, Slovenia, pp. 3612–3616.

[Ljubešić et al.2016b] Nikola Ljubešić, Katja Zupan, Darja Fišer, and Tomaž Erjavec. 2016. Normalising Slovene data: historical texts vs. user-generated content. In: *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, Bochum, Germany, pp. 146–155.

[Ljubešić et al.2015] Nikola Ljubešić, Darja Fišer, Tomaž Erjavec, Jaka Čibej, Dafne Marko, Senja Pollak, and Iza Škrjanec. 2015. Predicting the level of text standardness in user-generated content. In*: Proceedings of the 10th International Conference on Recent Advances in Natural Language Processing*, Hissar, Bulgaria, pp. 371–378.

[Logar Berginc et al.2012] Nataša Logar Berginc, Miha Grčar, Marko Brakus, Tomaž Erjavec, Špela Arhar Holdt, and Simon Krek. 2012. *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba* (The Gigafida, KRES, ccGigafida and ccKRES corpora of Slovene language: compilation, content, use.) Ljubljana, Slovenia: Trojina, zavod za uporabno slovenistiko, Faculty of Social Sciences.

[Lüngen et al.2016] Harald Lüngen, Michael Beißwenger, Eric Ehrhardt, Axel Herold, and Angelika Storrer. 2016. Integrating corpora of computer-mediated communication in CLARIN-D: Results from the curation project ChatCorpus2CLARIN. In: *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016),* Bochum, Germany, pp. 156–164. https://www.linguistics.rub.de/konvens16/pub/20_konvensproc.pdf.

[Margaretha and Lüngen2014] Eliza Margaretha and Harald Lüngen. 2014. Building Linguistic Corpora from Wikipedia Articles and Discussions. *Journal of language Technology and Computational Linguistics,* 29(2):59–82. http://www.jlcl.org/2014_Heft2/3MargarethaLuengen.pdf.

[Oostdijk et al.2013] Nelleke Oostdijk, Martin Reynaert, Véronique Hoste, and Ineke Schuurman. 2013. The Construction of a 500 Million Word Reference Corpus of Contemporary Written Dutch. In: Spyns, Peter; Odijk, Jan (eds*). Essential Speech and Language Technology for Dutch: Results by the STEVIN-programme*, Springer Verlag, Berlin, Germany, pp. 219-247.

[Panckhurst et al.2016] Rachel Panckhurst, Catherine Détrie, Cédric Lopez, Claudine Moïse, Mathieu Roche, and Bertrand Verin. 2016. *88milSMS*: A corpus of authentic text messages in French. [corpus] In: Chanier, Thierry (ed*). Banque de corpus CoMeRe*. Ortolang, Nancy, France. https://hdl.handle.net/11403/comere/cmr-88milsms.

[Poudat et al.2017] Céline Poudat, Natalia Grabar, Camille Paloque-Berges, Thierry Chanier, and Kun Jin. 2017. Wikiconflits: un corpus de discussions éditoriales conflictuelles du Wikipédia francophone. In: Wigham, C.R.; Ledegen, G. (eds.). 2017. *Corpus de communication médiée par les réseaux: Construction, structuration, analyse.* Collection Humanités Numériques. L'Harmattan, Paris, France, pp. 211-222.

[Riou and Sagot2016] Stéphane Riou and Benoit Sagot. 2016. *Etiquetage morpho-syntaxique du corpus FAVI* [corpus]. D'après Yun, H. & Chanier, T. (2014). Corpus d'apprentissage FAVI (Français académique virtuel international) [cmr-favi-tei-v1]. Banque de corpus CoMeRe. Ortolang, Nancy, France. http://hdl.handle.net/11403/comere/cmr-favi/cmr-favi-tei-v2

[Schiller et al.1999] Anne Schiller, Simone Teufel, Christine Stöckert, and Christine Thielen. 1999. *Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset).* Institut für maschinelle Sprachverarbeitung, University of Stuttgart, Germany. http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf.

[Schröck and Lüngen2015] Jasmin Schröck and Harald Lüngen. 2015. Building and Annotating a Corpus of German-Language Newsgroups. In*: Proceedings of the 2nd Workshop on Natural Language Processing for Computer-Mediated Communication / Social Media (NLP4CMC2015).* Essen, Germany, pp. 17-22. https://sites.google.com/site/nlp4cmc2015/program

[TEI P5] TEI Consortium (eds) (2007): TEI P5: Guidelines for Electronic Text Encoding and Interchange. http://www.tei-c.org/Guidelines/P5/.

[Verheijen and Stoop2016] Lieke Verheijen and Wessel Stoop. 2016. Collecting Facebook Posts and WhatsApp Chats. In: *Proceedings. Text, Speech, and Dialogue: 19th International Conference*, TSD 2016, Brno, Czech Republic, September 12-16, 2016, Springer International Publishing, Cham, Germany, pp. 249–58.

[Westpfahl and Schmidt2016] Swantje Westpfahl and Thomas Schmidt. 2016. *FOLK-Gold – A GOLD standard for Part-of-Speech- Tagging of Spoken German*. In: *Proceedings of the Tenth conference on International Language Resources and Evaluation (LREC16),* Paris, France, pp. 1493-1499.

[Wigham and Chanier2013] Ciara Wigham and Thierry Chanier. 2013. Interactions Between Text Chat and Audio Modalities for L2 Communication and Feedback in the Synthetic World Second Life. *Computer Assisted Language Learning*, 28(3):260-283. DOI:10.1080/09588221.2013.851702.

[Yimam et al.2013] 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),* Association for Computational Linguistics, Stroudsburg, USA, pp. 1–6.

MTAS: A Solr/Lucene based Multi Tier Annotation Search solution

Matthijs Brouwer
Meertens Institute
The Netherlands
matthijs.brouwer@
meertens.knaw.nl

Hennie Brugman
Meertens Institute
The Netherlands
hennie.brugman@
meertens.knaw.nl

Marc Kemps-Snijders
Meertens Institute
The Netherlands
marc.kemps.snijders@
meertens.knaw.nl

Abstract

In recent years, multiple solutions have become available providing search on huge amounts of plain text and metadata. Scalable searchability on annotated text however still appears to be problematic. With Mtas, an acronym for Multi-Tier Annotation Search, we add annotation layers and structure to the existing Lucene approach of creating and searching indexes, and furthermore present an implementation as Solr plugin providing both searchability and scalability. We present a configurable indexation process, supporting multiple document formats, and providing extended search options on both metadata and annotated text, such as advanced statistics, faceting, grouping and keyword-in-context. Mtas is currently used in production environments, with up to 15 million documents and 9.5 billion words. Mtas is available from GitHub[1].

1 Introduction

Many solutions providing search on both plain text and metadata rely on the inverted index based Apache Lucene[2]. The existing and popular extension Solr offers additional features such as distributed indexes, scalability, and load balanced querying to the construction of a sustainable and scalable infrastructure for these types of search requirements. However, for *annotated* textual resources these solutions appear less suitable due to the additional complexity introduced by the various annotation layers and limited options available within Solr and Lucene. Also, results and derived statistics are mainly based on numbers of documents, while often individual hits are required. There seems to be an increasing demand for solutions to these problems.

Several approaches are already available, amongst others from the CLARIN community, e.g. BlackLab, KorAP, SketchEngine, Corpus Workbench, PaQu, GrETEL, Corpuscle to name a few. Various considerations have led us to develop a new initiative in this area, most notably scalability and integrated metadata/annotation search. We provide an overview of functional requirements from our infrastructure projects and scientific users, and elaborate on development decisions taken in relation to existing solutions. After a short introduction of the implemented CQL support, we discuss performance and capabilities of statistics, faceting, grouping and termvectors. We conclude with performance, consistency checks and some suggestions on future work.

2 Requirements

In this section we describe the main high-level requirements that determine the general scope and direction of Mtas development. While it is certainly true that, depending on the envisaged use case or in comparison with other systems, additional requirements could be formulated that equally deserve attention, a number of specific strategic, functional and operational requirements provides the main

[1] https://github.com/meertensinstituut/mtas

[2] https://lucene.apache.org/

Matthijs Brouwer, Hennie Brugman and Marc Kemps-Snijders 2017. MTAS: A Solr/Lucene based multi tier annotation search solution. *Selected papers from the CLARIN Annual Conference 2016*. Linköping Electronic Conference Proceedings 136: 19–37.

foundation for our development ambitions. From a strategic perspective, multi tier annotation search represents one of the key components that supports the data management life cycle in our domain. Annotated text is essential to unlock (textual) data contents beyond the metadata level. Therefore we consider it imperative that we build up internal knowledge and experience in this area. Moreover, in order to achieve sufficient control and room for experimentation we strive for close collaboration with system providers or, if close collaboration proves impossible, investment in independent system creation. Close collaboration with researchers ensures that we invest in those functional areas that are immediately of interest to the research community we work with. Finally, our operational requirements are related to ease of deployment, testing and maintenance of the system. For example, ease of integration of new collections and the ability to quickly deploy various instances of the system for testing or production purposes are very important.

Besides the Nederlab project[3] (Brouwer, et al., 2014) serving as one of the primary use cases and application platforms for the system design, the development of Mtas is firmly situated in the CLARIN domain as part of the Dutch CLARIAH project. One of the major requirements therefore is the ability to include arbitrary (CMDI) metadata schemas in search processes. Considerable experience is at hand in making metadata available in metadata search processes, e.g. in the CLARIN VLO. This search domain however needs to be extended to include annotated text, containing multiple, often interdependent, annotation layers. These layers consist of normalized and spell-corrected texts, translations, lemma, part of speech (including feature lists), named entities, entity links to external knowledge bases such as DBpedia, chapters, paragraphs, sentences and other hierarchical annotations such as morphology or syntactic information. Given the myriad of annotation formats encountered in the domain, the system should be configurable to cope with a fair amount of different annotation formats.

At the level of individual annotation layers, support must be provided for multivalued attributes, differentiation between multiple set values (e.g. to cater for multiple tag sets simultaneously occurring in source documents) and full Unicode support at the value level. The system should support CQL (Corpus Query Language), possibly extended with additional features for higher order structures, such as for example hierarchies. The choice for Corpus Query Language is motivated by the fact that this is commonly used by various systems in the community, albeit with local differences in interpretation. Also, with the advances of the Federated Content Search program in CLARIN it is anticipated that Corpus Query Language will be adopted to extend the current SRU/CQL (Search/Retrieval via URL and Contextual Query Language) capabilities allowing for a more easy alignment with FCS activities.

With respect to result delivery, both (annotated) documents and keyword-in-context representations must be delivered, as well as statistical information regarding absolute and relative frequencies and hit distributions across result sets. Result set distributions must be calculated across each available metadata dimension, including time intervals, and not only over single but also multiple metadata dimensions, e.g. a distribution across both time and genre. Result sets may also be grouped according to result characteristics, such as grouping of all adjectives preceding a noun, to assist in determining collocations. Also, the system should be able to produce frequency lists across any result set and type of annotation; word forms, part of speech, named entities, etc. Finally, the system should be highly scalable, be able to work across multi-billion word corpora, be easily manageable and be freely available for use to a wide user community under an open source license[4].

2.1 Current solutions

A choice between search engines is often a balancing act between one's requirements and depends upon one's functional scope, corpus size, available expertise or conditions of use. Several systems designed for searching annotated text structures are currently available, each with its own strengths, weaknesses and track record, e.g. BlackLab (Reynaert, et al., 2014), KorAP (Banski, et al., 2013), Corpus Workbench (Evert & Hardie, 2011), Sketch Engine (Kilgarriff, et al., 2004), PaQu (Odijk, 2015), GrETEL (Vandeghinste, et al., 2014) and Corpuscle (Meurer, 2012).

Although many of these provide partial coverage of the listed requirements, as can be seen from our findings listed in Table 1, none of them provides a balanced coverage to be immediately applicable to

[3] https://www.nederlab.nl/

[4] https://github.com/meertensinstituut/mtas

our projects at hand. It thus became clear that, if any of these existing solutions were to be used, they would need to be modified to suit our needs. It should also be noted that the presented list is not considered to be exhaustive, but indicates the functional scope of the project. Other, non-functional, factors in the system choice were a preference towards a widely adopted and supported open source framework with clear design principles an active community maintaining the framework. This allows us to benefit from new insights gained and new progress made by the wider community and minimizes the risk of getting stuck in a dead-end program. Should such a framework reach its end-of-life then most likely it will be possible to secure a graceful migration path towards other systems.

As a development approach, new features were to be added using an iterative approach with short development cycles. This helps to identify risks in early stages of development and prevents over-engineering. 'Gold plating' is to be avoided, focusing on only delivering those features that are relevant to the use cases and research questions at hand.

Looking at the requirements it becomes clear that the *open source* and *scalability* requirements narrow the choice to only a limited number of systems. The Corpus Workbench is considered to be nearing its end-of-life judging from the new developments at IMS. Initial tests with Neo4J indicated that, even with adjustments, for graph databases such as Neo4j, performance and scalability is expected to remain problematic: the more general graph structure prevented us to take full advantage of the sequential nature of annotated text with reasonable response times. The Corpuscle system, besides its small user community and our inability to locate the source code, is considered to a rather exotic implementation being written in Common Lisp.

The BlackLab solution, being based also on Lucene, may seem to have some resemblances with our approach, although in Mtas we choose to take a completely different approach to represent distinctive annotation layers and hierarchical structure in Lucene. However, in our initial attempts to extend the BlackLab functionality, starting with taking advantage of the advanced scalability, sharding and other options provided by Solr, such as faceting, it became clear that the underlying architecture of the system prevented us from doing so without significantly altering the underlying code base. Rather than modifying the complete code base we choose to re-implement the system in such a manner that it was interoperable with Solr from the start.

Solr/Lucene is fast, scales well, and has a large basis of users as well as developers. The latter stands in sharp contrast to several existing corpus search and management systems, for which one or few developers have the task of maintenance and further development if the system. With Solr/Lucene one gets speed and scalability almost for free which makes it an interesting option as an implementation basis. Also, we had already gained considerable experience using Solr/Lucene for metadata and plain text indexing, it ties in well with existing infrastructure components and it provides good options for scalability and large corpus maintenance through its sharding functionality. Sharding refers to the possibility to create horizontal partitions of the data. Horizontal partition is a term that originates from the database community and refers to splitting one or more tables by row. In Solr, shards have one or more replicas and each replica is a core. A core refers to a single index and associated transaction log and configuration files. In our use cases, individual collections or sub collections can be indexed into separate cores and, using the sharding features, be addressed separately or collectively.

	Corpus Workbench	Sketch Engine	PaQu	GrETEL	BlackLab	Corpuscle	Neo4J	Solr	Solr + Mtas
Open source	✓	✗	✓	?	✓	?	✓	✓	✓
Highly scalable	✓	✓	✗	✗	✓	?	✓	✓	✓
Distributed search	?	?	✗	✗	✗	?	✓	✓	✓
Arbitrary (CMDI) metadata schemas	✗	✗	✗	✗	✗	?	✓	✓	✓
Annotated text	✓	✓	✓	✓	✓	✓	✗	✗	✓
- full support annotations	✗	✗	✓	✓	✗	✗	✗	✗	✓
- hierarchical structure	✗	✗	✓	✓	✗	✗	✗	✗	✓
- configurable mapping of input format on index	✗	✗	✗	✗	✗	✗	✗	✗	✓
- support FoLiA annotation format	✗	✗	✓	✗	✓	✗	✗	✗	✓
- corpus query language (CQL)	✓	✓	✗	✗	✓	✓	✗	✗	✓
- full statistics	✗	?	✗	✗	✗	✗	✗	✗	✓
- term vectors over any result document set	✓	✓	✗	✗	✗	✗	✗	✗	✓
- grouping	✓	✓	✗	✗	✓	✓	✗	✗	✓
- faceting	✗	✗	✗	✗	✗	✗	✗	✗	✓
- keyword in context (kwic)	✓	✓	✓	✓	✓	✓	✗	✗	✓

Table 1: Overview findings native coverage of our specific requirements for several existing solutions directed towards searching annotated text and/or structures.

3 Extending Lucene and Solr

Basic Lucene search functionality is based on the idea that data is grouped into documents. Each document consists of several fields and each field can have multiple values. Using this approach for metadata purposes, several values of *genre*, e.g. *fictie* and *proza*, can be associated with each document. For fields containing textual data, position information is available for each value. To this existing Lucene approach, we add annotations and structure by using prefixes to distinguish between text and different annotations. Annotations and structure are stored together with text in a separate designated type of Lucene field, thus providing simultaneous access to traditional Lucene fields for storing metadata features and Mtas enabled content. This provides a direct solution to store and search for annotations on individual words within a text, and only an adjusted tokenizer is needed to offer the correct token stream to the indexer. Ranges of words, distinct sets of words (e.g. named entities) and hierarchical relations are stored as a *payload* A payload, in Lucene terms, refers to an arbitrary array of bytes associated with a Lucene token at a certain position. Several additional extensions are used implementing different query strategies, most of them extend default Lucene methods. Our extension assumes a basic tokenization enriched with annotations on both single and multi-token levels. In most cases word level tokenization is used. It also possible to define other tokenizations, for example at the morpheme level, and in this case words will usually span multiple tokens. Table 2 provides a representation of various layers expressed in our index. Single and multi-token elements can be distinguished in the *position* column and parent hierarchy may be derived from the *parent* column. Prefixes displayed in this table are described through the configurable mapping (see 3.1).

Id	Offset		Position		Parent	Payload	Prefix	Postfix
72	1443	8100	0	5			s	s
0	1515	1536	0		72		t	Amsterdam
1	1515	1536	0		72		t_lc	amsterdam
2	1605	1616	0		3	1.0	feat.spectype	deeleigen
3	1540	1668	0		72	1.0	pos	SPEC
4	1674	1683	0		72		lemma	Amsterdam
53	5458	5617	0		72	0.885417	chunk	NP
56	6122	6324	0		72		entity	loc
5	1808	1822	1		72		t	is
6	1808	1822	1		72		t_lc	is
7	1894	1905	1		10	0.999891	feat.wvorm	pv
8	1938	1949	1		10	0.999891	feat.pvtijd	tgw
9	1984	1995	1		10	0.999891	feat.pvagr	ev
10	1826	2037	1		72	0.999891	pos	WW
11	2043	2052	1		72		lemma	zijn
54	5625	5777	1		72	0.993895	chunk	VP
…								
48	5019	5105	0		49		dependency.dep	
49	4880	5120	0	1	72		dependency	su
47	4936	5014	1		49		dependency.hd	
…								
66	7550	7628	1		68		dependency.hd	
67	7633	7714	4		68		dependency.dep	
68	7410	7729	1, 4		72		dependency	predc
…								

Table 2 Sample representation of Mtas index showing offsets, positions and postfix information for various prefixes.

Lucene uses an inverted index, storing the mapping from content, such as a word, to its location in a document for quickly retrieving search results and location in a text. While the inverted index plays an important role in most search operations, especially for dealing with multiple tiers in annotations it is also necessary to use forward indexes. These play an important role in result delivery processes such as keyword-in-context, lists and grouping functionality. We currently provide three main types of forward indexes for each available document, based on position, parent id and object id. These indexes are created and updated automatically when documents are added or deleted, or when cores are merged or optimized[5].

From a maintenance perspective, this approach provides the possibility to index collections separately into separate cores and simply activate new cores using Solr. Alternatively, separate cores can be merged into a single core as well. This is particularly useful when working with large data sets. One of our projects aims to make large Dutch annotated text corpora available to the scientific community. Using separate cores for the indexing process allows us to prepare these corpora in parallel and perform additional checks on metadata and content before merging or adding the new core to the set of Solr cores available for search and retrieval in the production environment.

3.1 Indexing and configurable mapping

The document indexing process itself is a complex process where the original text document is converted to a stream of tokens with possibly multiple tokens on the same position, addition of prefixes, interpretation of ranges and sets of positions as a token, assignment of unique subsequent ids

[5] A special codec extending the default postings format is used. By using this codec, the required files for the forward index are automatically constructed and managed.

to all tokens and finally the construction of individual payloads containing all the right references. Depending upon the processed annotation structures and requested search capabilities, a series of choices has to be made to index available documents. We provide a configurable tokenizer that has been tested against FoLiA, a WPL Sketch Engine like format and TEI among others. Configuration of this tokenizer can be specified in a separate file allowing search options to be adjusted and configured for specific needs. The indexer can thus be instructed to use a different indexing strategy for each individual file to be indexed. Also, the process may be instructed to differentiate between locally available files and remote ones.

The configurable tokenizer is particularly useful in situations where documents using multiple annotation formats are imported into the same index. Apart from the mapping challenges of multiple tag sets, the relevant information content that needs to be extracted from the annotated documents often occurs in different locations in the document. In our projects, by using configuration files put together to match both specific document structure and user requirements, we are also able to collect all information for more complex structures like e.g. entities, paragraphs and chunks from the documents, and include this in the token stream.

The indexer can be instructed about which configuration file to use when indexing a specific document type. In one of our current projects, this is used to index multiple annotation formats produced by different annotation services. Here, users transfer their textual documents to a personal workspace, request some processing service to work upon the document and the resulting annotated document is automatically indexed using this system. Since many of the annotation services produce different formats, this method at least provides the possibility of searching and retrieving such annotations from one uniform index. This method is also considered useful in combination with our archiving software allowing us to make the contents of the archive available not only at the metadata level, but also at the annotated content level while maintaining the flexibility to allow multiple annotation formats to be stored in our repository.

We also took direct advantage of this setup in one of our projects where three data sets were indexed with part of speech encodings from three different tag sets. Rather than using a runtime query expansion mechanism we decided to use one of the tag sets as a pivot, mapped all other tag sets onto the pivot and indexed both the original tag and pivot tag sets values in our index. Each word is thus annotated with multiple part of speech tags and in some cases, even multiple part of speech tags from the same set (e.g. V → V-fin or V-infin).

4 CQL support

Using the new approach based on prefixes and adjusted payloads, the default query parsing mechanisms of Solr and Lucene in most cases will not suffice. Our choice of query language support is furthermore largely motivated by the idea that this should match closely with current practices in the field. This reduces the learning curve for our potential end user community. This also has the practical advantage that front-end development may reuse some of the visual query construction mechanisms already available in the domain targeted at various proficiency levels (beginner, advanced, expert) of end users. Therefore, we support Corpus Query Language introduced by the Corpus Workbench. A CQL parser, based on JavaCC, has been developed mapping CQL queries onto the provided Mtas query methods. Not only does this language seem to be easily apprehensible by users with more specific search requirements, the syntax of this query language also directly matches the prefix/postfix structure we incorporated into the Mtas index structure. A query for a *word* with *part-of-speech* annotation *noun*, represented in the index as a single position token with prefix *pos* and postfix value *N*, is expressed in CQL as

```
[pos="N"]
```

while the search for a paragraph, represented in the index as a multiple position token with prefix *p*, can be performed using angular brackets

```
<p/>
```

The use of the and-operator & and the or-operator |, together with the use of parentheses, provides advanced options for more complex conditions on single words. Multiple conditions may be lined up into sequences, a question mark can be used to mark a part as optional; multiple occurrences of the same part may be indicated with a single number or a minimum and maximum between curly brackets, e.g.

```
[pos="LID" | lemma="the"][pos="ADJ"]{0,2}[pos="N"]
[pos="ADJ"]([word="," | word="and"][pos="ADJ"])?[pos="N"]
```

This can be even further extended by combining these constructed conditions on words and sequences to a new condition by using *containing* or *within* operators, e.g.

```
<entity="loc"/> within (<s/> containing [lemma="amsterdam"])
```

The conventions to search for words at the beginning or end of multiple token annotations, e.g. an adjective at the start of a sentence, or a noun within three positions before the end of a paragraph, closely follow the syntax as known from other formats using this angle bracket notation.

```
<s>[pos="ADJ"]
[pos="N"][]{0,2}</p>
```

By using a dash, the position of a word in the original document can be referred to, e.g. to get the first word of a document, or to query for an adjective within the first ten words

```
[#0]
[#0-9 & pos="ADJ"]
```

In addition to the standard CQL constructs shown above some additional extensions were made to the allow operations that were encountered in specific use cases under consideration while developing Mtas. One feature that was introduced is the ability to request the full prefix list from the system. Although not directly expressed in CQL, it is highly useful to be able to automatically extract this list given that the underlying index may contain arbitrary prefixes depending on the configuration settings while indexing. This list also distinguishes between single and multiple positions allowing to adjust any CQL query accordingly.

The not operator is supported by specifying an exclamation mark in front of the prefix

```
[!pos="ADJ"]
```

Many CQL implementations allow the user to put a word between double quotes as a short hand notation for querying for a single word using the bracket notation. However, since Mtas offers the user full freedom in choosing prefixes to distinguish the different annotation layers, such a notation would be ambiguous without defining the default prefix to apply for such requests. Therefore, requests like the following only can be formed when such a default prefix is provided

```
"the" [pos="ADJ"]?[pos="N"]
```

Furthermore, the multitude of annotation layers may result in queries not matching some results because of annotations unknown to the user. For example, some texts may contain anchors, indicators of some event occurring after the first and before the second word, that would for the following two examples cause the second query to have matches that a user unfamiliar with these anchors would have expected also to match the first query

```
[pos="ADJ"] [pos="N"]
[pos="ADJ"]<anchor/>?[pos="N"]
```

To overcome this problem, an optional *ignore query* can be provided together with each CQL expression, to define everything that should be ignored when searching for sequences and recurrences. By describing the anchor in such an ignore query, the first of the two expressions will now match exactly the same expressions as the latter.

In some of our use cases, the use of a lexicon service to expand queries was required. Instead of defining an explicit value between double quotes, we therefore allow the use of a variable as postfix within the condition of a single position token, where a list of possible values for this variable should be always provided, e.g.

```
[pos="ADJ"][lemma=$1]
```

where $1 will be replaced with items from a list, e.g.: `{"horse", "cow"}`

Although many of the basic queries for annotated texts seem to be covered by our implementation of CQL, especially more complex queries involving syntactic phenomena, such as dependencies, are expected to demand additional query language features to be able to take full advantage of the capabilities of the index.

5 Result delivery

One of the primary use cases for the system, the Nederlab project, currently provides access, both in terms of metadata and annotated text, to over 15 million items for search and analysis as specified in Table 3. Collections are added and updated regularly by adding new cores, replacing cores and/or merging new cores with existing ones. Currently, the data is divided over 23 separate cores. The Nederlab underlying hardware platform is a Dell PowerEdge R730 - Xeon E5 - 2630L v3 (1.8GHz) - 8 x 16 GB - 2 x 2 TB HDD with 67 GB of the available 128 GB memory assigned to Solr.

	Total	**Mean**	**Min**	**Max**
Solr index size	1,146 G	49.8 G	268 k	163 G
Solr documents	15,859,099	689,526	201	3,616,544

Table 3: Size and content of the Solr index consisting of 23 separate cores within the Nederlab project (January 2017).

	Total	**Mean**	**Min**	**Max**
Words	9,584,448,067	654	1	3,537,883
Annotations	36,486,292,912	2,488	4	23,589,831

Table 4: Number of available words and annotations for the 14,663,457 documents containing annotated text (January 2017).

For 14,663,457 of these documents, as described in Table 4, annotated text varying in size from 1 to over 3.5 million words is included. The remaining part of the 15,859,099 documents mentioned in Table 3 concerns descriptions of persons, volumes and other items for which only metadata is available.

On querying the index, Solr allows to filter documents with conditions on metadata in regular fields. By providing a parser plugin, this filtering is extended with the possibility to use CQL conditions on annotated text within reasonable time. Searching for all 1,944,167 documents containing an adjective followed by a noun[6] takes less than 4 seconds, searching for the 161.734 documents with a sentence containing both the word *amsterdam* and the word *rotterdam* takes 6 seconds. Additional restrictions on metadata only reduces the number of potential hits, and therefore result in faster search results.

Many of the features described below have been integrated into the working environment of one of our main infrastructure projects where the translation of statistical information to a user-friendly representation for the end user is performed using pie charts, time line views and other visualization methods. Notice that, although we tried to use both illustrative and realistic examples, the quality of the provided annotations in part of the resources is not quite up to standard, which may sometimes lead to unexpected results.

[6] Only 2.217.779 documents contain text with part of speech annotation.

5.1 Statistics

Whereas Solr only produces statistics on the number of documents, additional methods had to be implemented to produce, within the (filtered) set of documents, statistics on the number of words and the number of hits for specific CQL queries. Furthermore, computing statistics on the composition of these numbers within documents should be possible, e.g. statistics on the number of hits for a CQL query divided by the total number of words within each document.

Number of documents	138,152	Geometric mean	0.16290333781014
Sum	23894.977875106	Variance	0.002695471072453
Mean	0.17296150526309	Population variance	0.0026954515615442
Sum of squares	4505.2933656369	Standard deviation	0.051917926311179
Sum of logs	-250690.38070139	Median	0.17269981462327
Maximum	0.45167923235093	Skewness	0.0043594322359785
Minimum	0.00070521861777151	Kurtosis	0.59294319460155
Quadratic mean	0.18058553060646		

Table 5: Statistics for the number of adjectives followed by a noun divided by the number of nouns within all documents containing an adjective followed by a noun, and with at least 2000 words, computed in 53 seconds.

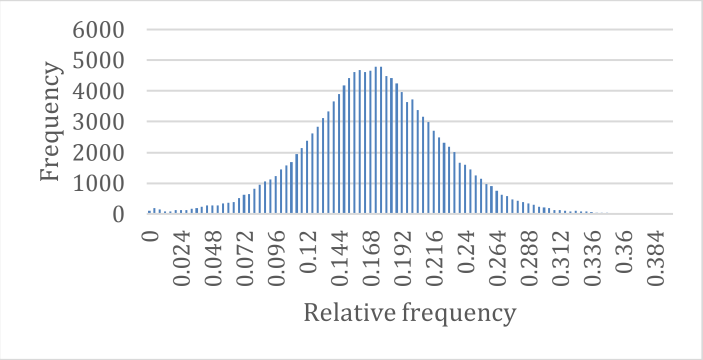


Figure 1: Frequency distribution for the number of adjectives followed by a noun divided by the number of nouns within all documents containing an adjective followed by a noun, and with at least 2000 words, computed in 56 seconds.

As illustrated in Table 5, several statistical properties can be computed, where more advanced items like *median*, *skewness*, and *kurtosis* tend to require more time, since not only a few aggregations but all individual values on document level have to be collected from the participating cores.

Besides these properties, also frequency distributions can be retrieved that can be used to create a graphical representation of the distribution of the studied value. Figure 1 demonstrates such a distribution for the example described in Table 5.

5.2 Faceting

Taking advantage of the available metadata, statistics can be computed for each occurring value of one or multiple metadata fields. This basically extends the available Solr options for faceting with the previously described statistical extensions.

Again, the mean number of adjectives followed by a noun divided by the number of nouns is computed, but now for all documents within a decade. In Figure 2 this mean value is plotted for each decade between 1270 and 2010 for all documents containing at least one noun.

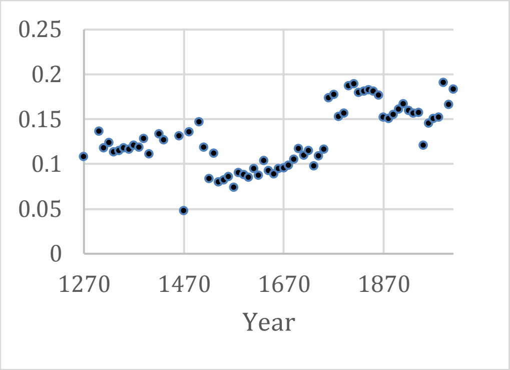


Figure 2: The distribution of the number of adjectives followed by a noun divided by the number of nouns: the mean value computed over all documents with publication year within the same decade is plotted against all 74 adjoining decades between 1270 and 2010 for all documents containing at least one noun.

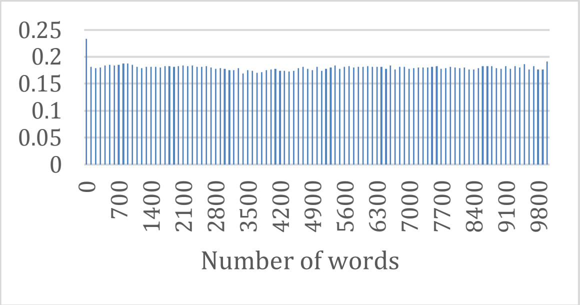


Figure 3: The distribution of the number of adjectives followed by a noun divided by the number of nouns. The mean value computed over all documents with size within the same range of size 100 is plotted against all 100 adjoining ranges of document sizes between 0 and 10,000 for all documents containing at least one adjective followed by a noun.

Besides based on classic metadata fields like year of publication, faceting can also be based on the number of words per document, which is directly derived from the annotated text during indexing. This is illustrated in Figure 3, where instead of using decades, the values found are grouped by and plotted against number of words.

Also, more advanced statistics are available, e.g. the standard deviation as a measure of spread around the mean value as has been illustrated in Figure 4.

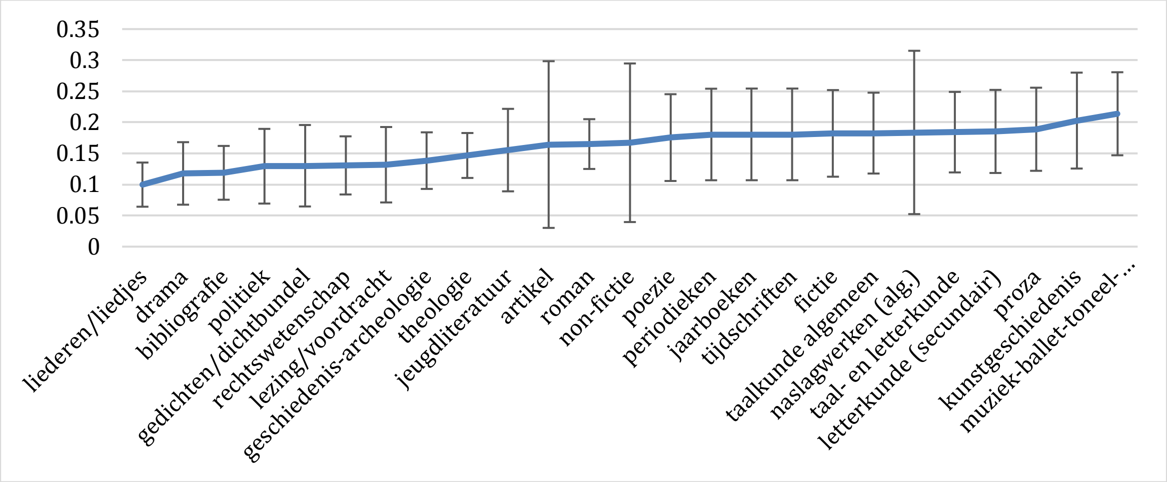


Figure 4: The distribution of the number of adjectives followed by a noun divided by the number of nouns. The mean value and standard deviation computed over all documents within the same genre is plotted for all genres, sorted ascending by mean value, for all documents containing at least one noun.

Finally, in Figure 5 the evolution of the distribution for the mean sentence length of documents within a decade is plotted, illustrating the possibility to study statistics over multiple dimensions.

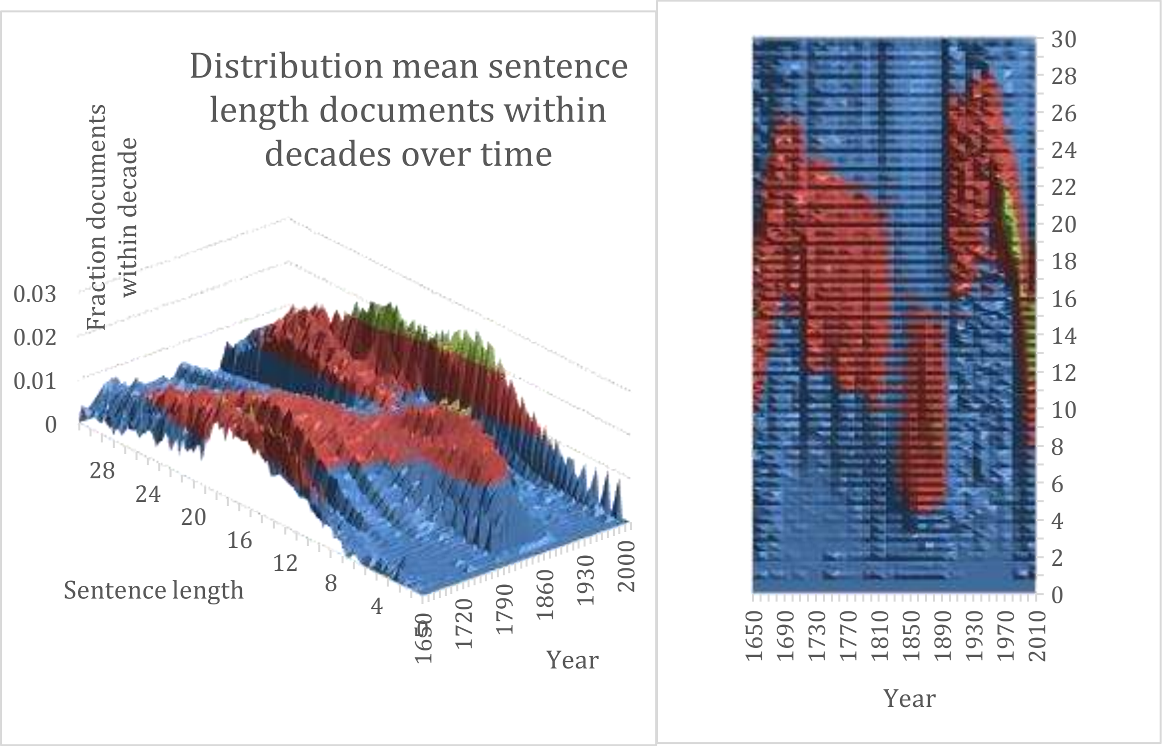


Figure 5: Evolution over time of the distribution for the mean sentence length of documents within a decade; Distribution is plotted for all documents containing at least one sentence and published in or after 1650, necessary data computed in 198 seconds.

5.3 Grouping

The previously presented possibilities of faceting can be seen as statistically grouping results based on metadata. Grouping of query results based on one or multiple annotation layers on the other hand produces lists of occurring values with number of occurrences and documents, sorted by frequency in descending order. These type of queries can be computationally quite expensive, especially for queries with large numbers of hits and also large numbers of distinct associated values for the annotation layer(s).

When grouping the occurring part-of-speech annotations associated with a query for the words *de* or *het*, the number of hits is large: 459,302,283 in 15,859,099 documents. However, there are only ten distinct associated values for the part-of-speech annotation layer, therefore performing this grouping is still possible within reasonable time. The result is illustrated in Figure 6 with, due to the large range, frequencies plotted on a logarithmic scale for each of the occurring values.

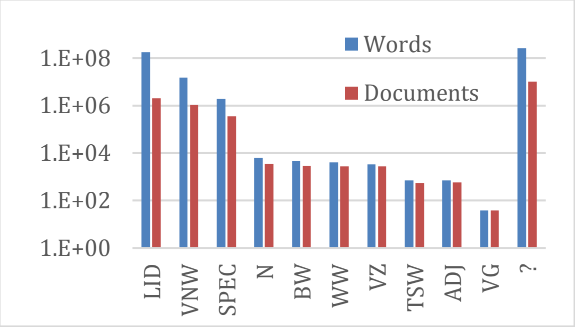


Figure 6: Number of occurrences and documents by grouping a query on “de” or “het” on the associated part-of-speech over all documents, computed in 209 seconds.

On grouping the occurring terms for adjectives followed by lemma *liefde* and for adjectives followed by lemma *haat*, the number of distinct associated values for the occurring terms, 7,400 and 2,605, is much higher. But, since the numbers of hits, 86,973 and 10,381, are substantially lower, performing such a grouping can still be performed within reasonable time. The result, the most frequent adjectives for lemma’s *liefde* and *haat*, are listed in Table 6 and Table 7

ADJ + “liefde”	documents	hits
grote	2,539	3,115
zyne	901	1,968
eene	1,213	1,867
vol	1,454	1,798
ware	1,337	1,749
myne	617	1,586
groote	1,182	1,504
christelijke	991	1,385
eeuwige	899	1,375
oude	1,038	1,166

Table 6: Grouping of the occurring terms for the 86,973 adjectives followed by lemma “liefde”. Computing the 7,400 unique values took 295 seconds, the 10 most frequent values are listed together with number of hits and documents.

ADJ + “haat”	documents	hits
vol	333	363
eeuwige	46	237
algemeenen	211	234
blinde	166	175
felle	149	160
diepe	122	133
fellen	117	128
doodelijken	115	121
onderlinge	105	115
ouden	100	113

Table 7: Grouping of the occurring terms for the 10,381 adjectives followed by lemma “haat”. Computing the 2,605 unique values took 181 seconds, the 10 most frequent values are listed together with number of hits and documents.

Taking advantage of the full possibilities offered by CQL, grouping can be used to retrieve more complex results, e.g. all person or location entities occurring within sentences containing the word *rembrandt*, as listed in Table 8 and Table 9. Notice the multiple token results in the list of person entities.

	documents	hits
rembrandt	3,475	15,009
rubens	354	506
van den	178	274
van gogh	173	206
vermeer	147	192
jan steen	130	156
van der helst	86	152
saskia	69	148
hals	83	144
shakespeare	100	138

Table 8: The 10 most frequent person entities within sentences containing Rembrandt, computed in 65 seconds.

	documents	hits
amsterdam	407	696
nederlandse	153	216
nederland	143	178
nachtwacht	122	170
holland	96	132
leiden	82	125
un	61	122
land	96	118
rijn	92	118
hollandse	69	92

Table 9: The 10 most frequent location entities within sentences containing Rembrandt, computed in 77 seconds.

Finally, grouping does not need to be restricted to a single layer of annotation, as can be seen when grouping on term, part-of-speech, and form for all occurrences of the lemma *zijn*. The 5 most frequent combinations occurring in documents from 1800 are listed in Table 10.

term	pos	tense	form	documents	hits
is	WW	present	pv	913	77,542
was	WW	past	pv	665	44,558
zijn	WW	present	pv	251	24,159
zijne	VNW	-	-	198	20,546
waren	WW	past	pv	456	14,794

Table 10: The 5 most frequent combinations of term, part-of-speech and form when grouping for occurrences of the lemma zijn in documents from 1800, computed in 29 seconds.

5.4 Termvector

A commonly requested feature for information retrieval systems working on text corpora is the ability to extract term lists from retrieved result sets. Our solution is capable of delivering termvectors on any of the annotation layers available in the index (words, lemmas, part of speech, named entities or otherwise) and, combined with the statistical features described above, deliver information on the distribution characteristics in the result set. The list of terms can be sorted on term or frequency, where the latter is more computationally expensive and complex when retrieving results over multiple cores, and can be restricted by a regular expression and/or a user defined set of words.

	Documents	Total	Mean	Median	Max	Mean	Median	Std. deviation	Kurtosis	Skewness
Term	**Frequency**					**Relative frequency**				
welke	4,170,151	12,242,628	2.94	1	3,996	0.0032	0.0021	0.0046	223.4	10.9
zijne	2,628,115	7,541,324	2.87	1	4,523	0.0030	0.0020	0.0036	56.1	5.29
hunne	2,291,385	5,237,717	2.29	1	3,994	0.0025	0.0016	0.0031	63.8	5.53
goede	2,268,787	4,081,615	1.80	1	1,318	0.0027	0.0015	0.0043	709.2	14.2
einde	1,954,052	3,351,655	1.72	1	893	0.0021	0.0012	0.0031	4932	27.3

Table 11: List of the 5 most frequent terms containing 5 letters and ending with e for all 14,663,457 documents. Besides the number of documents, and further statistics on the frequency, also statistics on the relative frequency within the documents are computed in 172 seconds.

Computing data for the list in Table 11, describing the most frequent terms containing 5 letters and ending with *-e* for all documents in all participating cores, took less than 3 minutes. Besides the number of documents, statistics on both frequency and relative frequency are included. The total length of the termvector, describing the total number of matching terms, is not computed by default, since this potentially is a very heavy operation.

5.5 Document

Although computing the full termvector over a set of documents can be quite expensive, as noted in the previous section, this type of computation is less excessive when only a single document is involved. In Table 12, the frequency distribution for a Dutch translation of the bible is illustrated, together with the ten most frequent terms for this document. Computing these results took approximately 6 seconds.

words	2,409,382
unique	47,421

term	**frequency**
ende	126,688
de	97,311
van	50,364
het	38,744
in	34,849
dat	33,020
die	30,010
den	29,271
te	27,758
en	26,516

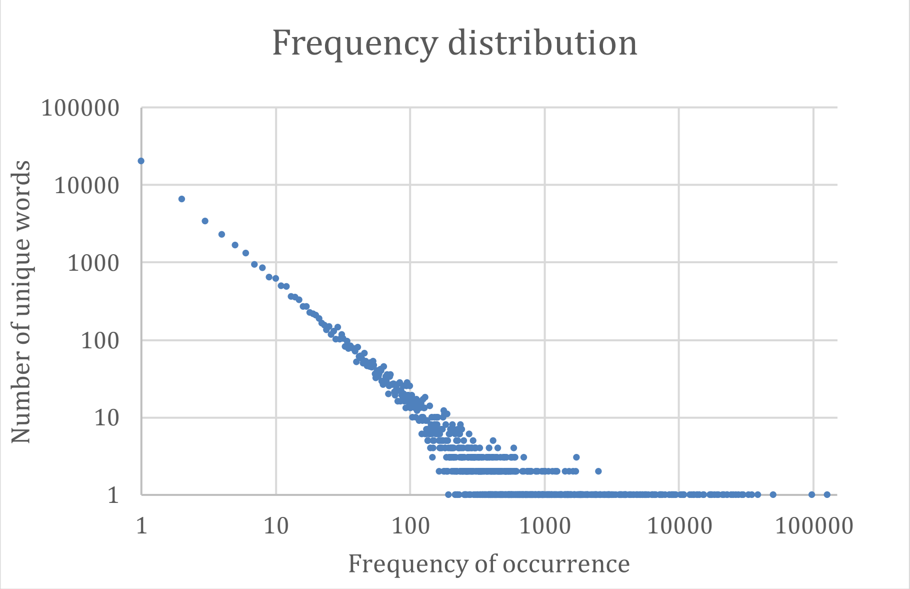


Table 12: Frequency distribution for a single document: the frequency distribution for a Dutch bible translation is computed within 6 seconds; the total number of words and total number of unique words together with the 10 most frequent terms are listed and the distribution of the frequency of occurrence is plotted, demonstrating behaviour as predicted by Zipf's law.

5.6 Keyword-in-context (kwic)

Using Mtas as a plugin, the default presentation of results from Solr, providing (part of) the list of matching documents and listing stored values for all or limited set of fields is extended by providing the option to list a set of matches to one or multiple CQL queries. This keyword-in-context like functionality provides the user with the option to investigate specified annotations on or around the location of hits within the annotated text. This includes multiple-position tokens, positions and hierarchical structure, as can be seen from the example in Figure 7 where such a kwic result from a query for

```
[pos="LID"][pos="ADJ"]
"Amsterdam"
```

is visualized. The application of a forward index, as described previously, makes the additional time needed to generate these representations almost always negligible compared to the time needed for the execution of the query involved.

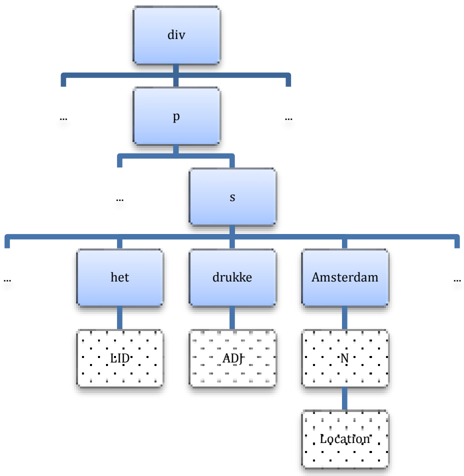


Figure 7:Keyword-in-Context result for a query to an article and and adjective followed by "Amsterdam"

5.7 Consistency checks

When developing Mtas, no suitable reference sets of resources, queries and results were available to test the provided functionality. Direct checks on consistency of the indexation process were therefore limited to manual tests on relatively small documents. However, much of the tokenization process can be tested indirectly with queries, using general knowledge on the structure of the resources. For example

- For most documents, the number of words satisfying the condition of being contained within a sentence, must equal the total number of words.
- For most part-of-speech annotated documents, the sum of the total number of words within each occurring part-of-speech value, must also equal the total number of words.

Furthermore, many aspects of the implemented Mtas functionality could also be tested by comparing results for specific queries. Consistency in these results from different methods, some of them even being native Solr functionality, does indirectly provide a test on those methods themselves. For example

- The number of documents for each term in the native Solr termvector should equal the number of documents in the Mtas termvector result, and also the number of hits for each separate term from the Mtas termvector should equal the number of hits in the Mtas statistics for a query to this specific term.
- The number of documents in native Solr facets should equal the number of documents in the corresponding Mtas facet, and also the number of hits within the Mtas facet response should be reproducible by requesting Mtas statistics with corresponding conditions on the metadata.

Finally, for the implemented Mtas functionality, consistency checks were done in comparing query results over separate cores with results where sharding was applied.

6 Performance

Performance measurements strongly depend on the number of documents, document size, type of query performed and available hardware options [7] Together with differences in implemented functionality, this makes it difficult to really compare Mtas performance with e.g. the solutions listed in Table 1. We tried to provide some indication of the performance by including the required search time in most of the previously introduced examples.

The advantages of a distributed setup in the process of adding and updating data have already been mentioned. The influence of distribution on performance of our implemented system can be illustrated more explicitly. For the graph in Figure 8, basic statistics for the number of sentences were computed multiple times for a setup with a single core, and for setups with multiple cores.

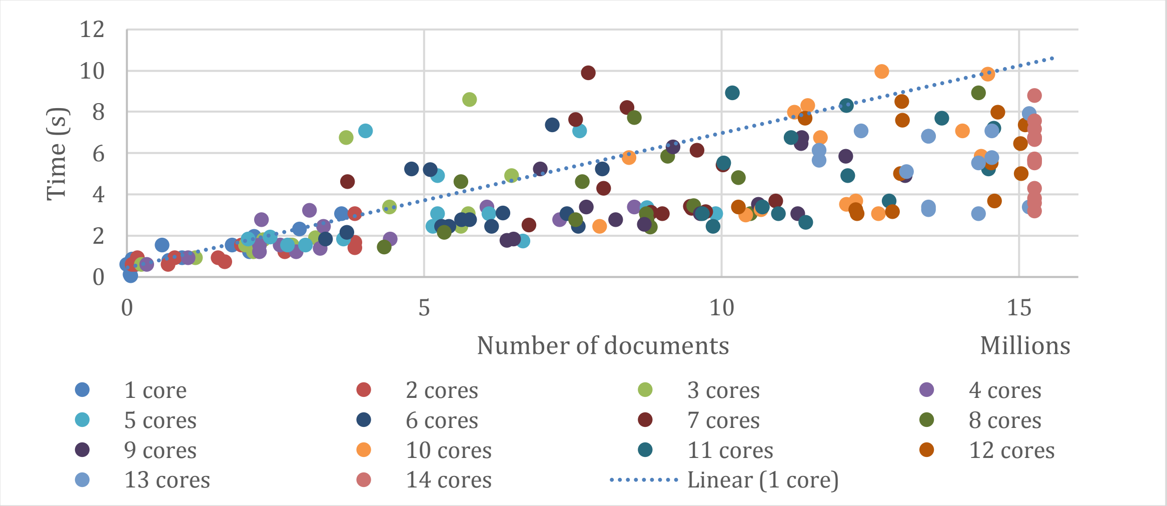


Figure 8: Measurements of query time for basic statistics on the number of sentences against number of matching documents for a single core setup, and for setups with multiple cores varying from two to fourteen.

As can be seen from the graph, most measured times for setups with multiple cores lie below the linear trendline from the single core setup. Total query time is likely strongly to be determined by disk access speed, where the spread in time possibly is caused by the availability time being influenced by disk access in the same location shortly before. The upper and lower limit in this band do not seem to be heavily influenced by the number of cores and/or documents.

[7] Underlying hardware platform is Dell PowerEdge R730 - Xeon E5 - 2630L v3 (1.8GHz) - 8 x 16 GB - 2 x 2 TB HDD; currently, with 67 GB of the available 128 GB memory assigned to Solr.

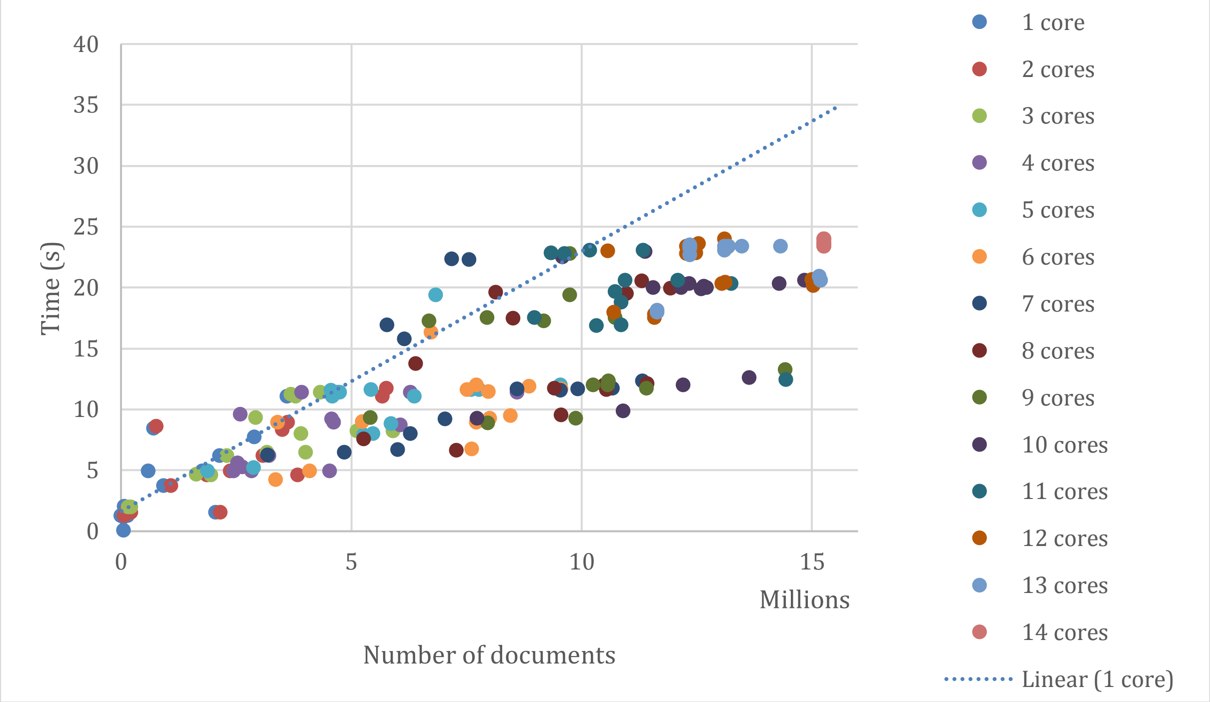


Figure 9: Measurements of query time for computing a termvector against number of matching documents for a single core setup, and for setups with multiple cores varying from two to fourteen.

Another illustration of performance for distributed queries is given in Figure 9, where the required time to compute a termvector sorted by frequency is measured, again in a single and multiple core situation. Again, most measured times for setups with multiple cores lie below the linear trendline from the single core setup. Furthermore, we seem to be able to distinguish two levels in this plot that can be explained by the algorithm used to efficiently compute a termvector over multiple cores. In this approach, sometimes a second termvector has to be computed by individual participating cores to be able to compute the required merged result, especially when the number of documents and/or participating cores increases.

7 Conclusion and future work

We provide a scalable Solr/Lucene based solution, capable of performing CQL queries across a range of annotation formats. Query capabilities have been extended into the statistical domain allowing gathering of statistical information from the retrieved result sets. Our system supports retrieval of termvectors across search results documents. Result delivery features key word in context, listings and groupings.

Although most of the requirements for e.g. the Nederlab project are probably sufficiently covered by the current implementation, multiple additional features seem to be desirable. By using technical or performance related considerations and by watching the search and analysis techniques applied in the research fields involved, several suggestions can be made:

- Exploring the hierarchical structure, already fully integrated in the index structure, is not covered very well in the CQL query language. Further development, e.g. integrating an additional query language to exploit this structure and possibly adjusting the index structure to new types of queries, should preferably be done in collaboration with specialized researchers and based on specific use cases.
- Queries containing CQL conditions seem to perform reasonably well, but little or no attention is paid to including e.g. the number of hits in determining the score value for each document.

Within the use cases at hand, currently no clear thoughts seem to be available with respect to the desired manners of weighing documents. Further adjustments to the scoring mechanism should be accompanied by theory and/or explanations to guarantee acceptance and understanding.

- Within the research areas of use cases involved, new techniques concerning clustering and analysis seem to gain popularity, especially when huge amounts of data are involved. Although some experiments related to these techniques are planned, these projects all seem to rely on very basic and often inefficient use of the possibilities offered by Mtas. Often, users plan to export potentially very large result sets and analyze them with the external tools they are used to. Integrating the computation of e.g. covariance matrices, and furthermore offering options to reuse the found factorizations in further search requests, although probably still keeping the cluster and factorization computations outside Mtas, seems a far more efficient approach, probably also directly applicable by other research projects.
- Whereas currently in Mtas annotated text is assumed to contain a basic granularity on word level, enriched with annotations on both single and multiple word level, some textual data does not completely fit into this scheme. Fully including annotated text containing a translation for example will be problematic, since translations will align probably on sentence or paragraph level, but not (always) on the level of words. Including a decomposition of words into syllables and morphemes also does not seem to fit the current structure.
- As illustrated in examples above, many statistical properties on the number of hits already can be determined. Less attention is paid to e.g. the distribution of these properties within single documents, bootstrapping methods and applying more advanced techniques in comparing documents, although these techniques do seem to applied regularly in the research areas involved.
- Producing termvectors over multiple cores is reasonably fast, but only regular expressions or explicit lists can be used to restrict the outcome. There seems to be a need to reduce these termvectors even further by using conditions on additional layers, e.g. only nouns. To achieve this, without falling back on far less performing grouping methods, adjustments to the indexation process have to be made. This may also improve the speed of other queries involving conditions on multiple annotation levels on the same position or word.
- To get the most relevant terms, TF-IDF for termvectors should be made available as statistic and sort condition, both on document level and for multiple documents within some configurable reference set.

Important for all future development seems to be to focus on a combination of performance and more advanced analysis techniques, preferably driven by use cases from active projects and in collaboration with researchers with experience and basic knowledge of the algorithm involved.

References

Banski, P. et al., 2013. KorAP: the new corpus analysis platform at IDS Mannheim.. s.l., s.n.

Brouwer, M. et al., 2014. Nederlab, towards a Virtual Research Environment for textual data.. s.l., s.n.

Brugman, H. et al., 2016. Nederlab: Towards a Single Portal and Research Environment for Diachronic Dutch Text Corpora.. s.l., ELRA, pp. 1277-1281.

Evert, S. & Hardie, A., 2011. Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. Birmingham, s.n.

Kilgarriff, A., Rychly, P., Smrz, P. & Tugwell, D., 2004. Itri-04-08 the sketch engine. Lorient, s.n.

Meurer, P., 2012. Corpuscle – a new corpus management platform for annotated corpora. In: G. Andersen, ed. Exploring Newspaper Language: Using the web to create and investigate a large corpus of modern Norwegian. s.l.:John Benjamins.

Odijk, J., 2015. Linguistic research with PaQu.. Computational Linguistics in The Netherlands, Volume 5, pp. 3-14.

Reynaert, M., Camp, M. v. d. & Zaanen, M. v., 2014. OpenSoNaR: user-driven development of the SoNaR corpus interfaces.. s.l., s.n., pp. 124-128.

Vandeghinste, Vincent & Augustinus, L., 2014. Making a large treebank searchable online. The SoNaR case.. Reykjavik, s.n., pp. 15-20.

What's in A Name? The Case of *Albanisch-Albanesisch* and Broader Implications

Erhard Hinrichs
Seminar für Sprachwissenschaft
Eberhard Karls Universität
Tübingen, Germany
erhard.hinrichs@uni-tuebingen.de

Alex Erdmann
Department of Linguistics
The Ohio State University
Columbus, Ohio, USA
erdmann.6@buckeyemail.osu.edu

Brian Joseph
Department of Linguistics
The Ohio State University
Columbus, Ohio, USA
joseph.1@dosu.edu

Abstract

This paper offers a use case of the CLARIN research infrastructure (Hinrichs and Krauwer, 2014) from the fields of historical linguistics and the history of linguistics. Using large electronically available corpora of historical English and German, it investigates differences in terminology used in the two languages when referring to the people and the language of Albania. The search and data exploration tools that are available for the DTA and the DWDS corpora as part of the CLARIN-D infrastructure (Hinrichs and Trippel, in press) make it possible to determine semantic change for the terminology under consideration. The paper concludes with a discussion of broader implication of the present use case for the use of historical corpora and the functionality of query tools needed for digital humanities research.

1 Introduction

The name for the country that lies on the western coast of the central part of the Balkan peninsula in south-eastern Europe, as well as for its people and its language, presents interesting variation in both German and, to a far lesser extent, English, raising questions about the nature and the chronology of the variation. The country in question is, in its usual form today in English, *Albania*, the people *Albanians*, and the language *Albanian*, and on the German side, the most usual terms nowadays are *Albanien*, *Albaner*, and *Albanisch*. However, if one looks at materials from a century ago, the picture is somewhat different in that variant forms of the substantival stem are rather widespread in German: *Albanese-* and *Albanier-* for the people and *Albanisch-* and *Albanesisch-* for the language. Even in English, in one author, linguist Leonard Bloomfield (Bloomfield (1914; 1933), the variant *Albanese* for the language name is encountered, as in the following quote from Bloomfield (1933, p. 14), with for emphasis *Albanese* added by the authors:

> *In the same way, finding all these languages and groups (Sanskrit, Iranian, Armenian, Greek,* ***Albanese****, Latin, Celtic, Germanic, Baltic, Slavic) resemble each other beyond the possibility of mere chance, we call them the Indo-European family of languages.*

Since Bloomfield's first academic mentor in linguistics was the Austrian-born Indo-Europeanist Eduard Prokosch and since Bloomfield spent part of his postdoctoral training with leading Indo-European scholars at the University of Leipzig and at the University of Göttingen in 1913-14, one cannot help but wonder whether Bloomfield's choice of the term *Albanese* in place of *Albanian*, the term used by other

Erhard Hinrichs, Alex Erdmann and Brian Joseph 2017. What's in a name? The case of albanisch-albanesisch and broader implications. *Selected papers from the CLARIN Annual Conference 2016.* Linköping Electronic Conference Proceedings 136: 38–53.

contemporary English-speaking scholars, has its roots in the German scholarly tradition. The hypothesis that Bloomfield borrowed the term *Albanese* from German scholarly tradition presupposes that the lemma *Albanese* was, in fact, the preferred way to refer to people of Albania in German at the beginning of the 20th century. This in turn raises the question about the usage patterns in German of the nouns *Albaner*- versus *Albanese*- and the related adjectival forms of *Albanisch*- and *Albanesisch*- at that time.

With the increased availability of large electronic historical language corpora, it has become significantly easier to trace the usage patterns of words and to document changes in word meaning over time. In the present paper, three electronic collections of historical and contemporary German will be consulted to answer these questions and to shed some light on the variation noted above in both German and English: the Google Books collection of digitized English and German books (henceforth: GBCE and GBCG, respectively), the Deutsche Text Archiv (henceforth: DTA; www.dta.de) (Geyken et al., 2011), and the corpus of the Digitales Wörterbuch der deutschen Sprache (www.dwds.de) (Geyken, 2007), both available at the CLARIN Center at the Berlin-Brandenburg Academy of Sciences (BBAW) as part of the CLARIN-D research infrastructure.

The remainder of this contribution is structured as follows: Section 2 contrasts the usage of the term *Albanese* in English and German by consulting the Corpus of Historical American English (COHA) (Davis, 2012), the Google books collections for English and German (Michel et al., 2012) and DTA collections for German. Section 3 utilizes the DiaCollo tool (Jurish 2015) to trace changes in meaning over time for the German words under consideration. Sections 4–6 discuss some methodogical issues and broader implications of the present use case and summarize the results.

2 Comparative Study of Historical Corpora for English and German

A comparative diachronic study of the terms *Albania* versus *Albanien*, *Albanian*/*Albanese* versus *Albanisch*/*Albanesisch*, and *Albanians* versus *Albaner*/*Albanesen* needs to consult historical corpora of both English and German. Since the focus is on American English, the COHA corpus of Historical American English is the most relevant English data source for the present investigation. For historical German, the DTA corpus collections with texts ranging from 1600 to 2000 is used as a data source. Both the COHA and the DTA corpus collections are linguistically annotated and include lemma and part-of-speech information. In addition to the two linguistically annotated corpora, the Google Books collections for English and German were consulted with the help of the Google Ngram viewer. The inclusion of these collections as data sources is motivated by the size of the Google Books collections.

2.1 Results for the COHA corpus of Historical American English

The COHA corpus is a balanced corpus of 400 million words with texts ranging from 1810 to 2000. It is currently the largest corpus of its kind and contains texts from the following genres: fiction, academic writing, magazines and newspapers. For the search string *albanian*[1] , COHA returns 387 occurrences in total, with 10 data points for the 19th century. The query term *albanese* yields a total of 28 occurrences for the following decades (with frequencies shown in parentheses: 1830(1), 1880 (1) 1940 (12), 1950 (4), 1960 (1), 1970 (3), 1980 (5), and 1990 (1). Examination of the linguistic context for each occurrence reveals that only the two data points from the 19th century refer to a person from the country of Albania. All other data points refer to someone named Albanese. These findings show that mere frequency counts can be quite misleading and need to be followed up with an inspection of the context of use for each occurrence or require high-quality named-entity tagging that would identify the proper name usage of the search term.

2.2 Results for the Google Books Collection for English and German

Figure 1 presents the results for all word forms of *Albanian*- and *Albanese*- for the GBCE corpus of English and confirms, as expected, that the former outranks the latter by a wide margin for entire period covered by the GBCE. As is the case for the COHA corpus, almost all data points for *Albanese* in the GBCE concern persons with the last name *Albanese*, rather than persons from Albania.

[1] The search terms for *albanian* and for *albanese* need to be submitted in all lowercase letters in the COHA interface.

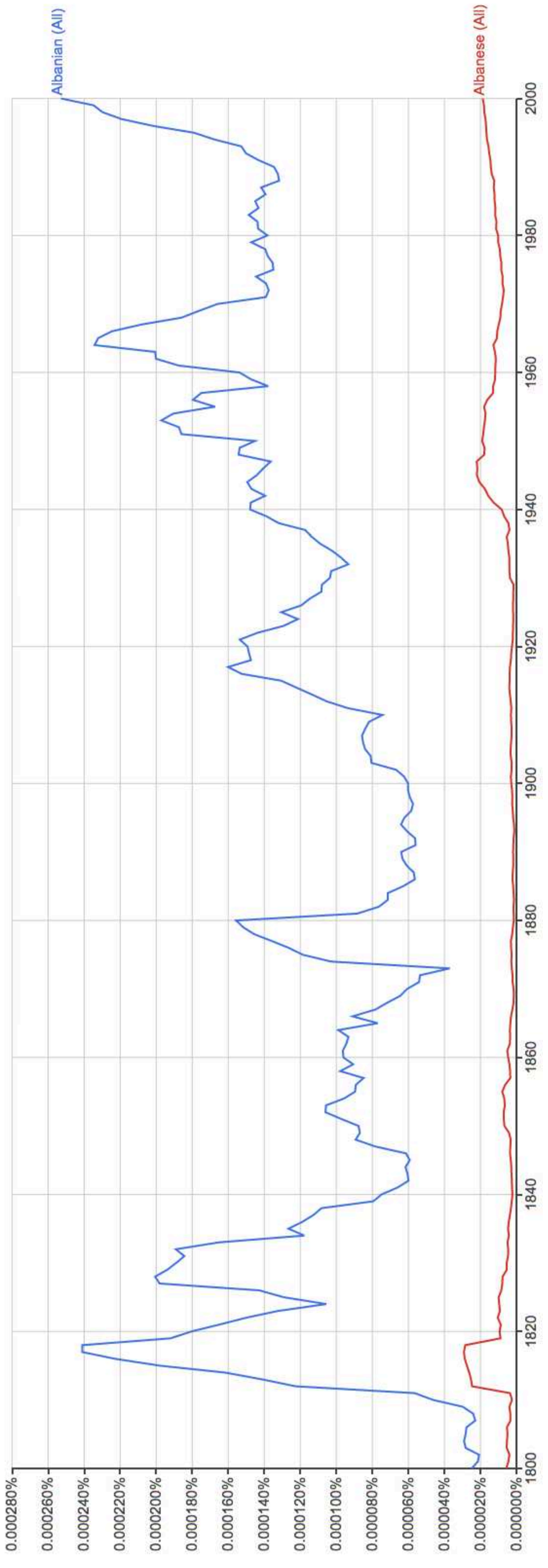


Figure 1: GBCE search results for *Albanian*/*Albanese*.

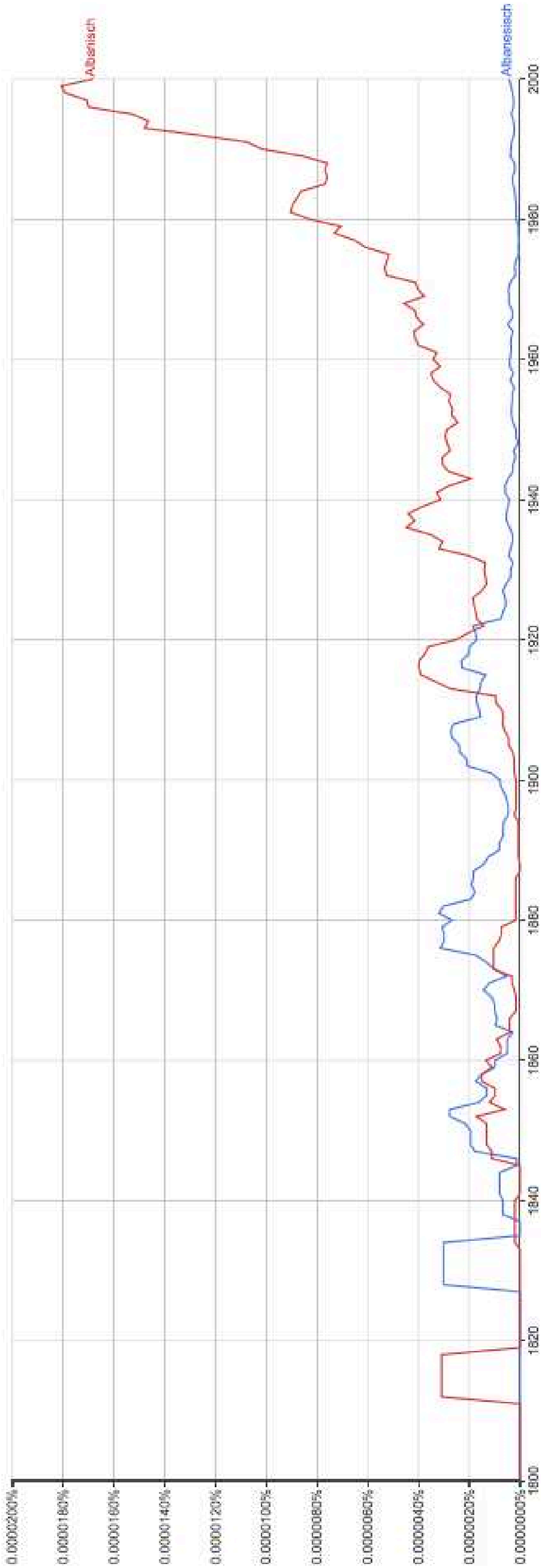


Figure 2: GBCG search results for *Albanisch*/*Albanesisch*.

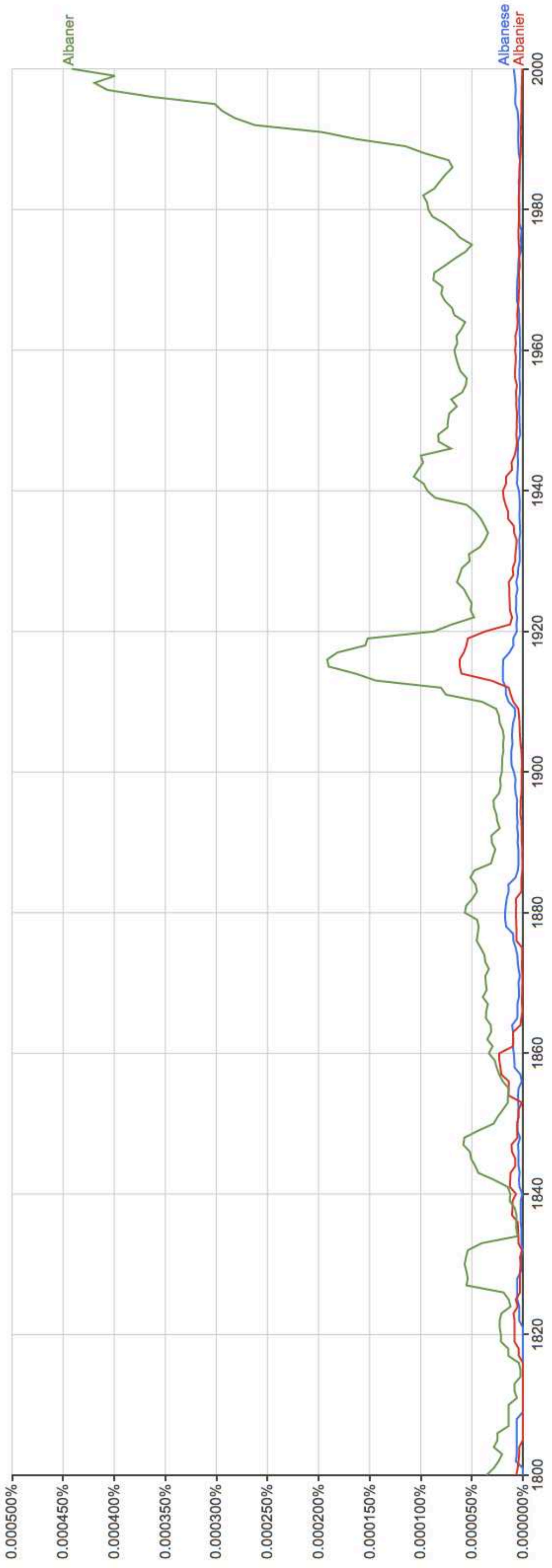


Figure 3: GBCG search results for *Albaner*/*Albanier*/*Albanese*.

Lemma	Frequency Count	Earliest DP	Latest DP
Albanien	109	1627	1913
Albanisch	97	1650	1913
Albanesisch	19	1789	1913
Albaner	61	1663	1913
Albanier	21	1661	1881
Albanese	36	1789	1913

Table 1: DTA query results.

Figure 2 shows the results for all word forms of *Albanisch*- and *Albanesisch*- for the GBCG. *Albanesisch*- outranks *Albanisch*- in relative frequency for most of the 19th century and up until 1914 and then shows a steady decline for the remainder of the century. This result increases the likelihood that Bloomfield may have adopted this term from his German-speaking academic teachers and during his postdoctoral stay in Germany in 1913/14. However, the search results for the nouns *Albaner*- and *Albanese*- in Figure 3 differ from the results in Figure 2 in that the former outranks the latter for entire period covered by the GBCG.

Are we to conclude from Figure 3 that *Albaner* was the preferred term of reference for persons from Albania, with *Albanese* and *Albanier* secondary variants? The mere frequency counts in Figure 3 do not suffice to give a reliable answer to this question. Rather, close inspection of the linguistic contexts for each occurrence of the terms in question is required to determine the intended referent. While the Google Books Ngram Viewer provides links to the digitized objects for each occurrence found for the search terms under consideration, there are a number of limitations due in part to Google's proprietary page ranking algorithm and in part to copyright restrictions. Copyright restrictions prohibit easy and complete inspection of the underlying digitized texts since for some sources only metadata can be provided. Presentation of the data via page rank, rather than by chronological order, makes it difficult to easily detect systematic changes in word meaning for the search terms in question.

2.3 Results for the DTA Corpus

The DTA contains German texts ranging from 1610 to 1900. The texts have been digitized and transliterated, using a high-precision double-keying method. The archive is still under construction. The version used for the present study dates from September 2016 and consists of 142,348,468 lexical tokens with 993,828,135 Unicode characters that are taken from 595,929 digitized pages and 2,448 different published works. The texts represent different genres, including novels and other literary works, scientific and journalistic texts.

The DTA corpus does not suffer from the same limitations as the GBCG. Search results can be rendered in ascending or descending chronological order with open access to all digitized texts via a web application supporting any web browser; seamless linking of facsimiles, digitized object data with the search term highlighted in red in its surrounding context, as well as complete and high-quality metadata all support a comprehensive and reliable inspection of the entire data set. Table 1 provides the frequency counts for the same set of words investigated in the GBCG corpus.

Inspection of the linguistic contexts for all DTA data points reveals that the adjectival and nominal uses of [*a*|*A*]*lbanesisch*- refer to the country or the language spoken in Albania, and all instances of *Albanese* and *Albanier* refer to persons from Albania. By contrast, all instances of the lemma [*a*|*A*]*lbaner*- in the DTA refer to people or locations north of Rome and not to people from Albania, which is the present usage of this lemma. Typical bigrams found in the DTA include *Albaner See* ('Alban lake'), *Albaner Gebirge* ('Alban mountains'), *Albaner Könige* ('Alban kings') as local rivals of the Roman Empire.

Unlike the other three terms, the lemma *Albanisch* has two distinct senses in the DTA, with some of its uses referring to entities related to the territory north of Rome and other instances referring to entities related to Albania. Examples (1) and (2) illustrate these distinct uses, with the term *albanische* in (1) referring to the location north of Rome and in (2) referring to the language spoken in Albania.

(1)

Dass	*Alba*	*als*	*Haupt-*	*und*	*Muttergemeinde*	*galt,*	*ist*	*gewiss,*	*und*	*bloss*	*in*	*diesem*	*Sinn*
That	Alba	as	main	and	mother community	featured	is	certain	and	only	in	this	sense

wird	*Rom*	*auch*	*als*	*albanische*	*Colonie*	*bezeichnet.*
is	Rom	also	Alban	colony	considered	

‘That Alba featured as main and mother community is certain, and only in this sense is Rome considered an Alban colony.

source: Mommsen (1854), p. 29

(2)

Die	*albanische*	*Sprache*	*ist*	*der*	*älteste*	*griechisch*	*aeolische*	*Dialect.*
the	Albanian	language	is	the	oldest	Greek	aeolic	dialect

‘The Albanian language is the oldest Greek aeolic dialect.’

source: Libelt (1828), p. 430

The use of the term *Albanesisch* to refer to the language spoken in Albania may therefore at least partly be motivated by the well-attested and well-motivated pragmatic strategy of trying to avoid ambiguity. Such an ambiguity would have arisen if the term *Albanisch* would have been used instead. The fact that the bigram *albanesische Sprache* occurs with higher frequency than the *albanische Sprache* in the GBCG corpus, as shown in Figure 4, provides cross-corpus evidence for this strategy being at work. Notice also that the bigram *albanesische Sprache* maintains higher frequency in the GBCG corpus for about the same time period during which the unigram *albanesisch* outnumbers the *albanisch*.

Likewise, the choice of the terms *Albanese-*, and *Albanier-* to refer to the people from *Albania* and of the term *Albaner-* to refer to people from a region north of Rome suitably avoids ambiguity of reference.

For the period covered by the DTA corpus, the language of Albania was referred to as *Albanesisch* and the people were referred to as *Albanesen* or *Albanier*. These corpus findings support the hypothesis that Bloomfield’s use of the English term *Albanese* may be due to his close contacts with German scholars who would have used the German cognate. What still remains to be accounted for is the linguistic change that took place in the 20th century, when the term *Albanesisch* was replaced by *Albanisch* as the name of the language, and the term *Albaner* replaced *Albanesen* and *Albanier* as the name of the people of Albania. The DWDS corpus and the tools DiaCollo (Jurish, 2015), which are made available by the CLARIN center at the BBAW make it possible to trace these two changes.

3 Tracking Semantic Change in the DTA and DWDS Collections

The web application DiaCollo collects for a given query term sets of collocates for regular time slices within a text collection. Changes in collocation behavior of a target word are one diagnostic for detecting changes in word meaning over time since the choice of collocates help to disambiguate the meaning of a word.

Figure 5 contrasts the collocates found by DiaCollo for the word *albanisch* in the DTA and DWDS corpus collections. DiaCollo supports continuous word cloud animations for the entire time interval chosen for a particular query. For the query at hand, the time interval is specified by the parameter-value setting `DATE(S):1670-2010`. A continuous DiaCollo animation over the entire interval is available at URL `http://kaskade.dwds.de/dstar/public/diacollo/`. In this paper, we can only show individual frames from this more complete animation. The upper panel in Figure 5 shows the noun *Erz* (‘ore’) as the only collocate in the DTA texts for the decade (`SLICE:10`) starting with 1890; the lower panel provides the five strongest collocates (`KBEST:5`) for the term *albanisch* for the decade of 1960-1970. The collocates are shown as word clouds, which are one of the visualization options offered by the DiaCollo tool and chosen for this query by the parameter setting `FORMAT:cloud`. In the queries shown in Figure 5, the collocates are grouped by lemmas (rather than word forms) and filtered by the part-of speech label `NN` (short for: *normal noun*). This label is part of the Stuttgart-Tübingen tagset (STTS; Schiller et al. (1995)) that has been used for morph-syntactic tagging of the DTA and DWDS corpus collections.

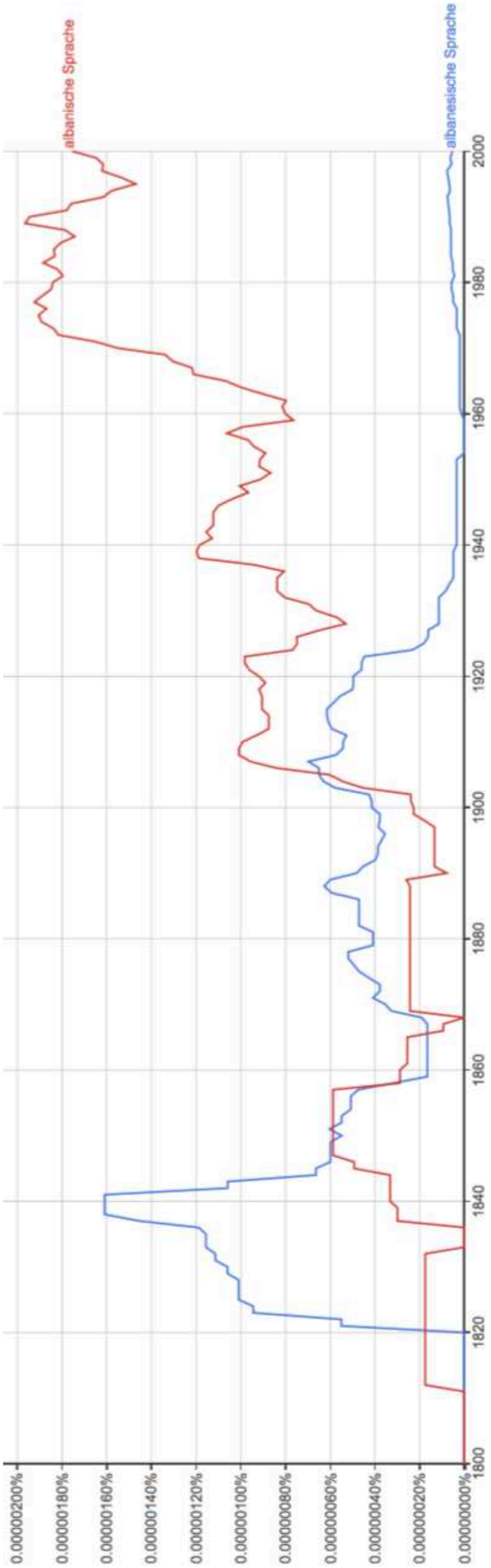
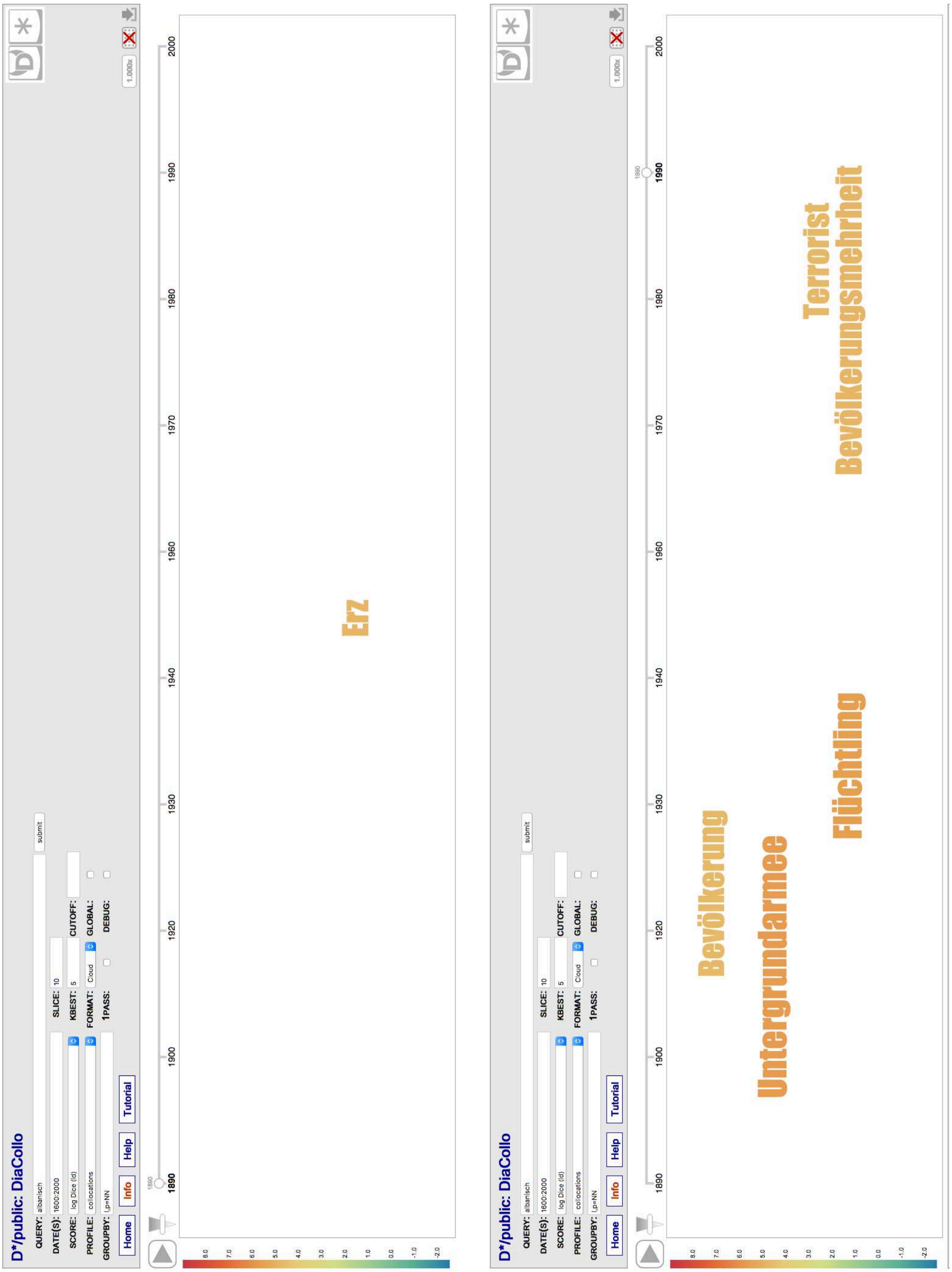


Figure 4: Bigram Comparison of *albanische Sprache* und *albanesische Sprache* in the GBCG.



D*/public: DiaCollo

QUERY: albanisch
DATE(S): 1600:2000
SLICE: 10
SCORE: log Dice (ld)
KBEST: 5
CUTOFF:
PROFILE: collocations
FORMAT: Cloud
GLOBAL: □
GROUPBY: l,p=NN
1PASS: □
DEBUG: □

Erz

D*/public: DiaCollo

QUERY: albanisch
DATE(S): 1600:2000
SLICE: 10
SCORE: log Dice (ld)
KBEST: 5
CUTOFF:
PROFILE: collocations
FORMAT: Cloud
GLOBAL: □
GROUPBY: l,p=NN
1PASS: □
DEBUG: □

Bevölkerung
Untergrundarmee
Flüchtling
Terrorist
Bevölkerungsmehrheit

Figure 5: Collocations of *albanisch* in the DTA and the DWDS corpora: decades of 1890 (left panel) versus 1960 (right panel).

Collocate Noun	From	To
Erz ‘ore’	1890	1910
Patriarchat ‘patriarchy’	1910	1920
Regierung ‘government’	1910	1920
Nationalversammlung ‘national assembly’	1920	1930
Aufständische ‘rebels’	1940	1960
Telegrafenagentur ‘telgraph agency’	1940	1960
Front ‘battle line’	1940	1960
Regierung ‘government’	1940	1960
Grenze ‘border’	1940	1960
Kp Führer ‘Communist Party leader’	1950	1970
Parteiführer ‘Party leader’	1950	1970
Stalinist ‘Stalinist’	1950	1970
Spalter ‘divider’	1950	1970
Marxismus-Leninismus	1950	1970
Ausschluß ‘exclusion’	1970	1980
Partei ‘party’	1970	1980
Volk ‘people’	1970	1980
Antrag ‘petition’	1970	1980
Serbe ‘Serbian’	1980	1990
Jahr ‘year’	1980	1990
Flüchtling ‘refugee’	1990	2000
Bevölkerung ‘population’	1990	2000
Bevölkerungsmehrheit ‘population majority’	1990	2000
Untergrundarmee ‘underground army’	1990	2000
Terrorist ‘terrorist’	1990	2000

Table 2: DTA and DWDS query results for NE collocates of *albanisch*.

Collocation strength is measured by a suite of statistical scores that includes the scaled log-Dice coefficient (`SCORE:log Dice (ld)`). The scaled log-Dice score is defined by the following equation, due to Rychlý (2008), where f_1,f_2, and f_{12} present the raw frequency counts of the collocate, the query term, and of the joint occurrences of the query term and the collocate, respectively.[2]

$$score_{ld} = 14 + log_2(\frac{2 * (f_{12} + \epsilon)}{((f_1 + \epsilon) + (f_2 + \epsilon))} \tag{3}$$

The color spectrum, shown to the left of the word cloud panels in Figure 5, is correlated with the log Dice scores from -2 to 10, in ascending order of collocation strength. Hence, the color used to display a given collocate in a DiaCollo word cloud indicates the collocation strength of the word: *Flüchtling* ‘refugee’ and Untergrundarmee ‘underground army’ are, therefore, the strongest collocate for *albanisch* for the decade 1990-2000 shown in the lower panel in Figure 5.

While the DiaCollo search spans over the time frame of 1610 - 2000, so as to include the time coverage of the DTA and the DWDS, the first decade that yields a common noun with sufficient collocation strength for *albanisch* is the one beginning with 1890. This could either mean that prior to 1890 the lemma *albanisch* did not occur with sufficient frequency itself or that the set of co-occurring lemata was too widely dispersed. Table 2 lists the set of common noun (NN) collocates identified by DiaCollo for the time period from 1890 to 2000 and records the decade during the first and last occurrence for each

[2] Other score functions available in Diacollo include pointwise mutual information and binomial log-likelihood ratio.

collocate. Inspection of the linguistic contexts, in which *albanisch* co-occurs with the noun *Erz* ('ore') reveal that the ore referred to comes from the Alban region north of Rome. By contrast, for all collocates listed for the 20th century in Table 2, *albanisch* is linked to the country of Albania. This suggests that the change in meaning originated at the turn of the 19th and 20th century. The frequent change in the five strongest collocates per decade is indicative of the many changes in the history of Albania during 20th century.

The second semantic change involving the term *Albaner* can also be traced with the help DiaCollo tool. The upper panel in Figure 6 shows the noun *Römer* ('Romans') as the only collocate in the DTA texts for the decade (`SLICE:10`) starting with 1670; the lower panel provides the five strongest collocates (`KBEST:5`) for the term *Albaner* for the decade of 1990-2000.

The disjoint sets of collocates between the decades starting with 1670 and 1990 indicate that the meaning of the term *Albaner* has shifted from referring to people or other entities associated with a territory north of Rome, to the people or other entities from the Balkan country Albania, with the collocates *Serbe* ('Serb'), *Kfor*, *Kfor-Soldat* ('Kfor-soldier'), *Provinz* ('province'), and *Vertreibung* ('forced migration') all salient lemmas at that time, due to the Balkan wars.

Collocate Noun	**From**	**To**
Römer 'Roman'	1670	1880
Gebirge 'mountain range'	1700	1970
Stein 'stone'	1700	1970
Berg 'mountain'	1910	1990
Jahr 'year'	1910	1990
Serbe 'Serb'	1910	2000
Provinz 'province'	1990	2000
Vertreibung 'forced migration'	1980	2000
Kfor 'Kfor'	1980	2000
Kfor-Soldat 'Kfor soldier'	1980	2000
Friedenstruppe 'peace keeping force'	1990	2000

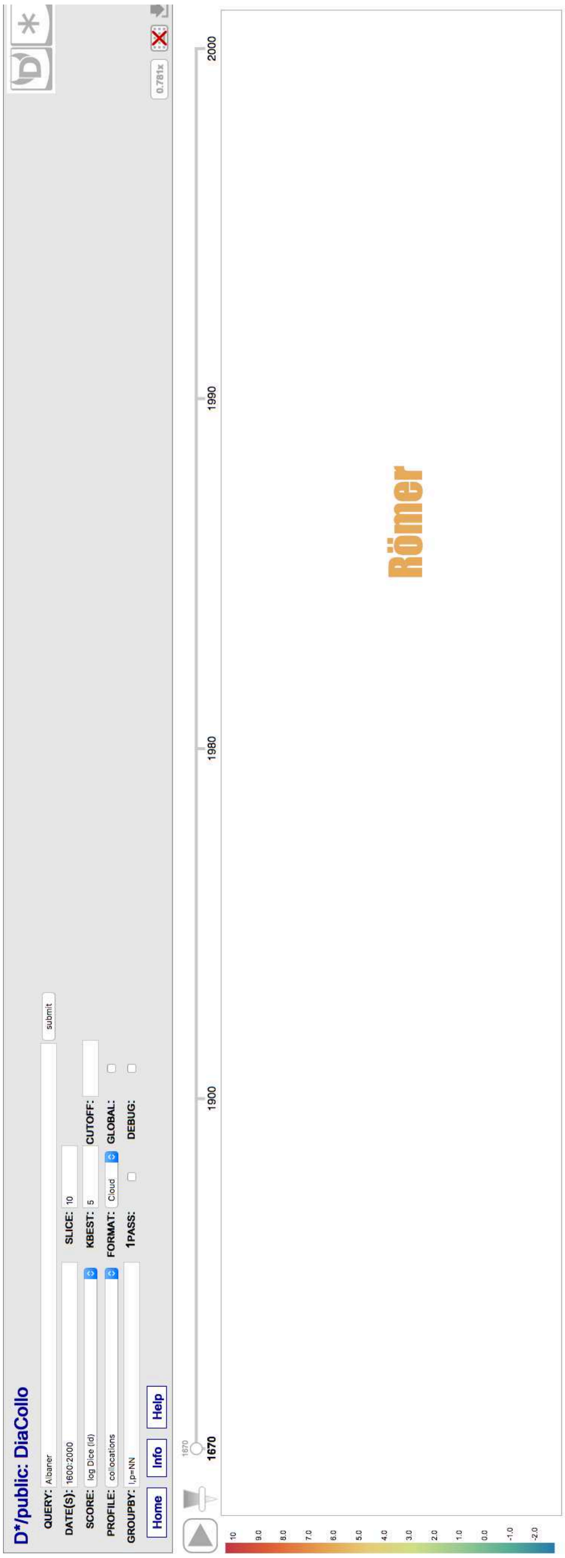
Table 3: DTA and DWDS query results for NN collocates.

Collocate Noun	**From**	**To**
Rocca 'Rocca'	1850	1970
Jugoslawien 'Jugoslavia'	1860	1980
Kosovo 'Kosovo'	1910	2000
Mazedonien 'Macedonia'	1980	1990
Pristina 'Pristina'	1980	1990
Rugova 'Rugova'	1980	1990
UCK 'UCK'	1980	1990
Kosovska Mitrova 'Kosovska Mitrova'	1990	2000
Serbien 'Serbia'	1990	2000

Table 4: DTA and DWDS query results for NE collocates.

Table 3 lists the set of common noun (NN) collocates identified by DiaCollo for the time period from 1670 to 1990 and records the decade during of first and last occurrence for each collocate. While most collocates are clearly indicative of a particular reading of the term *Albaner*, the collocates *Gebirge*, *Stein*, *Berg*, and *Jahr* are not. Examination of the linguistic contexts of the collocates reveals that with the exception of *Jahr*, where *Albaner*- refers to persons from Albania, for all other collocate nouns the query term refers to the Italian region north of Rome.

The term *Serbe* is the first collocate in chronological order that clearly shows the shift in meaning for



D*/public: DiaCollo

QUERY: Albaner

DATE(S): 1600:2000

SCORE: log Dice (ld)

PROFILE: collocations

GROUPBY: l,p=NN

SLICE: 10

KBEST: 5

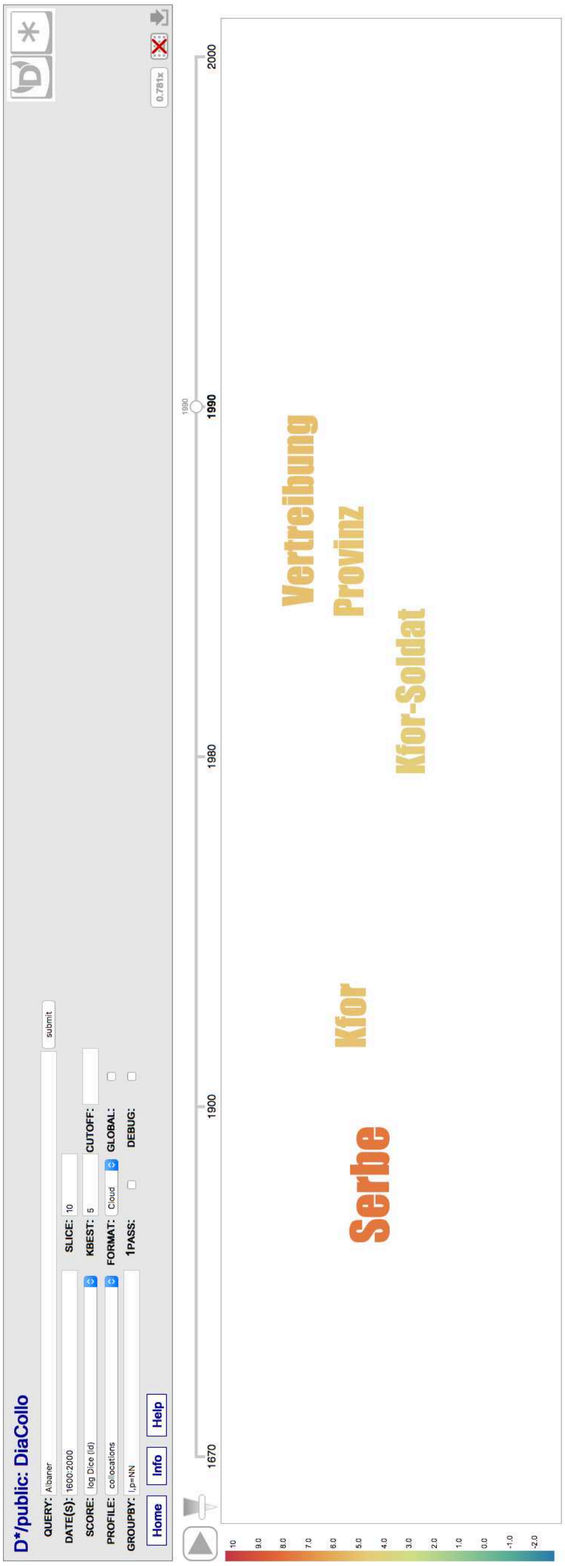
FORMAT: Cloud

CUTOFF:

GLOBAL: □

DEBUG: □

1PASS: □



D*/public: DiaCollo

QUERY: Albaner

DATE(S): 1600:2000

SCORE: log Dice (ld)

PROFILE: collocations

GROUPBY: l,p=NN

SLICE: 10

KBEST: 5

FORMAT: Cloud

CUTOFF:

GLOBAL: □

DEBUG: □

1PASS: □

Figure 6: Collocations of *Albaner* in the DTA and the DWDS corpora: decades of 1670 (left panel) versus 1990 (right panel).

the term *Albaner*. Additional data points can be obtained by a DiaCollo query that filters for collocates belonging to part-of-speech category NE (short for: proper names in the STTS tagset). The results of this query are shown in Table 4. Among the NE collocates, the first occurrence of the collocate Jugoslawien for the query term *Albaner* provides evidence that the shift in meaning is starting already in the second part of the 19th century.

Data mining of the DWDS corpus of the 20th century provides additional evidence of a transitional period between the uses of the term *Albanese* at the beginning and of the term *Albaner* at the end of the 20th century. The DWDS contains a total of 33 occurrences of the term *Albanier* between 1914 and 1991. We suspect as well that homonomy avoidance, documented as a driving force in some semantic change (Hock and Joseph 1996: 224), may have been at work here.

4 Methodological Issues and Wider Implications

This corpus study was prompted by the writings of the American linguist Leonard Bloomfield and aimed at answering a specific research question concerning Bloomfield's use of the English term *Albanese*. The findings of the study have allowed us to track changes in the German lexicon for the language and the people of Albania. While these results are valuable in their own right, it is worthwhile to reflect on some lessons learnt in the course of this investigation. We will concentrate on those aspects and methodological issues which go beyond this particular use case and are applicable more widely to corpus studies based on diachronic language data.

In the present study, we consulted different corpora of various sizes and with varying degrees of linguistic post-processing, including spelling normalization, lemmatization, part-of-speech tagging, as well as collocation and bigram analysis. With Google's Ngram Viewer and DiaCollo word clouds we also utilized two kinds of visualizations to highlight relevant patterns in the data of interest. Such visualizations are essential, given the size of the Google Books and DTA corpus collections, and their utility extends well beyond the type of linguistic study that we are engaged in here to other areas of digital humanities research. In fact, they have given rise to a text-mining approach in its own right under the name of *culturomics*[3] and have have been widely applied in recent years to detect social dynamics of various kinds and different humanities disciplines.

As the proponents of culturomic methods have pointed out themselves, it is important to be aware of the limitations of the Google Books corpus data[4]. These limitations include errors due to optical character recognition (OCR), metadata quality, the opportunistic data collection method, and lack of lemmatization for the Google Books corpus for German. Historical texts are particularly prone to OCR errors, and metadata quality can be unsystematic if the texts included in the corpus are not first editions. Since there are no published evaluation results on these two issues for the Google Books corpus, its reliability is unknown. For the DTA, published results on both matters are available. (Haaf et al., 2013) overall accuracy rate of 99.9909% for a balanced DTA text sample, which implies on average 91 transcription errors in one million characters. The data collection policy for the DTA adheres to the principle that only first editions of individual texts are included in the DTA collection so as to ensure highly reliable metadata. The reliability of metadata is particularly important for search terms that occur with low frequency in a given corpus, as is the case for the set of lexical items under investigation in this paper. Unreliable metadata, due to automatic harvesting methods and/or lack of information about first editions, can lead to rather distorted results about historical trends in word usage and frequency.

Another difference between the Google Books collection for German and the DTA corpus collections concerns the lack of linguistic analysis and annotation for the former. The Google Books corpus for German is not lemmatized, and Google Ngram queries do not support regular expressions. Taken together, this means that only word form frequencies and no lemma frequencies can be displayed in the Google Ngram Viewer. Such limitations do not apply to DTA queries since the DTA data are lemmatized and the DTA's query language DDC (Jurish et al., 2014) supports searching for word forms (tokens) and lemmata. For morphologically rich languages like German this functionality is essential.

[3] See, inter alia, (Michel et al., 2012; Lieberman et al., 2007).

[4] See `//www.culturomics.org/Resources/faq` for a more in-depth discussion.

The search functionalities for the Google Books collection for German and for the DTA corpus differ not only in expressivity of the underlying query languages, but also in terms of the information that these two well-designed web applications convey. The main goal of the Google Ngram viewer is to visualize changes in the frequencies of ngrams over time. This is what makes it so attractive for diachronic corpus studies. The main focus of the DTA query interface, on the other hand, is on the seamless rendering and browsing of different textual views: a keyword-in-context view for a particular query along with a pointer to the relevant section of the underlying manuscript and its transcription. The keyword in context presentation of each data point is essential for a careful examination of the meaning of the query term and avoids the danger that arises if only unigram frequencies can be compared. In the course of the present investigation, we encountered precisely this type of situation, when the query term *Albanese* was a proper name, rather than the type of referent we were interested in. This potential error was only detected by consulting the source text in the Google Books collection. However, due to copyright restriction such double-checking is not always possible.

The above discussion shows that corpora such as the DTA, whose construction is quite labor-intensive, due to the amount effort required in double-keying, spelling normalization, linguistic annotation, and manual metadata creation have distinct advantages in data accuracy and reliability over the Google book corpus collections. However, this does not mean that the Google Books collections are irrelevant for diachronic linguistic studies. They are very useful as a secondary source of information that help to double-check the validity of results obtained from corpora such as the DTA.

5 Further Applications

While this study documents the ways in which certain words have waxed and waned in their use and frequency, with consequences for their meaning, there are wider implications that go beyond those important lexical details. In particular, the value of the corpora consulted and of the search tools they provide has clearly been demonstrated by the results that they allow for. At the same time, these results show that there are limitations on lexeme-based searches, in that our understanding of the developments that the *Alban(es)*- lexical items underwent crucially emerged from an examination of the context for each item, provided by the corpora and tools, disambiguating Italian *Albaner* from Balkan *Albaner*. These developments in turn provided some insight into mechanisms for semantic change viewed “up close” in a relatively short time span. Finally, it is a well-known problem in dealing with names of peoples and of groups that one and the same group can have multiple names in different, even related, traditions (e.g. *Deutscher*, *German*, *allemand*, etc.); this problem is acute in the case of group names from the distant past. The example of *Alban(es)*- shows how it is possible to untangle multiple names for the same referent through careful corpus searches and accompanying manual work. The ability to do so enhances, for instance, the prospects of undertakings like the Herodotos Project (`https://u.osu.edu/herodotos/`), aimed at developing a comprehensive listing of group names mentioned in Classical sources and modeling the networks of those groups.

In support of the Herodotos Project, Erdmann et al. (2016) employ a Named Entity Recognition system that identifies textual references to group names and the personal and place names with which they cooccur, but they have yet to disambiguate the nature and context of each reference. Like the *Alban(es)*- lexical items, there are many references in Ancient Latin or Greek corpora that can map to multiple concepts, meaning that one name could refer to any one of several groups, given the context. Conversely, there are many concepts that map to multiple references, and furthermore, the very identity of these concepts and the nature of these mappings can change over time, just as *Albaner* evolved from referring to an area near Rome to referring to the Albanian people. The same data-driven approach combined with manual analysis employed in this paper can elucidate such evolving reference-concept mappings, enabling projects like the Herodotos Project to better understand the relationships between the named entities it extracts from historical texts.

One example of a case study of interest to the Herodotos Project is provided by the term *Thebes* and references *Thebans* in Greek. *Theban*, of course, refers to people from Thebes, a geographically based designation. However, the geography is ambiguous in that there is a Thebes in ancient sources in Egypt and one in Greece (specifically in the region called Boeotia), as well as a few others too, so that Greek Θηβαῖοι (Thēbaîoi) could in principle refer to people from either city and in fact any use of the stem Θηβα- (Thēba-) would be potentially ambiguous. Typically contextual information can disambiguate various instances of the stem Θηβα-. For instance, in Iliad 14.113-4, Diomedes says:

(4) πατρὸς δ᾽ ἐξ ἀγαθοῦ καὶ ἐγὼ γένος εὔχομαι εἶναι /
patròs d’ ex agathoû kaì egō génos eúkhomai eînai /

Τυδέος, ὃν Θήβῃσι χυτὴ κατὰ γαῖα καλύπτει
Tydéos, hòn Thēbēsi khytē kata gaîa kalyptei

‘I too can declare my stock to be from a noble father, Tydeus, whom the heaped earth in Thebes now covers’

In this case, the reference to the Greek hero Tydeus, within the immediate context here, serves to locate the referent of Θήβῃσι ‘in Thebes/DAT.PL’, as the Greek Thebes, not the Egyptian one. Similarly, reference in Homeric epic to “Thebes with one hundred gates” indicates a reference to the Egyptian city, not the Greek one.

There are many more cases like the Thebes case, since many colonies were named after the home city of the colonizers; for instance, there is a Κύμη (Kymē) in Euboea, an island just east of central Greece, but there is also one in southern Italy which was founded by settlers from the Greek city. Contextual cues such as those leveraged by the DiaCollo tool can similarly differentiate which place is referenced in such cases.

One final example comes from the opening of Caesar’s *De Bello Gallico* in which he asserts that one of the groups inhabiting Gaul is known as *Celti* in their language, and *Galli* in Latin (Caesar, *BG* 1.1). In other words, Caesar not only maps two references to one concept, but also clarifies who is more likely to use which reference. In this case, gathering collocations for both references in a range of texts and analyzing the output would shed critical light on perspectives and biases before and after Caesar’s campaign into Gaul. The sentiment of such collocations would demonstrate how *Galli*-reference users depicted the Gauls as compared to *Celti*-reference users. A diachronic analysis would address questions such as whether these biases converged or diverged after Caesar’s campaign and how long convergence/divergence took in addition to other such questions. Perhaps the two references are actually better modeled as referring to two distinct but related concepts: the romanticized idea of *Celti* versus the barbarian caricature of the *Galli*. This stance can be supported or rejected with such an investigation, and developing a language-agnostic version of DiaCollo would greatly contribute to the feasibility and success of such investigations.

6 Conclusion

The methodology demonstrated here, combining automatic collocation analysis and manual inspection of the results can both identify and shed light on complex relationships between lexical items and the concepts they refer to, as well as how those reference mappings evolve over time. While DiaCollo enables this process to be very effective in German, a comparable language-agnostic tool will need to be developed in order to address many cases of interest in other languages without requiring potentially problematic translation. Regardless of the future of such technology, it is clear that homonymous named entities and entities with multiple names present a major challenge for attempts to automatically infer information about them. We present here a framework for addressing such challenges in a manner sophisticated enough to inform and impact qualitative research in the humanities.

Acknowledgements

The research of the first author was supported by the CLARIN-D grant provided by the German Ministry for Education and Research (BMBF). All authors express their gratitude to three anonymous reviewers for their detailed comments on an earlier version of this paper, which was presented at the CLARIN Annual Conference in Aix-en-Provence. Special thanks go to Bryan Jurish of the Berlin-Brandenburg Academy of Sciences for his extensive help with the use of the DiaCollo tool.

References

[Bloomfield1914] Leonard Bloomfield. 1914. *Introduction to the Study of Language*. Henry-Holt, New York.

[Bloomfield1933] Leonard Bloomfield. 1933. *Language*. Henry-Holt, New York.

[Davis2012] Mark Davies. 2012. Expanding Horizons in Historical Linguistics with the 400 Million Word Corpus of Historical American English. *Corpora*, 7:121–157.

[Erdmann et al.2016] Alexander Erdmann, Christopher Brown, Brian Joseph, Mark Janse, Petra Ajaka, Micha Elsner and Marie-Catherine de Marneffe. 2016. Challenges and Solutions for Latin Named Entity Recognition. *Proceedings of the Language Technologies for the Digital Humanities Workshop* in conjunction with the *26th International Conference on Computational Linguistics* (COLING-2016), December 2016.

[Geyken2007] Alexander Geyken 2007. The DWDS Corpus: A Reference Corpus for the German Language of the 20th Century. C. Fellbaum ed. *Collocations and Idioms: Linguistic, lexicographic, and computational aspects*. Bloomsbury Academic, London. p. 23-41.

[Geyken et al.2011] Alexander Geyken, Susanne Haaf, Bryan Jurish, Matthias Schulz, Jakob Steinmann, Christian Thomas und Frank Wiegand. 2011. Das Deutsche Textarchiv: Vom historischen Korpus zum aktiven Archiv. S. Schomburg et al. eds. *Digitale Wissenschaft. Stand und Entwicklung digital vernetzter Forschung in Deutschland*. pp. 157-161.

[Haaf et al.2013] Susanne Haaf, Frank Wiegand, and Alexander Geyken 2013. Measuring the Correctness of Double-Keying: Error Classification and Quality Control in a Large Corpus of TEI-Annotated Historical Text. *Journal of the Text Encoding Initiative* (jTEI) 4.

[Hinrichs and Krauwer 2014] Erhard Hinrichs and Steven Krauwer 2014. The CLARIN Research Infrastructure: Resources and Tools for E-Humanities Scholars. *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (LREC-2014), May 2014, pp. 1525–31.

[Hinrichs and Trippel in press] Erhard Hinrichs and Thorsten Trippel in press. CLARIN-D: eine Forschungsinfrastruktur für die sprachbasierte Forschung in den Geistes- und Sozialwissenschaften. *Bibliothek Forschung und Praxis*; Vol. 41.1 (April 2017).

[Hock and Joseph1996] Hans Henrich Hock and Brian Joseph. 1996. *Language History, Language Change, and Language Relationship. An Introduction to Historical and Comparative Linguistics*. Mouton de Gruyter (2nd edn., 2009), Berlin.

[Jurish et al.2014] Bryan Jurish, Christian Thomas, and Frank Wiegand. 2014. Querying the Deutsches Textarchiv. In: U. Kruschwitz, F. Hopfgartner, and C. Gurrin eds.: *Proceedings of the Workshop MindTheGap 2014: Beyond Single-Shot Text Queries: Bridging the Gap(s) between Research Communities* (co-located with iConference 2014, Berlin, 4. März, 2014), p. 25–30.

[Jurish2015] Bryan Jurish. 2015. DiaCollo: On the Trail of Diachronic Collocations. K. De Smedt ed. *Proceedings of the CLARIN Annual Conference 2015*. Wrocław, Poland, 15th-17th October, pp. 28-31.

[Libelt1828] Karol Libelt. 1828. *Wykłady Humboldta na uniwersytecie Berlińskim: notaty prelekcyj tych po uczniu Jego Karolu Libelcie* [= Nachschrift der 'Kosmos-Vorträge' Alexander von Humboldts in der Berliner Universität, 3.11.1827–26.4.1828].

[Lieberman et al.2007] Erez Lieberman, Jean-Baptiste Michel, Joe Jackson, Tina Tang, and Martin Nowak Quantifying the Evolutionary Dynamics of Language. *Nature* 449 (2007).

[Michel et al.2012] Jean-Baptiste Michel, Yuan Kui Shen, Aviva P. Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak and Erez Lieberman Aiden 2012. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*, DOI: 10.1126/science.1199644.

[Mommsen1854] Theodor Mommsen. 1854. *Römische Geschichte. Bd. 1: Bis zur Schlacht von Pydna*. Leipzig, Germany.

[Rychlý2008] Pavel Rychlý. 2008. A Lexicographer-friendly Association Score. *Proceedings of the Second Workshop on Recent Advances in Slavonic Natural Language Processing RASLAN 2008*, pp. 6–9.

[Schiller et al.1995] Anne Schiller, Simone Teufel, and Christine Thielen. 1995. *Vorläufige Guidelines für das Tagging deutscher Textcorpora mit STTS*. Technical Report. Universität Stuttgart, Institut für maschinelle Sprachverarbeitung, and Seminar für Sprachwissenschaft, Universität Tübingen.

[Zhang2015] Sarah Zhang. 2015. The Pitfalls of using Google Ngram to study Language. *Science* 10.12.15.

Polish Read Speech Corpus for Speech Tools and Services

Danijel Koržinek and **Krzysztof Marasek** and **Łukasz Brocki** and **Krzysztof Wołk**
Polish-Japanese Academy of Information Technology,
Warsaw, Poland
(danijel,kmarasek,lucas,kwolk)@pja.edu.pl

Abstract

This paper describes the speech processing activities conducted at the Polish consortium of the CLARIN project. The purpose of this segment of the project was to develop specific tools that would allow for automatic and semi-automatic processing of large quantities of acoustic speech data. The tools include the following: grapheme-to-phoneme conversion, speech-to-text alignment, voice activity detection, speaker diarization, keyword spotting and automatic speech transcription. Furthermore, in order to develop these tools, a large high-quality studio speech corpus was recorded and released under an open license, to encourage development in the area of Polish speech research. Another purpose of the corpus was to serve as a reference for studies in phonetics and pronunciation. All the tools and resources were released on the the Polish CLARIN website. This paper discusses the current status and future plans for the project.

1 Introduction

Much of the data used in Humanities and Social Sciences (HSS) research is stored in the form of audio recordings. Examples of this include radio and television programmes, interviews, public speeches (e.g. parliament, public events), lectures, movies, read literary works and other recordings of speech. This data contains valuable information from many aspects of HSS research. This encompasses both the linguistic (with the emphasis on vocabulary and pronunciation) and sociological (emphasis on speaker) points of view. During our project, we have met many scientists that have shown interest in processing either already available data and corpora, or would like to process recordings (e.g. interviews) they intended to make in the future.

The main issue with processing acoustic data is that it is more expensive and time consuming than, for example, traditional, textual data. It demands both the know-how and lots of effort to achieve comparable results. That is why it is often overlooked by researchers who either do not have the time or the funding to deal with such issues. Our primary goal was to create free and accessible solutions for researchers from the HSS community.

Similar efforts in other Clarin consortia already exist, like WebMAUS (Kisler et al., 2016) speech segmentation services at LMU, AVATech (Lenkiewicz et al., 2012) by Max Planck Institute and Fraunhofer Institute which provide video and audio processing services including speech segmentation, VAD and speaker diarization, and TTNWW (CLARIN-NL, 2013) which includes speech transcription services for Dutch. It is worth noting that many of them (although not all) are language dependent, requiring a re-implementation of these services in individual countries.

This paper will first describe the tools and corpora created during the project. Next, it will describe a few of the existing and planned applications of the tools and services. Finally, it will describe the plans of development for the upcoming years.

Danijel Koržinek, Krzysztof Marasek, Łukasz Brocki and Krzysztof Wołk 2017. Polish read speech corpus for speech tools and services. *Selected papers from the CLARIN Annual Conference 2016.* Linköping Electronic Conference Proceedings 136: 54–62.

2 Speech Tools

One of the earliest decision during the project was to release all the tools in the form of web services, rather than downloadable applications. There are many advantages to this: ease of use (no installation required), better support, stable environment and performance. However, a few disadvantages as well: more effort required from the consortium, increased response time if many people use the platform, issues with releasing sensitive data. Most of these have been addressed individually and by releasing the source code of the tools for ambitious individuals.

The main website located at `http://mowa.clarin-pl.eu` was divided into three sections: speech corpora downloads, grapheme-to-phoneme (G2P) conversion and the rest of the speech processing tools. The reason for removing the G2P from the rest of the tools is because it uses a different set of modalities (i.e. text-to-text) from the rest of the tools (audio and optionally text into text).

The tools were selected from the basic pipeline used in speech processing and speech recognition. Rather than hiding them, it was decided to expose the intermediary steps in the pipeline as standalone services. Figure 1 shows a typical speech processing pipeline used to produce the information shown at the bottom of that graph. The grey-filled blocks are the services that were exposed as standalone services. The remaining three blocks were omitted either because they were already available in other forms (language modeling) or because they were deemed not too useful for HSS research (audio analysis, acoustic modeling).

2.1 Grapheme-to-phoneme conversion

This tool allows converting any text written in the orthographic (i.e. written) form into its phonetic (i.e. spoken) form. It is one of the primary steps in any process that involves speech data. but may also serve as a tool outside of the acoustic speech processing context.

The tool is created using a rule-based system. It accepts any form of text, although it does not perform text normalization (it does not expand numbers, dates or abbreviations automatically). The system is completely rule based and contains a list of exceptions for names, foreign and other atypical words. A statistical system, based on the Sequitur (Bisani and Ney, 2008) tool, is also available but due to available data, it does not outperform the rule based system in any way. The tool can generate both word lists (with multiple pronunciations) and a canonical transcription of the text.

There are several enhancements planned for the future of this tool. The foremost includes the normalization, mentioned above. This problem is particularly difficult for inflected languages (e.g. Slavic languages). This work is already in progress as of writing this paper (Brocki et al., 2012). Other improvements would include adding different forms of phonetic alphabets (while retaining the same rule-set) and possibly adding other levels of annotation (e.g. accents and syllabification). Some uses may also benefit from using a graph-based representation, but these (and other) improvements will be added pending further interest from the community.

2.2 Speech-to-text alignment

Speech alignment is one of the most useful tools available. It is used to align a sequence of words to the provided audio recording of speech. This can be understood simply as automatically generating a set of time-codes, when both the audio and its transcription are known. It is a very useful tool because it can be used to easily look up specific events in large sets of recordings. It also makes possible to compute statistics related to the duration and other characteristics of individual speech events.

The tool was created, based on the SailAlign (Katsamanis et al., 2011) concept, in order to work efficiently with long audio files. The engine is constructed around the Kaldi toolkit (Povey et al., 2011), just like most of the tools in this paper, but the main work-flow is managed using a set of libraries written in Java. The alignment is produced both on the level of words and phonemes. The tool currently only generates outputs in the form of a Praat TextGrid file (Boersma and others, 2002), but others could be added in the future. The service also generates a link to the new EMU-webApp website (Winkelmann and Raess, 2014), which allows viewing the result of the segmentation directly in the browser (see figure 3).

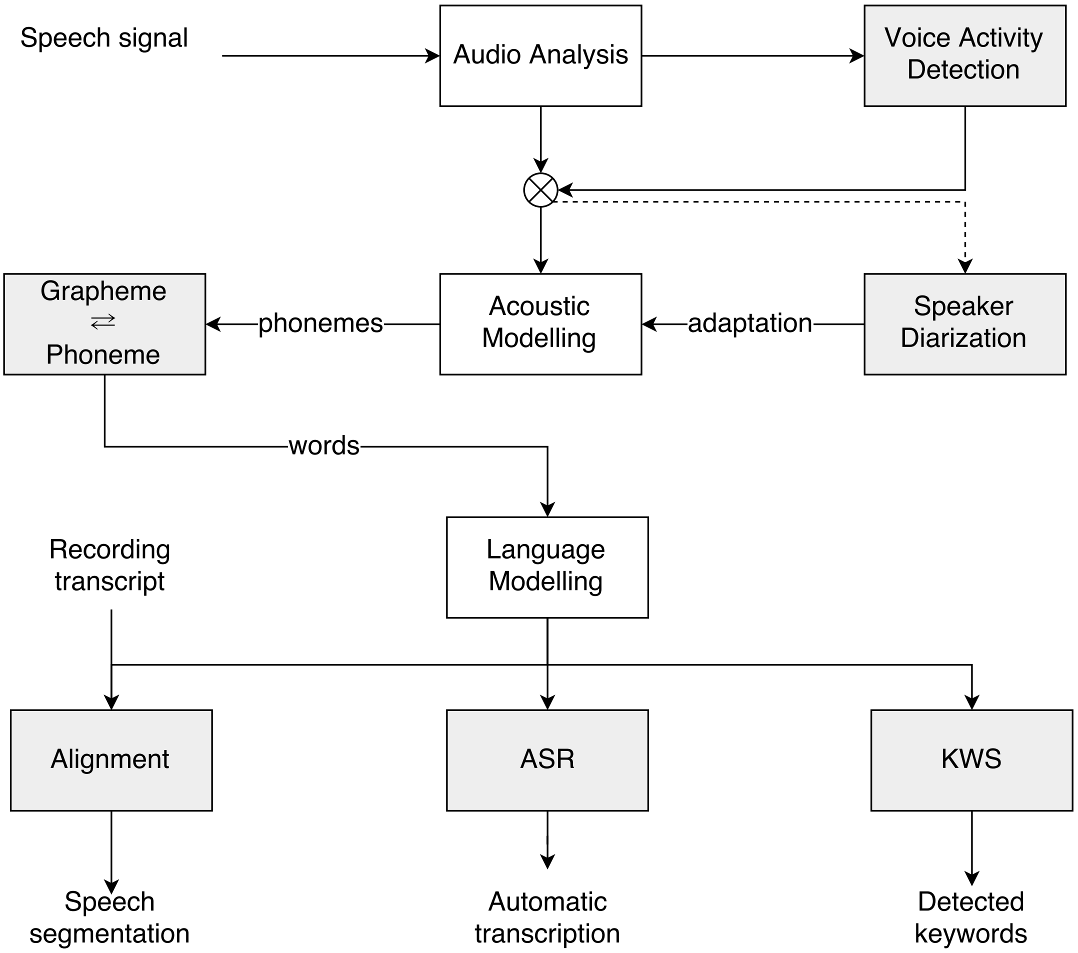


Figure 1: The pipeline of several speech processing mechanisms. The input signal on top is processed to produce the information on the bottom. The grey blocks were exposed as services in the CLARIN-PL infrastructure.

```
f S tS e b Z e S I ni e x S on S tS b Z m i ft S tsi i ni e
i S tS e b Z e S I n s t e g o s w I ni e v u w g o p I t a
p a ni e x S on S tS u p o ts u S p a n t a g b Z en tS I
v g on S tS u
```

Figure 2: An example of a transcription of the poem “Chrząszcz” by Jan Brzechwa, as produced by the grapheme-to-phoneme tool. Uses a slightly modified version of the Polish SAMPA phonetic alphabet.

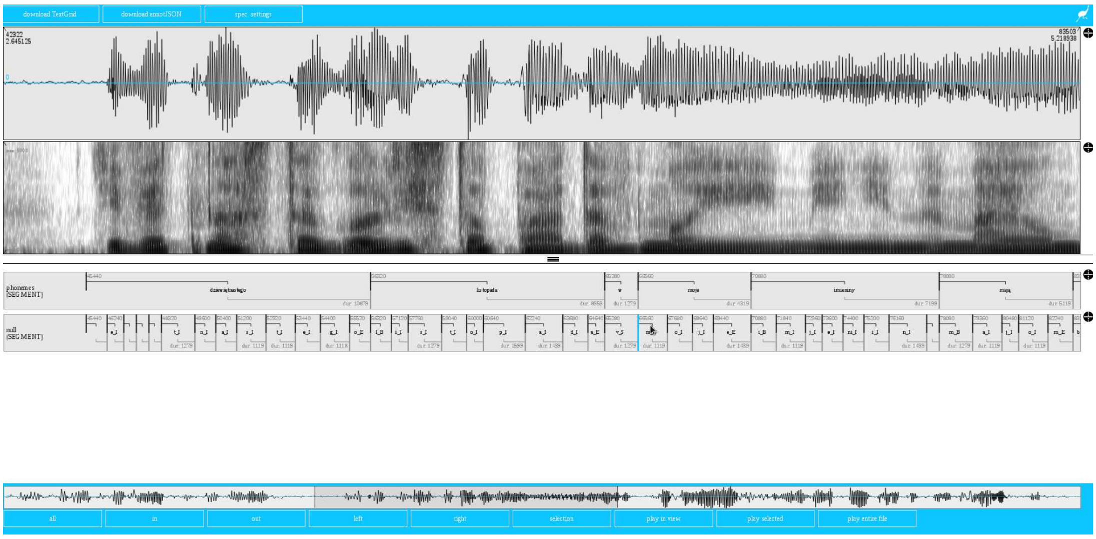


Figure 3: An example of a segmentation generated by the speech-to-text alignment service, displayed in the EMU-webApp interface.

For the future, several improvements are planned. A better acoustic model, possibly based on ANNs is going to be implemented. Adaptation of both the acoustic and the language model could also be beneficial to the overall process, especially when it comes to noisy data. The tool works fine for clean and predictable data, but it still produces errors or fails entirely for very noisy or otherwise low SNR signals.

Finally, a UI enhancement could be created to allow the user to manually invoke the re-alignment of a certain portion of the output, while also allowing to fix some of the information (like the orthographic or the phonetic transcription). This could turn the already very useful, but sometimes imperfect fully automatic tool, into a perfect semi-automatic tool.

2.3 Voice activity detection

Voice activity detection (VAD) is frequently found as a pre-processing step of many speech processing tools. Its purpose is to isolate portions of audio that contain speech from those that contain other types of acoustic events, like silence, noise or music. Apart from the aforementioned use as a pre-processing step, it can also be useful as an indexing tool for large quantities of audio. This tool is completely language and domain independent, although it may fail with very noisy data.

This tool was constructed using an Artificial Neural Network based classifier that performs VAD in an online manner. The non-speech data is further analyzed using an SVM classifier to try and classify types of noise. This last step was not developed very thoroughly and performs rather poorly, depending on the data, but given a proper use-case and training data, it could be modified to work better.

The VAD component was used extensively during previous projects and was already known to perform reasonably well for real-world data. A simple experiment confirmed a fairly high level of recall ($\sim$99%) with a not so good precision ($\sim$58%), which was a conscious decision (preferring not to lose any speech, while sometimes accepting non-speech falsely). The subsequent tools (ASR modules) can deal with a small amount of noise in the data, but would suffer greatly if any speech was omitted.

2.4 Speaker diarization

This tool is used to segment a large audio file into portions spoken by individual speakers. There are several types of speaker related segmentation strategies that can be performed: speaker change detection recognizes only the segments where different speakers are talking, speaker diarization additionally anno-

```
że 5.91 0.3 7228.28
że 20.21 0.35 5301.86
że 20.21 0.13 5266.03
że 1.11 0.13 4021.23
że 1.23 0.17 4014.55
że 0.79 0.12 3494.49
że 28.29 0.17 1822.69
że 16.6 0.08 0
listopada 7.43 0.58 3877.51
listopada 29.26 0.5 2541.87
polityki 11.27 0.63 7678.28
```

Figure 4: The output of the keyword spotting tool, searching for the words "że", "listopada" and "polityki" in an example recording. Each line contains one occurrence of a keyword with the following fields: word, start time, duration, keyword likelihood. Note that the short word "że" often occurrs as a part of longer words and is thus erroneously detected multiple times.

tates which segments belong to the same speaker and speaker identification recognizes exactly who the speaker talking in each segment is (e.g. by name). Our tool only does the second algorithm.

It is mostly useful for adaptation of various tools and models to individual speakers but some researchers have mentioned that they would like to use it for other types of analyses that require speaker segmentation. Our tool is based around the LIUM (Meignier and Merlin, 2010) toolkit and just like the previous one, it is completely language independent. Other toolkits were also tested during the project (for example SHoUT(Huijbregts, 2006)) but LIUM seemed to perform best on real-world data.

2.5 Keyword spotting

Often times, an accurate transcription of audio material is not necessary because we are only interested in individual words appearing in the text. Keyword spotting (KWS) is a process that takes an audio file with a list of keywords and generates a list of their occurrences with in the audio file.

Our system is based around the Kaldi toolkit, but it is also expanded to support an open vocabulary scenario. Given the limited language model vocabulary size, it would be impossible to predict all the words that people may be looking for. Therefore, our system uses a combination of words and syllables, so when a word out of vocabulary needs to be found, its syllable representation is used instead. This makes the tool sometimes more useful than full speech transcription, because it can deal with words that are out-of-vocabulary (OOV), but is more prone to errors when given very short keywords. If the word is phonetically a part of another word it may still be recognized as a separate word.

A small corpus was prepared to test this component and the overall precision was very high ($>\sim$95%) with the recall being reasonably high for known words ($\sim$82%) and low for words that were OOV ($\sim$20%). It seems that the syllable model worked well sometimes but still needs improvement to deal with OOVs.

2.6 Automatic speech transcription

This tool uses an Automatic Speech Recognition (ASR) system (based on the Kaldi toolkit) to generate a probable orthographic transliteration of audio recording of Polish speech. Initially, this tool was not planned for inclusion in the project but due to overwhelming interest, it was added in the second part of the project. The current system uses our Euronews model for recognizing broadcast news (Marasek et al., 2014) and in order for it to be useful for other types of recordings, it has to be adapted to the proper domain. More details on the developed architecture is given in section 3.1 below.

3 Speech Corpus

In order to produce most of the tools mentioned in the previous section, a large set of good quality recordings is required. This is usually expensive to produce and even if such data is available for purchase from third-parties, it is usually very expensive and unobtainable by most researchers. Prior to our work, there was no free, high quality, large-vocabulary audio corpus of Polish speech. Our goal was to create such a corpus and release it on an open license, both for commercial and non-commercial use.

The corpus was recorded in a studio environment using two microphones: a high-quality studio microphone and a typical consumer audio headset. The corpus consists of 317 speakers recorded in 554 sessions, where each session consists of 20 read sentences and 10 phonetically rich words. The size of the audio portion of the corpus amounts to around 56 hours, with transcriptions containing 356674 words from a vocabulary of size 46361. In addition to the studio corpus, a smaller corpus of telephony quality was also recorded. It contains 114 sessions, amounting to around 13 hours of recorded speech.

Both the studio and telephone quality corpora were released in two forms. The first one is the EMU database (Cassidy and Harrington, 2001), which allows for easy lookup of data and even some statistics thanks to the integration with the R platform. Unfortunately, the current version of the system relies on downloading the rather sizable corpus locally onto the computer. A new release of the corpus is planned using the more modern EMU Speech Database Management System (EMU-SDMS) that works in the web browser. This will make the corpus much more convenient, since it won't require downloading of all the data.

3.1 Baseline speech recognition system

Given that the main purpose of preparing the corpus was the development of speech tools, it seemed fit to deliver the corpus in a form that would make it easy to make such a tool. Since most of the tools mentioned in the previous chapter rely on the Kaldi speech recognition system, a baseline setup for developing such a tool was constructed. This setup was designed to replicate the approach used in the official project for other languages. The main idea is to have all the tools necessary in one folder, with one main script performing the training of the full system, from start to finish. If the user doesnâĂŹt wish to modify anything, they can simply run the one script and wait a few hours to get a fully-trained, fully-working large-vocabulary continuous speech recognition system. Users are, however, encouraged to read the comments in the script and figure what is actually being performed.

The corpus was split into a training and test portion, roughly 90% and 10% respectively: 56 random sessions were chosen to be in the test set, and the remaining 499 session were stored in the test set. A trigram statistical language model was trained using a large collection of texts in Polish collected from various online sources and interpolated on the transcriptions from the training portion of the corpus only. Since the source material for this portion of the setup cannot be distributed freely, the trained language models are provided as a download.

The acoustic models were trained using the standard GMM training procedure: first the monophone (mono) models were trained, followed by triphone (tri2a). The standard feature front-end is then replaced by an LDA multi-splice set (tri2b), followed by adapting the models to speakers (tri3b). Next, the silence models are are retrained and the phonetic dictionary is rescored (tri3b-sp). Then an experiment is performed with increasing the number of mixtures (tri3b-20k) and with using MMI training (tri3b-mmi). This last one is also repeated using a larger beam, followed by rescoring using a much larger language model, to give the final result of 7.37% word error rate (WER). If we use the lattice from the wide beam stage and instead of rescoring look for the best sequence of words (in other words, if we had the ideal language model), we can get a score as low as 3.23% WER.

In addition to the standard GMM acoustic models, two artificial neural network (ANN) based systems were also tested. The time-delay neural network (TDNN) system achieves a score significantly better than GMMs and the LSTM is even slightly better than that. The LSTM (being a recurrent ANN model) is however much slower to train and the marginal improvement in WER may not be worth it for most people.

WER %	experiment
30.06	mono
17.56	tri1
16.75	tri2a
15.75	tri2b
13.50	tri3b
13.10	tri3b-sp
12.88	tri3b-20k
12.41	tri3b-mmi
11.64	+wide beam
7.37	+large LM rescoring
3.23	oracle of wide beam

Table 1: GMM acoustic model results.

WER %	experiment
9.25	TDNN
5.91	+large LM rescoring
2.83	oracle
8.91	LSTM
5.78	+large LM rescoring
2.61	oracle

Table 2: ANN acoustic model results.

4 Applications

A couple of projects have already utilized our tools and resources for their own uses. Our speech alignment tool was used by a consortium partner in order to further annotate the corpora on their Spokes platform (Pezik, 2015). The studio speech corpus was used in a paper by a Czech research team (Nouza et al., 2015). We have also managed to cooperate with a team from the Institute of Applied Linguistics at the Warsaw University on their project titled “Respeaking - the process, competences and quality” (project code NCN - OPUS6 -2013/11/B/HS2/02762). Finally, one of the most interested groups were researchers of sociology interested in automatic transliteration of sociological interviews. We managed to receive several hours of recordings by a group of researchers from the The Cardinal Wyszyński University in Warsaw. Some preliminary results show promise but more work is needed to achieve success.

We intend to open several new areas of applications in the future. The new project will concentrate mostly around these three domains: parliamentary speeches, historical early and mid-20th century news segments and improved systems for the transliteration of sociological interviews.

5 Future plans

With the project being prolonged for the two more years, several improvements are planned. The main focus will be on creating working speech recognition solutions for the aforementioned domains. To achieve this, certain tools, like the G2P conversion including text normalization and possibly other modules, like speaker diarization and VAD, will have to be improved. The biggest improvements, however, will lie in the speech recognition engine, itself. Many experiments are planned, including various adaptation techniques, Deep Neural Network for acoustic modeling (Vu et al., 2014), Recurrent Neural Networks for language modeling (Mikolov et al., 2013).

No new corpora will be recorded, although lots of data will have to be collected, in order to adapt the tools to their respective domains. It is unclear whether all of the data will be released for other researchers, due to legal concerns. Our primary intention will be to improve the services available on our website and to provide the trained models and tools for free, for others to use as they deem necessary.

6 Acknowledgments

The activities described in this paper were funded by the Clarin-PL project. This project was partially supported by the infrastructure bought within the project “Heterogenous Computation Cloud” funded by the Regional Operational Programme of Mazovia Voivodeship.

References

[Bisani and Ney2008] Maximilian Bisani and Hermann Ney. 2008. Joint-sequence models for grapheme-to-phoneme conversion. *Speech Communication*, 50(5):434–451.

[Boersma and others2002] Paulus Petrus Gerardus Boersma et al. 2002. Praat, a system for doing phonetics by computer. *Glot international*, 5.

[Brocki et al.2012] Łukasz Brocki, Krzysztof Marasek, and Danijel Koržinek. 2012. Multiple model text normalization for the polish language. In *International Symposium on Methodologies for Intelligent Systems*, pages 143–148. Springer.

[Cassidy and Harrington2001] Steve Cassidy and Jonathan Harrington. 2001. Multi-level annotation in the emu speech database management system. *Speech Communication*, 33(1):61–77.

[CLARIN-NL2013] CLARIN-NL. 2013. Ttnww - tst tools voor het nederlands als webservices in een workflow. `https://portal.clarin.nl/node/1964`. [Online; accessed 2016-09-27].

[Huijbregts2006] Marijn Huijbregts. 2006. Shout speech recognition toolkit.

[Katsamanis et al.2011] Athanasios Katsamanis, Matthew Black, Panayiotis G Georgiou, Louis Goldstein, and S Narayanan. 2011. Sailalign: Robust long speech-text alignment. In *Proc. of Workshop on New Tools and Methods for Very-Large Scale Phonetics Research*.

[Kisler et al.2016] Thomas Kisler, Uwe Reichel, Florian Schiel, Christoph Draxler, Bernhard Jackl, and Nina PÃűrner. 2016. Bas speech science web services - an update of current developments. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

[Lenkiewicz et al.2012] Przemyslaw Lenkiewicz, Eric Auer, Oliver Schreer, Stefano Masneri, Daniel Schneider, and Sebastian Tschöpe. 2012. AvatechâĂŢautomated annotation through audio and video analysis. In *LREC 2012: 8th International Conference on Language Resources and Evaluation*, pages 209–214. European Language Resources Association.

[Marasek et al.2014] Krzysztof Marasek, Krzysztof Wołk, Danijel Koržinek, Łukasz Brocki, and Ryszard Gubrynowicz. 2014. Spoken language translation for polish. *Forum Acousticum*.

[Meignier and Merlin2010] Sylvain Meignier and Teva Merlin. 2010. Lium spkdiarization: an open source toolkit for diarization. In *CMU SPUD Workshop*, volume 2010.

[Mikolov et al.2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

[Nouza et al.2015] Jan Nouza, Petr Cerva, and Radek Safarik. 2015. Cross-lingual adaptation of broadcast transcription system to polish language using public data sources. In *7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, Poland*, pages 181–185.

[Pezik2015] Piotr Pezik. 2015. Spokes-a search and exploration service for conversational corpus data. In *Selected Papers from the CLARIN 2014 Conference, October 24-25, 2014, Soesterberg, The Netherlands*, number 116 in Linköping Electronic Conference Proceedings, pages 99–109. Linköping University Electronic Press.

[Povey et al.2011] 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.

[Vu et al.2014] Ngoc Thang Vu, David Imseng, Daniel Povey, Petr Motlicek, Tanja Schultz, and Hervé Bourlard. 2014. Multilingual deep neural network based acoustic modeling for rapid language adaptation. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7639–7643. IEEE.

[Winkelmann and Raess2014] Raphael Winkelmann and Georg Raess. 2014. Introducing a web application for labeling, visualizing speech and correcting derived speech signals. In *LREC*, pages 4129–4133.

Discovering Resources in the VLO: A Pilot Study with Students of Translation Studies

Vesna Lušicky
Centre for Translation Studies
University of Vienna, Austria
vesna.lusicky@univie.ac.at

Tanja Wissik
Austrian Academy of Sciences, Austria
and
University of Graz, Austria
tanja.wissik@oeaw.ac.at

Abstract

Common Language Resources and Technology Infrastructure (CLARIN) provides access to language resources for scholars in the humanities and social sciences. In theory, scholars and students of Translation Studies may be assumed to be active data providers of language resources, as well as prolific users of the CLARIN services. However, data show that the uptake of CLARIN services by this user group is rather low. This paper investigates the needs of the students of Translation Studies and evaluates the CLARIN Virtual Language Observatory (VLO) from their perspective. It is based on a pilot study applying open and closed situated user assignments and an evaluation of the VLO service. The results provide insights into the needs of this user group and give suggestions to data and service providers that could increase the adoption of CLARIN services by the user group.

1 Introduction

E-research has transformed the process of research and has become a more ubiquitous research practice. CLARIN (Common Language Resources and Technology Infrastructure) aims at providing sustainable access for researchers in the humanities and social sciences to digital language data and tools. As observed in other service-oriented e-research infrastructures (Chunpir et al., 2015), the phase of development, setting-up and running of the CLARIN infrastructure and services was followed by the requirement to conduct studies into user involvement and user experience. The survey of user involvement in CLARIN was presented (Wynne, 2015) at the CLARIN Annual Conference 2015, showing user activities by discipline. Rather surprisingly, Translation Studies were not listed among the disciplines[1] (Wynne, 2015) of the users involved in CLARIN services.

At least some branches of Translation Studies, especially Corpus-based Translations Studies (Baker, 1993, 1995; Laviosa, 2002; Fantinuoli and Zanettin, 2015)[2] and Computational Translation Studies[3], are carried out with digital methods and tools They not only heavily rely on various languages resources, e.g. parallel and comparable corpora, translation memories, terminology resources, and lexica for research purposes, but they also generate both mono-, bi- and multilingual language resources (Budin, 2015). Language resources are also extensively used and generated by translation practitioners and by

[1] Probably included in the category *Other humanities*.

[2] Corpus-based theoretical and descriptive research in Translation Studies has investigated topics such as translation universals and norms, ideology and individual translator style, and corpus-based tools and methods are included in the curricula of translation training institutions (see Fantinuoli and Zanettin, 2015).

[3] In this paper, Computational Translation Studies are understood as a paradigm and research methodology that takes place at two levels as proposed by Budin (2015): Translations Studies carried out with computational methods and Translation Studies investigating computational processes, for example from a sociological or cognitive perspective (for instance to optimize the human-computer-interaction in translation workflows).

Vesna Lušicky and Tanja Wissik 2017. Discovering resources in the VLO: A pilot study with students of translation studies. *Selected papers from the CLARIN Annual Conference 2016.* Linköping Electronic Conference Proceedings 136: 63–75.

trainers and students in translation training. For these reasons, the absence of documented users from the field of Translation Studies in the study mentioned above appears noteworthy and calls for further investigation.

The specific needs and requirements of this user group that could contribute to a higher uptake of the CLARIN services, in particular Virtual Language Observatory (VLO), by this user group are being investigated in this study. In no way suggesting that the study's results can immediately be generalised, as at this stage the investigation is a pilot study of real world research (Robson and McCartan, 2016) and didactic in action, this paper sets out to show the specific needs and requirements of a specific user group, and to give impulses for studies into user experience of the CLARIN service VLO.

The paper is structured as follows: First, we briefly present selected considerations that were vital for the design of this pilot study in section 2. In section 3, we elaborate on the actual pilot study and present the objectives of the study, the data collection process and settings, and participants. We also illustrate the task design in detail. This is followed by a presentation of results related to each task, the interpretation of results, and discussion. We conclude by summing up the key findings and give an outlook for potential further research.

2 Background

2.1 Language resources in Translation Studies and translation practice

The production, compilation, use, and re-use of various language resources, such as mono- and multilingual corpora, translation memories, terminology resources, and lexica is well-documented and explored in settings related to translation: in translation practice (*inter alia* Beeby et al., 2009; Bowker, 1998, 2002 and 2004; Gallego-Hernández, 2012; Wilkinson, 2005), in translation training (*inter alia* Kenny, 2007; Krüger, 2012, Kübler, 2003; Maia, 2003) and directly in the research in Translation Studies (*inter alia* Baker, 1993, 1995; Granger, 2003, Fantinuoli and Zanettin, 2015).

Computer-assisted translations tools (CAT tools) are used by 73 per cent of translation practitioners (Ehrensberger-Dow et al., 2016). This implies that these practitioners are also highly active in generating and reusing a particular type of language resources, namely translation memories, which are "a very specialised kind of parallel corpus, and are usually relevant, reliable and well integrated into the translation workflow. Of course, translators do not have a translation memory ready for all occasions" (Zanettin, 2012:247). In such cases, translation practitioners either build their ad hoc corpora or look for stable corpora (Sánchez-Gijón, 2003) and other language resources, such as terminological resources and lexical resources. Curated one-stop entry-points where these types of datasets could be found via search activities, addressing specific needs of translation practitioners and scholars, could, therefore, find a good acceptance by this user group.

2.2 Repositories and catalogues

Since creating digital resources from scratch is often time-consuming and expensive, re-use of the existing data and resources is recommended. To re-use the existing resources, researchers, and also other potential users, have to be aware of the existence of suitable resources and need "efficient ways to navigate to the language resources that matter, whatever the selection criterion is" (Van Uytvanck et al., 2012). Various portals, repositories, and catalogues that originate in various e-research projects and initiatives could also provide entry points to the datasets usable in the scope of Translation Studies. Among them are rather general repositories and catalogues that cater to diverse user groups, e.g. ELRA catalogue and META-SHARE, and catalogues curated for specific sub-types of tasks – as language resources originally created with a specific purpose are not always generally reusable – and user groups, such as commercial users in the field of machine translation in the case of the LT-Observe catalogue (Maegaard et al., 2016).

VLO is one component of the CLARIN research infrastructure that falls into the former group, addressing a broad range of researchers in the humanities and social sciences. It is a metadata-based portal for language resources, providing multiple views on metadata for linguistic data and software and trying to give a consistent online overview of the data that is available in a variety of CLARIN Centres (Van Uytvanck et al., 2010; Van Uytvanck et al., 2012). The VLO offers faceted search (language, subject, collection, format, resource type, organisation, continent, national project, country, keyword, modality, data provider, genre) and string search (Odijk, 2014).

2.3 Approaches to search activities

Search activity, i.e. search for new or additional data, whether due to a pertinent practical need or in the scope of the first stage of scholarly research (Kemman et al., 2014) is often the first entry point and contacts of users with the services of an infrastructure, and may have an impact on their further engagement with the infrastructure, and willingness to contribute resources, etc. Optimal alignment of the entry points of repositories and catalogues to user requirements also implicate that approaches to search activities and specific user group competences, and requirements are understood.

From the perspective of information retrieval, search activities are commonly divided into two broad categories: lookup and exploratory (Marchionini, 2006), although lookup tasks can be embedded in exploratory tasks and vice versa. Lookup search is assumed to have precise search goals, such as finding facts to answer a specific question. Exploratory search includes a variety of qualitative definitions, so this search is naturally faceted (Wildemuth and Freund, 2012), and has open-ended search goals and inexact task requirements. This approach to search activities corresponds to the functionalities of the VLO and seems to be suited as an underlying premise when designing tasks to investigate user search activities and requirements of the service.

Another factor that should be considered is the competence of users to identify information requirements, find relevant information, and evaluate the search results. Expertise plays a significant role in the approaches that users utilise to seek information. The distinction should be made between technical and domain expertise. Jenkins et al. (2003) observe that whereas each of the dimensions is valuable, users are most likely to succeed in search activities when both are present. These factors should be considered when developing tasks for a user-centred study or investigation of user requirements. Students, i.e. participants in this pilot study, are often classified as *experts-in-training* (Hagemann, 2016).

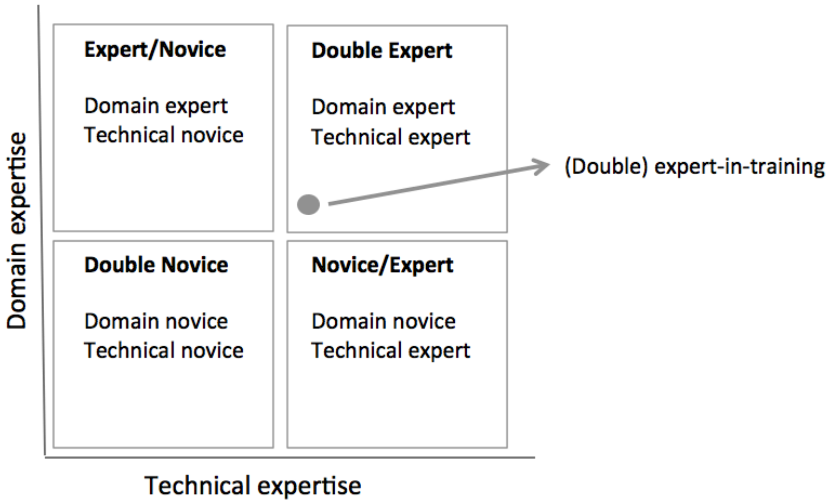


Figure 1: Two dimensions of expertise: domain and technical (adapted from Russel-Rose and Tate, 2013).

2.4 Users' search competences

Present day translation training is informed by both academic research and professional experience. The training incorporates situated learning (Risku, 2016), by which it emulates the actual translation practice through the use of authentic resources, tools, assignments and processes relevant for translators. These curricula often follow one of the most exhaustive translation competence models, that of the European Master in Translation framework[4], which covers six main competences (see Fig. 2), among others the

[4] The network aims at promoting quality in translation training and is led by the Directorate General for Translation of the European Commission. The label is awarded only to academic translation programmes meeting admission criteria (currently 64 programmes). However, the competences developed in the framework are widely integrated also in curricula of the programmes that had not been awarded the official label. For more information, see https://ec.europa.eu/info/european-masters-translation-emt_en (accessed 1.4.2017).

information mining competence, and the technological competence. Information mining competence includes, inter alia:

- identifying information and documentation requirements,
- extracting and processing information for a given task,
- evaluating the reliability of sources,
- effectively using search engines (EMT Expert Group, 2009).

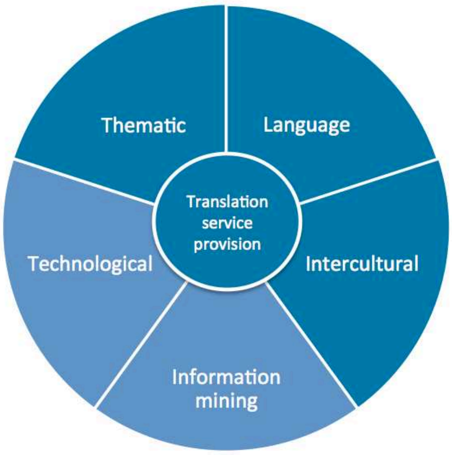


Figure 2: Competences in present translation training: Wheel of competence of the European Master of Translation (based on EMT Expert Group, 2009).

Although e-research practices (such as the role of digital curation practices) are not explicitly integrated into present day academic translator curricula, there is a substantial overlap between information mining and technological competences, and research competences, therefore translation courses can be expected to train some research competences indirectly (Vandepitte, 2013; Austermühl 2016). As translation training caters to both prospective translators-practitioners and researchers in Translation Studies, students were treated as *double experts-in-training* and investigated as users in this pilot study.

3 Pilot study with students of Translation Studies

3.1 Objectives

As discussed above, scholars and students of Translation Studies use language resources in their research and practical work. For this reason, the objective of the pilot study was to investigate which selection criteria matter to users in Translation Studies when they try to navigate through a research infrastructure to the language resources that they need. Secondly, we wanted to test one of the services of the CLARIN infrastructure, namely the VLO faceted browser, as it offers the exploratory functionalities for language resources (Odijk, 2014), and to find out the perceived quality of the results in the VLO by this user group. Lastly, the objective of this pilot study was to determine how students as prospective translators and researchers in Translation Studies would engage with the service and what is needed to ensure a higher uptake of the service by this user group.

3.2 Data collection setting and participating users

Data collection for this study was carried out at two Austrian universities in five courses with students of Translation Studies at the BA and MA levels during two academic years (winter semester 2015, summer semester 2016[5], and winter semester 2016[6]). The course at the BA level is recommended for students in their 5th semester. Therefore the majority of the students participating in the study were in their final year of the BA studies. The majority of the students at the MA level participating in this study

[5] References to the VLO refer to the version 3.3.

[6] References to the VLO refer to the version 4.0.2.

was in their second semester, or at even in more advanced stages of their MA studies. The courses that served as the platform for this study are obligatory for all students with all working language combinations in the program. This means that we had unique access to users with a broad range of working language combinations and could, therefore, cover a spectrum of languages with various degrees of support (Rehm and Uszkoreit, 2012). The participants' working languages ranged from rather traditional combinations (e.g. German, English, French, Spanish, Polish, Italian) to Arabic and Austrian Sign Language. Details on participants' working languages are given in the next section.

All five courses focused on practical aspects of specialised translation, emulating the actual practice through the use of authentic resources, tools, assignments and processes. The format of the face-to-face sessions was designed to cover selected topics on language resources, and an introduction to e-infrastructures, repositories, catalogues and similar services. None of the participants had previous experience with searching for language resources in repositories or catalogues.

In the study design, the length of time of the study and each task needed particular attention. This was a critical issue because the participants were performing activities that took some length of time to complete. Moreover, search tasks can be exhausting both mentally and physically (Kelly, 2009), and prolonged search activities may negatively impact the results. For these reasons, as well as due to time limitations (length of an academic session, availability of students limited to one semester) and logistical limitations (number of available computer working stations), the data collection was conducted in several phases, and with a different setup and number of participating users. Each phase focused on one specific task as elaborated in the next section.

Each phase of the study was initiated by a face-to-face information session, in which the overall objectives and logistical details of the study were presented to the students by the course trainers. A written task description was also presented to participants to read until they understood it thoroughly. Participants were given time to ask questions and to opt out of the study. The observance of the anonymity of the data collected was explained and stressed. To keep the search process natural, we did not ask the participants to think aloud, but collected self-reported research results and comments as described in detail in the section below.

3.3 Task design

Task design utilised qualitative and quantitative approaches combining mainly open and to a limited extent closed tasks to operationalize exploratory search activities by double experts. This method allowed the findings to be identified both through pre-formulated research questions and through the formulation of newly raised topics of interest that had not been anticipated during the planning phase of the study.

We merged tasks into three task groups (one task group per study phase), two of them with two subtasks each to address specific research questions separately (see Tab. 1). We instructed the participants to inform us when they had completed each task; however, each task had a maximal time limit for completion.

Selection criteria can be understood as usability, i.e. a set of criteria that facilitate human decision-making (Maegaard et al. 2016). Based on this set of criteria, users conduct their exploratory search for language resources. As there is a little insight into the selection criteria for language resources preferred by the investigated user group, the first task (T1) aimed at establishing the set of selection criteria, based on which the users may decide if a language resource is relevant and operationally reusable for their purpose. The participants (n_0=25) were given an open assignment to identify the criteria without linking it to a specific service or e-infrastructure to ensure a minimal bias towards their selection of criteria. The users were asked to provide a weighted list of their selection criteria.

The tasks T2a and T2b were based on the criteria identified in T1. The objective of T2a was to establish how many unique language resources relevant for this user group could be found in the CLARIN VLO that cannot be found in other repositories and catalogues. In this task, the participants (n_1=25) were asked to query portals (CLARIN VLO, META-SHARE, and ELRA), and catalogues (LT-Observe, Opus) for language resources in their working languages (Croatian, Czech, English, French, German, Italian, Polish, Romanian, Russian) based on preselected criteria, identify them and document their research results in a form. The participants were asked to search for all relevant types of languages resources: mono-, bi-, and multilingual corpora, parallel and comparable corpora, translation memories, terminological resources, lexical resources, etc. This task also included self-reporting the information where the language resource had been found, and where the metadata had been extracted from. The participants were instructed not to valorise the metadata by cross-pollinating them from various sources or improving them by additional research, but to assign a score from 1 to 5 (1 for very high quality, 5 for very low quality) to the perceived quality of the metadata provided for each resource found in different portals and catalogues in task T2b (What is the perceived quality of the metadata of the LRs found in the VLO?).

Phase	Task	Task objective	Type of task	Data collected
Definition of selection criteria	T1	When searching for LRs, which criteria are essential for the user group?	Open	Criteria sets with descriptors
Identification of language resources based on selection criteria	T2a	How many (unique) LRs relevant for this user group can be found in VLO?	Semi-open (task completion based on pre-defined criteria)	Number of resources; list of identified resources with descriptors of criteria based on found metadata records
	T2b	What is the perceived quality of the metadata of the LRs found in VLO?	Closed	Values 1-5 (1 for very high quality, 5 for very low quality)[7]
Discovering language resources in VLO	T3a	When searching for multilingual LRs in VLO, what metadata is missing according to the user group?	Open	Metadata with descriptors
	T3b	How does the user group perceive the interface and the functionalities of the VLO while searching for (multilingual) LRs?	Open	Written self-reports

Table 1: Overview of the task design with task objectives, type of task, pre-task activities, and data collected in each task.

The third task group (T3a, T3b) solely concentrated on the CLARIN VLO. The participating users were briefed on the aims and objectives of CLARIN, the main principles and search functionalities of the VLO (e.g. faceted browsing, textual queries, and advanced querying), and were given a demonstration. The third task group was conducted twice as the task was first carried out with the VLO refer to the version 3.3, which became obsolete less than a month later, and was replaced with a more sophisticated upgrade. Therefore, we decided to run the tasks T3a (When searching for multilingual LRs in the VLO, what metadata is missing according to the user group?) and T3b (How does the user group perceive the interface and the functionalities of the VLO while searching for (multilingual) LRs?) again in the VLO version 4.0.2. The participants (n_2=14) that performed the task in the VLO version 3.3 worked with the following languages: Arabic, Bosnian/Croatian/Serbian, English, French, German, Italian, Spanish, Russian, and Sign Language. The participants (n_3=21) that performed the task in the VLO version 4.0.2 worked with Bosnian/Croatian/Serbian, English, French, German, Italian, Polish, Spanish, Russian, and Welsh.

Both times the participants were asked to run a search for multilingual language resources in the VLO. They were given basic preselected search criteria to yield comparable results. In task 3a (When searching for multilingual LRs in the VLO, what metadata is missing according to the user group?) they were asked to list further categories of metadata that could be useful from the perspective of Translation Studies to be included as "[…] one of the main purposes of metadata is to enable discovery of a resource"

[7] The inverted score scale was adopted for practical reasons: The participants were already familiar with this scoring system as it mirrored the academic marks.

(Odijk, 2014). In addition, they were asked to provide a comment on the satisfaction with the search functionalities (T3b).

4 Results and discussion

In this section we present the results of all the tasks described in section 3, followed by a brief discussion.

4.1 Selection criteria as a basis for explanatory search in repositories

As the investigated users often need very specific types of language resources, e.g. when they work in a particular language combination or a particular domain, they need text and terminology from the domain, and language resources exactly in the languages in question, etc. The results of the task T1 thus unsurprisingly showed that the users highly valued the information on the *Language(s) covered* by the resource, but also the criterion *Representativeness of the domain* (see Fig. 3).

Related to this criterion are also two further criteria: the *Reliability of the resource* (for explanation see below) and *Up-to-dateness* of the resource. Participants described the criterion *Up-to-dateness* as the fact that a resource captures the latest data relevant to the domain, language, etc. This is deemed essential in specific domains (e.g.) legal, or after a spelling reform.

As certain language resources are preferably used directly in a computer-assisted translation tool, *Format* (e.g. tmx) was the third most identified criterion. A separate criterion *Downloadable* might appear redundant. However, it should be taken into account that the criteria should not be evaluated separately, but as a set of criteria to reflect the complexity of exploratory search activities.

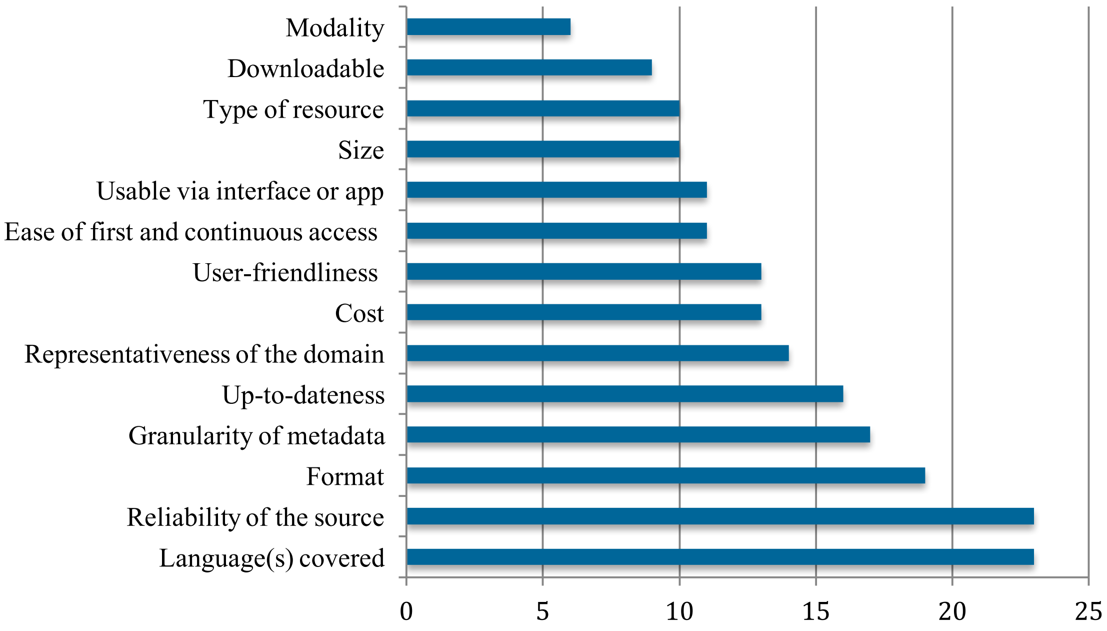


Figure 3: Selection criteria for language resources from the perspective of Translation Studies students (n=25) (results related to task T1).

It should also be emphasised that the selection criteria do not directly correspond to metadata describing a language resource. Metadata records are interpreted by the user and checked against the set of criteria to facilitate the decision-making and if needed, abort or continue the exploratory search. For example, reliability can be understood as the degree of authoritativeness of the originator and is perceived as an important indicator of the quality of the resource, especially for certain specialised translation assignments in legal, administrative, but also technical domains. In the absence of this metadata category, reliability could be derived from the combination of metadata records.

4.2 Relevant language resources and metadata

Based on the criteria defined in task T1, the participants queried repositories and catalogues (task T2a), which resulted in identifying of 210 relevant resources[8] in total, found in all repositories and catalogues (CLARIN, META-SHARE, ELRA, LT-Observe). The raw dataset included multiple identifications of the same language resource as more than one language was queried in this task. Some multilingual resources were identified several times, for example, the resource *Europarl Parallel Corpus* fulfilled the criteria of the search for Spanish as well as for English. A subsequent clean up of the doublets covering several languages (e.g. Europarl Parallel Corpus, FAO Glossary of Biotechnology for Food and Agriculture, JRC-Acquis Multilingual Parallel Corpus, and others) was needed and removed 33 per cent of entries from the raw dataset. A comparison of all the identified relevant language resources in the cleaned up dataset showed that the majority of the identified resources could be found in more than one repository or catalogue. Ten language resources were uniquely found through CLARIN VLO, all but one of them being text corpora. The participants identified a wide variety of resources: monolingual, bilingual and multilingual corpora, translation memories, terminological resources, and lexical resources.

The median of the perceived quality of the metadata (1-5, 1 for very high quality, 5 for very low quality) for the resources uniquely found through the VLO (task T2b) was 3. Overall, participants repeatedly assigned high quality scores for the quality of the metadata found in the catalogue LT-Observe[9]. It should be noted that this is a catalogue that consists of identified language resources for machine translation scenarios, based on evidence-based usability. The metadata in the catalogue has been validated and valorised by human experts (Maegaard, 2016).

4.3 Discovering language resources through CLARIN VLO and user requirements

The focus of the last task group was exclusively on the VLO, and bi- and multilingual resources (e.g. comparable corpora, parallel corpora, translation memories, lexical resources, terminological resources) found through the VLO. The answers of the participants were abstracted and grouped into categories (see Fig. 4 and Fig. 5)[10].

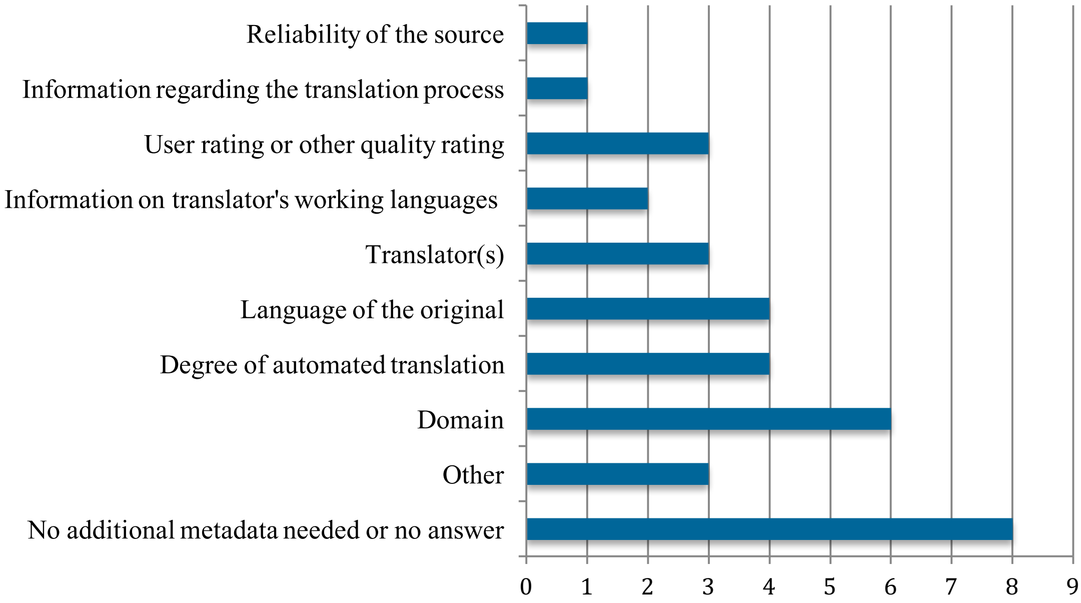


Figure 4: Desirable additional metadata for bi- and multilingual language resources from the perspective of Translation Studies students (n=35) (results related to task T3a).

[8] Multiple selections in more than one language were possible.

[9] http://www.lt-innovate.org/lt-observe/resources-list

[10] Multiple answers were possible.

The most requested additional metadata information voiced by participants was *Domain* (see Fig. 4). The participants also listed additional metadata that they considered useful in relation to bi- and multilingual resources: for example, in parallel data the information what is the original text and what is the translation. Further requested information was the degree of automated translation, i.e. whether the resource had been translated by a human translator, by a human translator utilising computer-assisted translation tools, or by a machine translation system. In addition to this, the information regarding the translation process could be useful, e.g. whether the output of automated translation was post-edited, etc. Some participants would also find the information about the translator(s) (e.g. name, details about the translator's working languages) useful. This metadata might not be applicable for all types of resources since we concentrated solely on bi- and multilingual resources in this task.

Regarding the search functionalities, the majority of participants were satisfied with the given search functionalities. Considering the importance of the selection criteria *Language* and due to the fact that the investigations in Translation Studies are usually conducted contrastively between two specific languages, the participants were not satisfied with the fact that it was not possible to search for language combinations (Fig. 5). The participants suggested that a multi-value selection for a single facet (e.g. language) would be very desirable. Regarding functional desiderata, the users suggested that functionalities that would make storing of search queries and search results possible. Furthermore, localisation of the search interface into different languages would be useful.

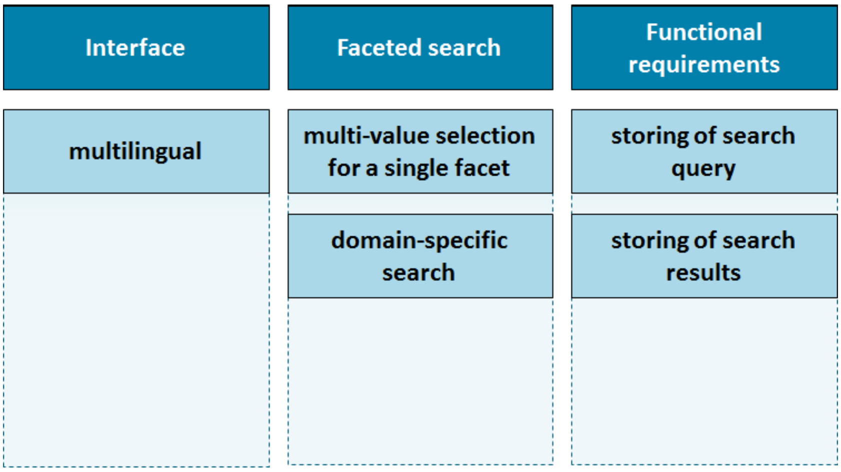


Figure 5: Desirable functional and search requirements, and interface features from the perspective of Translation Studies students (results related to task T3b).

The repetition of the task T3b (How does the user group perceive the interface and the functionalities of the VLO while searching for (multilingual) LRs?) in the VLO version 4.0.2 generated almost identical results. Therefore, the results of the repetition of this task support the initial user requirements established in VLO version 3.3 towards the interface, faceted search, and functional requirements.

4.4 Discussion

What can we learn from the obtained user search and selection criteria, users' performance in finding relevant language resources in the CLARIN VLO, as well as their comments on the search functionalities and requirements? To a large extent, the metadata fields provided in the CLARIN VLO (language, type, genre, licence, time coverage, etc.) help and support the users to decide whether the resource is relevant and useful for their task. Since relevant language resources seem to be also explored through other portals and services, the users might prefer them over CLARIN if they establish that they suit better their user requirements. This may result in a lower user involvement and may also have an effect on the involvement of this user group as data providers.

The information about user requirements of this user groups, specifically about desiderata in terms of metadata, could be of interest for providers of bi- and multilingual language resources, especially intrinsically translation-related language resources, such as translation memories, and would allow them to fine-tune the metadata provided for a better searchability in VLO, and thus a higher re-use of the resource.

Based on the pilot study with the students of Translation Studies, we derived to the following desiderata of the features in the VLO:

- Multi-value selection within a facet: Multi-value selection for a single facet (e.g. language) could support a better user experience for those users, who want to narrow down their search to a certain language pair. Users in Translation Studies are often interested in a contrastive comparison of language resources in a specific language combination.
- Domain-specific search: User interested in specialised translation and languages for special (specific) purposes, often need language resources from a particular domain, for instance, legal or medical domain. The VLO faceted search currently supports the search by subject. The granularity and depth of facet *Subject* vary considerably (e.g. legal documents vs. the legal prescriptions concerning hunting). As observed in this study, the search does not support users unfamiliar with the topics covered, who want to explore the resources intuitively by domain, or by classification system. Clustering of subjects into domains would encourage a more explorative approach to the discovery of resources in the VLO.
- Multilingual (localised) interface: Users searching for language resources in languages other than English may not have a strong command of English, or may simply be inclined to search in the language(s) of the potential language resource. This seems to occur more often when searching for a multilingual language resource (e.g. Italian-German terminological resource) in languages other than English.
- Multilingual (localised) metadata: Multilingual metadata in languages other than English would support and complement the multilingual search interface. Multilingual metadata would be presumably needed for those language resources, in which English is not one of the languages of the resource. Reusing the data by national projects and national consortia, as well as automatization of the process, could support this endeavour.
- Storing of search queries or search preferences: This feature could support repetitive search sessions, e.g. for comparative or didactic purposes, but would require a user account.[11]
- Storing of search results: This feature could support a repetitive or on-going discovery of language resource in the VLO but would require a user account.[12]
- Valorisation of metadata: Awareness for the multifaceted needs of various user groups should be raised among data providers. For instance, the information what is the original and what is the translation would be valuable.

Since the completion of the pilot study, new features have been introduced in the VLO version 4.1.0[13]. It is now possible to select multiple values within a facet, thus broadening the selection with the operator *OR*, or narrowing down the selection with the operator *AND*. Broadening the selection with the operator *OR* is the default setting of the multiple value selection behaviour[14].

As deduced from the pilot study, the new feature will greatly improve the usability for users searching for language resources in a certain language pair or language combination. From the perspective of the users in Translation Studies, the search could be optimised by the default operator *AND*, as users typically want to narrow down their search to a specific language combination. Alternatively, the preference for the default setting could be stored in search queries or search query preferences as suggested above. We would also suggest making the advanced search feature more prominent (e.g. position it on the top of the interface) as it has a major impact on the user experience and the users' satisfaction with the search functionalities.

[11] Regarding sharing of the search query, the current version of the VLO 4.1.0. supports the following: bookmarking the search query, copying the link and sharing the link via email.

[12] Regarding sharing of the search results, the current version of the VLO 4.1.0. supports the following: bookmarking the search query, copying the link and sharing the link via email.

[13] https://vlo.clarin.eu/about (accessed 8.4.2017).

[14] https://vlo.clarin.eu/help (accessed 8.4.2017).

In addition to the suggested features for the VLO outlined above, outreach activities designed for users in Translation Studies (students, researchers, trainers, and practitioners) would help to make the VLO service and other CLARIN services more known in the community. A wider recognition of the CLARIN services among the Translation Studies community could result in bringing a new user group on board as well gaining active providers of language resources.

5 Conclusion and outlook

This pilot study addressed the needs of the students of Translation Studies as prospective translators and researchers in Translation Studies, focusing primarily on Corpus-based Translation Studies and Computational Translation Studies, as one of the user groups of the CLARIN service VLO. Their assessment of the gaps in terms of the usability of the service was investigated, and suggestions were made for possible optimisation. It was established that the resources found through the VLO would need some additional metadata information, especially bi- and multilingual language resources, in order to be better suited for reuse by researchers, trainers, and students in Translation Studies. Although the metadata that is not generated by the data provider cannot be added to the VLO by third parties, awareness for the multifaceted needs of various user groups should be raised among data providers. This especially applies in cases, in which the resources provided had been generated by translators, translation scholars and translation students, to ensure a higher uptake of the VLO service as well as other CLARIN services by this user group. Moreover, outreach activities tailored for users in Translation Studies would help to make the VLO service and other CLARIN services more known in the community and would help to gain a user group as well as a data provider group.

Due to the specific nature of the modern translation training, which emulates the actual translation practice, and covers a wide array of competences, the present pilot study could be a starting point for further research on the specific needs of the users from Translation Studies of the CLARIN services. Dissemination activities targeting translation scholars, students, trainers, and translators would increase the visibility and the uptake of the CLARIN services by these user groups.

The study also discussed and implemented considerations and preparation that should be taken into account when designing tasks for exploratory search in the VLO or similar services. User requirements of similarly under-represented user groups could take this pilot study as their departing point. In addition to task completion and self-reporting, further avenues of user search behavior could be explored to arrive at precise and complementary data, such as task completion time, query time, query length, scroll depth, cumulative click, etc. to better understand users' search activities and adapt the CLARIN services to evidence-based user requirements.

Acknowledgements

This work crucially uses data, tools, and services made available through the CLARIN infrastructure.

References

[Austermühl2016] Frank Austermühl. 2016. Recherche und Arbeitsmittel. In Mira Kadric and Klaus Kaindl (eds.), *Berufsziel Übersetzen und Dolmetschen: Grundlagen, Ausbildung, Arbeitsfelder.* Tübingen: Gunter Narr, 200–217.

[Baker1993] Mona Baker. 1993. Corpus Linguistics and Translation Studies: Implications and Applications. In Mona Baker, Gill Francis and Elena Tognini-Bonelli (eds), *Text and Technology: In Honour of John Sinclair*. Amsterdam, John Benjamins, 233–250.

[Baker1995] Mona Baker. 1995. Corpora in Translation Studies: An Overview and Suggestions for Future Research. *Target,* 72, 223–244.

[Beeby et al2009] Allison Beeby, Inés Patricia Rodríguez, and Pilar Sánchez-Gijón (eds.). 2009. *Corpus Use and Translating: Corpus Use for Learning to Translate and Learning Corpus Use to Translate*. John Benjamins Publishing.

[Bowker1998] Lyanne Bowker. 1998. Using Specialized Monolingual Native-Language Corpora as a Translation Resource: A Pilot Study. *Meta*, 43.4, 631–651.

[Bowker2002] Lyanne Bowker and Jennifer Pearson. 2002. *Working with Specialized Text: A Practical Guide to Using Corpora.* Routledge.

[Bowker2004] Lyanne Bowker. 2004. Corpus Resources for Translators: Academic Luxury or Professional Necessity?. *TradTerm,* 10, 213–247.

[Budin2015] Gerhard Budin. 2015. Digital Humanities, Language Industry, and Multilingualism – Global Networking and Innovation in Collaborative Methods. In Martin Forstner and Hannelore Lee-Jahnke (eds.), *CIUTI-Forum- 2014*. Boston, USA: Peter Lang, 423–448. DOI: http://dx.doi.org/10.3726/978-3-0352-0290-8.

[Chunpir et al2015] Hashim Iqbal Chunpir, Thomas Ludwig, and Dean N. Williams. 2015. Evolution of E-Research: From Infrastructure Development to Service Orientation. In Aaron Marcus (ed.), *Design, User Experience, and Usability: Interactive Experience Design: 4th International Conference, DUXU 2015, Proceedings*. Cham: Springer, 25–35. http://dx.doi.org/10.1007/978-3-319-20889-3.

[Ehrensberger-Dow et al2016] Maureen Ehrensberger-Dow, Andrea Hunziker Heeb, Gary Massey, Ursula Meidert, Silke Neumann, and Heidrun Karin Becker. 2016. An International Survey of the Ergonomics of Professional Translation. *ILCEA Revue de l'Institut des Langues et des Cultures d'Europe et d'Amérique, 27*.

[EMT Expert Group2009] EMT Expert Group. 2009. *Competences for Professional Translators, Experts in Multilingual and Multimedia Communication.*

[Fantinuoli and Zanettin.2015] Claudio Fantinuoli and Federico Zanettin. 2015. Creating and Using Multilingual Corpora in Translation Studies. In Claudio Fantinuoli and Federico Zanettin (eds.), *New Directions in Corpus-based Translation Studies*. Berlin: Language Science Press, 1–10.

[Gallego-Hernández2015] Daniel Gallego-Hernández. 2015. The Use of Corpora as Translation Resources: A Study Based on a Survey of Spanish Professional Translators. *Perspectives: Studies in Translatology*, 23.3, 375–391.

[Granger2003] Sylviane Granger. 2003. The Corpus Approach: A Common Way forward for Contrastive Linguistics and Translation Studies. In Sylviane Granger, Jacques Lerot and Stephanie Petch-Tyson (eds.). *Corpus-based Approaches to Contrastive Linguistics and Translation Studies.* Amsterdam & Atlanta: Rodopi, 17–29.

[Hagemann2016] Susanne Hagemann. 2016. (Non-)Professional, Authentic Projects? Why Terminology Matters. In Don Kiraly (ed.), *Towards Authentic Experiential Learning in Translator Education.* Mainz University Press, 33–51.

[Jenkins et al2003] Christine Jenkins, Cynthia L. Corritore, and Susan Wiedenbeck. 2003. Patterns of Information Seeking on the Web: A Qualitative Study of Domain Expertise and Web Expertise. *IT & Society,* 1.3 (2003), 64–89.

[Kelly2009] Diane Kelly. 2009. Methods for Evaluating Interactive Information Retrieval Systems with Users. *Foundations and Trends in Information Retrieval*, 3 (1–2), 1–224.

[Kemman et al2014] Max Kemman, Stef Scagliola, Franciska de Jong, and Roeland Ordelman. 2014. Talking with Scholars: Developing a Research Environment for Oral History Collections. In Łukasz Bolikowski, Vittore Casarosa, Paula Goodale, Nikos Houssos, Paolo Manghi, and Jochen Schirrwagen (eds.), *Theory and Practice of Digital Libraries -- TPDL 2013 Selected Workshops*. Springer International Publishing, 197–201.

[Kenny2007] Dorothy Kenny. 2007. Translation Memories and Parallel Corpora: Challenges for the Translator Trainer. In Dorothy Kenny and Kyongjoo Ryou (eds.), *Across Boundaries: International Perspectives on Translation Studies.* Cambridge Scholars Publishing, 192–208.

[Krüger2012] Ralph Krüger. 2012. Working with Corpora in the Translation Classroom. *Studies in Second Language Learning and Teaching,* 4, 505–525.

[Kübler2003] Natalie Kübler. 2003. Corpora and LSP translation. In Federico Zanettin, Silvia Bernardini and Dominic Stewart (eds.), *Corpora in Translator Education*. Manchester: St Jerome, 2003, 25-42.

[Laviosa2002] Sara Laviosa. 2002. *Corpus-based Translation Studies: Theory, Findings, Applications.* Amsterdam – New York, NY: Rodopi B.V.

[Maegaard et al2016] Bente Maegaard, Lina Henriksen, Andrew Joscelyne, Vesna Lušicky, Margaretha Mazura, Sussi Olsen, Claus Povlsen, and Philippe Wacker. 2016. Providing a Catalogue of Language Resources for Commercial Users. In *Proceedings of LREC 2016, Tenth International Conference on Language Resources and Evaluation*, 449–456.

[Maia2003] Belinda Maia. 2003. Some Languages are More Equal than Others. Training Translators in Terminology and Information Retrieval Using Comparable and Parallel Corpora. In Federico Zanettin, Silvia Bernardini and Dominic Stewart (eds.), *Corpora in Translator Education*. Manchester: St Jerome, 43–53.

[Marchionini2006] Gary Marchionini. 2006. Exploratory Search: From Finding to Understanding. *Communications of the ACM - Supporting Exploratory Search*, 49(4), 41–46.

[Mellange2006] Mellange. 2006. MeLLANGE Corpora & E-learning Survey Results. *ITI Bulletin.*

[Odijk2014] Jan Odijk. 2014. *Discovering Resources in CLARIN: Problems and Suggestions for Solutions.* Utrecht University Repository, Netherlands.

[Rehm and Uszkoreit2012] Georg Rehm and Hans Uszkoreit. 2012. *Europe's Languages in the Digital Age. White Paper Series*. Springer.

[Risku2016] Hanna Risku. 2016. Situated Learning in Translation Research Training: Academic Research as a Reflection of Practice. *The Interpreter and Translator Trainer*, 10(1), 12–28.

[Robson and McCartan2016] Colin Robson and Kieran McCartan. 2016. *Real World Research*. Chichester, West Sussex: Wiley.

[Russell-Rose and Tyler2013] Tony Russell-Rose and Tate Tyler Tate. 2013. *Designing the Search Experience: The Information Architecture of Discovery*. Morgan Kaufmann Publishers Inc.

[Sánchez-Gijón2002] Pilar Sánchez-Gijón. 2002. Aplicaciones de la Lingüística de Corpus a la Práctica de la Tradución. Complemento de la Traducción Asistida por Ordenador. *Terminologie et Traduction*, 2, 84–106.

[Van Uytvack et al2010] Dieter Van Uytvanck, Claus Zinn, Daan Broeder, Peter Wittenburg, and Mariano Gardelleni. 2010. Virtual Language Observatory: The Portal to the Language Resources and Technology Universe. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation*, 900–903.

[Van Uytvanck2012] Dieter Van Uytvanck, Hermann Stehouwer, and Lari Lampen. 2012. Semantic Metadata Mapping in Practice: The Virtual Language Observatory. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, 1029–1034.

[Vandepitte2013] Sonia Vandepitte. 2013. Research Competences in Translation Studies. *Babel*, 59 (2), 125–148. DOI:10.1075/babel.59.2.

[Wildemuth and Freund2012] Barbara M. Wildemuth and Luanne Freund. 2012. Assigning Search Tasks Designed to Elicit Exploratory Search Behaviors. *Human Computer Information Retrieval (HCIR) Symposium*, 4, 1-10.

[Wilkinson2005] Michael Wilkinson. 2005. Using a Specialized Corpus to Improve Translation Quality. *Translation Journal,* 9.3.

[Wynne2015] Martin Wynne. 2015. *User Involvement.* Presentation at the Clarin Annual Conference 2015. https://www.clarin.eu/sites/default/files/20151016-CAC-04-Wynne-User-Involvement-CAC2015-05.pdf (accessed 1.4.2017).

[Zanettin1998] Federico Zanettin. 1998. Bilingual Comparable Corpora and the Training of Translators. *Meta: Journal des Traducteurs / Meta: Translators' Journal*, 43, 616–630.

[Zanettin2012] Federico Zanettin. 2012. *Translation-Driven Corpora Corpus Resources for Descriptive and Applied Translation Studies.* Taylor and Francis.

TalkBank and CLARIN

Brian MacWhinney
Department of Psychology
Carnegie Mellon University, Pittsburgh USA
macw@cmu.edu

Abstract

TalkBank promotes the use of corpora, web-based access, multimedia linkage, and human language technology (HLT) for the study of spoken language interactions in a variety of discourse types across many languages, involving children, second language learners, bilinguals, people with language disorders, and classroom learners. Integration of these materials within CLARIN provides open access to access a large amount of research data, as well as a test bed for the development of new computational methods.

1 Introduction

The TalkBank system (http://talkbank.org) is the world's largest open access repository for spoken language data. It provides language corpora and resources for a variety of research topics in Psychology, Linguistics, Education, Computer Science, and Speech Pathology. There are currently seven funded TalkBank components. The National Institutes of Health (NIH) funds the development of the CHILDES database (http:childes.talkbank.org) for the study of child language development (MacWhinney, 2000), PhonBank (phonbank.talkbank.org) for the study of phonological development (Rose & MacWhinney, 2014), AphasiaBank (aphasia.talkbank.org) for the study of language in aphasia (MacWhinney & Fromm, 2015), and FluencyBank (fluency.talkbank.org) for the study of the development of fluency and disfluency in children and language learners (Bernstein Ratner & MacWhinney, 2016). The National Science Foundation (NSF) provides additional funding for FluencyBank, as well as funding for HomeBank (http://homebank.talkbank.org) with daylong audio recordings in the home (VanDam et al., 2016). The National Endowment for the Humanities (NEH) and the Deutsche Forschungs Gesellschaft (DFG) have provided funding for web-based access to materials from Classical Latin and Historical German (http://sla.talkbank.org).

In addition to these seven funded projects, TalkBank has developed resources for TBIBank (traumatic brain injury), RHDBank (right hemisphere damage), ASDBank (autism), DementiaBank (dementia), CABank (Conversation Analysis) (MacWhinney & Wagner, 2010), SamtaleBank (Danish), GestureBank (gesture), SLABank (Second Language Acquisition) (MacWhinney, 2015b), BilingBank (bilingualism), ClassBank (classroom interactions), and TutorBank (human tutors). All these resources use a common transcript format called CHAT which is used by the CLAN analysis programs and other open-access resources. Except for some of the corpora from clinical areas and the daylong recordings from the home, these resources are available without passwords.

TalkBank includes 348 corpora contributed by researchers across all these fields. After corpora have been contributed, they undergo additional reformatting, curation, indexing, annotation, and linkage to media. The result is a unified open-access database with a fully consistent system of transcription and annotation across all corpora. We believe that this type of data integration with open access is important for maximizing the value of the corpora contributed to TalkBank, and that this method can serve as a model for other CLARIN data sites.

In 2014, the TalkBank center at Carnegie Mellon University in Pittsburgh became a CLARIN-B site, and in 2016 it became a CLARIN-K site. TalkBank is the first CLARIN site outside of Europe.

Brian MacWhinney 2017. TalkBank and CLARIN. *Selected papers from the CLARIN Annual Conference 2016*. Linköping Electronic Conference Proceedings 136: 76–89.

This paper will summarize the principles underlying the design of TalkBank, the ways in which TalkBank has implemented CLARIN standards, and how it can provide resources for the CLARIN community.

2 The Motivation for TalkBank

Most language resources derive from written sources, such as books, newspapers, and the web. It is relatively easy to enter such written data directly into computer files for further linguistic (Baroni & Kilgarriff, 2006) and behavioral (Pennebaker, 2012) analysis. On the other hand, preparation of spoken language data for computational analysis is much more difficult. Despite ongoing advances in speech technology (Hinton et al., 2012), collection of spoken language corpora still depends on a time-consuming process of hand transcription. Because of this, the total quantity of spoken language data available for analysis is much less than that available for written language, although face-to-face conversation is the original and primary root of human language. Furthermore, unplanned spoken language (Givon, 2005; Redeker, 1984) includes many prosodic features, gestural components, reductions, and hesitation phenomena that further complicate transcription and analysis.

Because of its conceptual centrality, there are several major disciplines that examine aspects of face-to-face communication. These include Psycholinguistics, Development Psychology, Applied Linguistics, Phonology, Theoretical Linguistics, Conversation Analysis, Gestural Studies, Human-Computer Interaction, Social Psychology, Speech and Hearing, Neuroscience, Evolutionary Biology, and Political Science. To understand language acquisition, second language learning, language attrition, language change, language disorders, sociolinguistic variation, persuasion, and group communication, we will need to combine methods and insights from each of these disciplines. Through such comparisons, and by examining language usage across a range of timescales (MacWhinney, 2015a), we can address core issues such as: how language is learned, how it is processed, how it changes, and how it can be restored after damage.

Like written language (Biber, 1991), the forms of spoken language vary enormously from situation to situation (Hymes, 1962). However, individual speakers can operate smoothly within each of these varied contexts. This means that, to fully understand human language and its role in human culture, we need to compare language use across many situations, forms, and participants. Because of this diversity, contrasts between practices in individual laboratories and disciplines have made the forms of transcription and coding for spoken language corpora remarkably unstandardized, making it difficult to construct comparisons across corpora. The first goal of TalkBank is to provide a system that bridges across these differences by providing an inclusive standard that recognizes all the features required for these specific disciplinary analyses. To achieve this goal, TalkBank has elaborated the CHAT transcription standard.

The development and extension of the CHAT transcription standard represents a necessary precondition to the central goal of TalkBank, which is to encourage and support data-sharing across all the language sciences. In the physical sciences, the process of data-sharing is taken as a given. However, until recently, data-sharing has not been adopted as the norm in the social sciences. This failure to share research results – much of it supported by public funds – represents a huge loss to science. Researchers often cite privacy concerns as reasons for not sharing data on spoken interactions. However, as illustrated at http://talkbank.org/share/irb/options.html, TalkBank provides many ways in which data can be made available to other researchers, while still preserving participant anonymity.

3 Many Banks in One

TalkBank is composed of 17 component banks, each using the same CHAT transcription format and database organization standards. This section describes the contents and each of these component language banks. The homepage at http://talkbank.org provides links to each of these 17 banks, as well as related resources.

3.1 CHILDES

The CHILDES (Child Language Data Exchange System) database at http://childes.talkbank.org is the oldest of TalkBank's component banks. Brian MacWhinney (CMU) and Catherine Snow (Harvard School of Education) began the CHILDES system in 1984 with funding from the MacArthur Foundation. Snow organized a meeting in the (appropriately named) town of Concord, Massachusetts at which many of the major figures in the child language field agreed on the basic principles for sharing child language data. In the early 1980s, researchers were just beginning to use personal computers and transcribed data was still stored in 9-track tapes, punch cards, and floppy disks. The Internet was not generally available for data transmission, so data was shared by mailing CD-ROM copies to members. At that time, there was no thought that the transcripts might eventually be linked to audio or video. As a result, researchers often destroyed or recycled their audio recordings. Since that early beginning, CHILDES has grown in coverage, membership, and output. Since 1987, the project has been funded by NIH with some additional support from NSF. The table at the end of this section shows that there now are over 7000 published articles based on the use of data or programs from CHILDES. This work extends across the areas of phonology, morphology, syntax, lexicon, narrative, literacy, and discourse.

Using CHILDES data and methods, researchers have evaluated alternative theoretical approaches to comparable data. For example, the debate between connectionist models of learning and dual-route models focused first on data regarding the learning of the English past tense (MacWhinney & Leinbach, 1991; Marcus et al., 1992; Pinker & Prince, 1988) and later on data from German plural formation (Clahsen & Rothweiler, 1992). In syntax, emergentists (Pine & Lieven, 1997) have used CHILDES data to elaborate an item-based theory of learning of the determiner category, whereas generativists (Valian, Solt, & Stewart, 2009) have used the same data to argue for innate categories. Similarly, CHILDES data in support of the Optional Infinitive Hypothesis (Wexler, 1998) have been analyzed in contrasting ways using the MOSAIC system (Freudenthal, Pine, & Gobet, 2010) to demonstrate constraint-based inductive learning. In these debates, and many others, the availability of a shared open database has been crucial in the development of analysis and theory. Based on these contributions, CHILDES serves as a model and inspiration for next-generation data-sharing projects in child development such as Databrary (http://databrary.org) and Wordbank (http://wordbank.stanford.edu).

3.2 PhonBank

During the first two decades of work on the CHILDES system, it was frustratingly difficult to adapt computer transcripts for the study of children's phonological development. Researchers used ASCII-based system such as ARPANET, SAMPA, PHONASCII, and UNIBET. However, application of these systems across languages was difficult and error-prone. The LIPP system (Nathani & Oller, 2001) solved some of these problems, but the proprietary nature of its font encoding made it difficult to integrate into transcripts, and it provided no linkage to media. With the introduction of Unicode in the 1990s and the promulgation of fonts supporting data entry for IPA such as Arial Unicode and the SIL Unicode IPA fonts (http://fonts.sil.org), it became increasingly easier to represent children's phonological productions in a standardized way. Building on this opportunity, Yvan Rose (Memorial University, Newfoundland) and Brian MacWhinney (CMU) initiated the PhonBank project. Working with a consortium of researchers in child phonology, and supported now for over 10 years by grants from NICHD, the PhonBank project has accumulated 40 corpora of early child phonological productions across 12 languages, all transcribed in IPA along with the target language forms and linked directly to the audio record. These new corpora are available in two formats: CHAT and Phon, and these two formats subscribe to the single underlying CHAT XML Schema that guarantees complete interoperability. Files in CHAT transcript format can be analyzed using the CLAN programs which we will describe later. Files in Phon format can be analyzed using the Phon program. Phon provides all the basic analyses required in the study of child phonology for tracking the growth of segments, features, prosodic patterns, and phonological processes. In addition, Phon incorporates the full source code of Praat (http://praat.org), making it possible to run Praat's acoustic analysis directly inside Phon and storing the results in the Phon transcript.

3.3 HomeBank

HomeBank, which began in 2015, is one of the newest components of TalkBank. It is supported by a grant from the National Science Foundation to Anne Warlaumont (UC Merced), Mark VanDam (Washington State University), and Brian MacWhinney (CMU). The primary data in HomeBank are daylong (i.e. 16-hour) audio recordings collected from children in the home through use of the LENA recording system (http://www.lena.org). This system uses a small digital recording device sewn into a child's vest. The LENA software processes the captured audio to identify who is speaking when, but it does not attempt to recognize words. The output of this processing includes a text file in LENA's ITS format and the associated WAV file. To include these data in HomeBank, we use the LENA2CHAT conversion program in CLAN (http://childes.talkbank.org/clan) to output CHAT format. Researchers then select segments of these huge CHAT files for detailed language transcription. HomeBank currently includes 3.5 TB of these audio recordings and this number will soon grow well beyond this.

Because these data have no transcripts, we cannot provide public access to segments that may include potentially embarrassing material. Researchers interested in working with the non-public versions of these data must undergo careful debriefing regarding this issue before they are given access. To make at least some of this huge quantity of material publicly available, our students and research assistants listen through complete recordings to spot any questionable material, which they then tag in the CHAT transcript with a code for later silencing. Determining what should count as embarrassing material in these natural contexts is itself an interesting research topic.

Even without transcripts, these recordings can address many issues regarding the language environment of the young child. How much input is the child receiving and when? Do children who receive more input acquire language more quickly and does that help them in later years? How much responsivity do different adults show to child vocalizations? How do a child's intonational patterns change over time? These and many other questions can be addressed even without additional coding. However, when these recordings are accompanied by video or when various new methods for automatic analysis are used, the data can address an even broader range of research questions. For example, we are currently working with Florian Metze (Metze, Riebling, Warlaumont, & Bergelson, 2016) to apply the Speech Recognition Virtual Kitchen (SRVK) methodology (http://speechkitchen.org) to the CHAT and audio files derived from LENA. InterSpeech 2017 includes a challenge to see how well the SRVK methodology can diarize these recordings and identify the various speakers. If this methodology proves to be as good as that provided by the LENA system, we will work to make it available through open source, and we will work to create inexpensive recording devices that can be used with this non-proprietary software.

3.4 AphasiaBank

Aphasia involves the loss of language abilities, often arising from a stroke or embolism. This condition affects nearly 2 million people in the United States alone, making it the most common adult communication disorder. To improve our understanding of language usage and recovery in aphasia, NIH has been funding the AphasiaBank project for 10 years. Unlike the other language banks, AphasiaBank emphasizes the collection of data based on a tightly specified elicitation protocol. This protocol requires that the investigator follow a script in terms of asking questions and eliciting narratives. The detailed components of the protocol can be found at http://aphasia.talkbank.org/protocol. Using this standardized protocol, we have collected, transcribed and analyzed 402 hour-long interviews from persons with aphasia (PWAs) and 220 age-matched control participants. All transcripts are linked to the video at the utterance level and can be played back using the TalkBank browser over the web. Analysis of these materials have generated 256 publications, and the videos are used as teaching materials in universities and clinics throughout the English-speaking world. AphasiaBank also has smaller numbers of recordings for French, Cantonese, Spanish, and German, collected through translations of the protocol and the protocol materials into these languages.

We plan several extensions of AphasiaBank. First, we will record and transcribe increasingly naturalistic interactions in both group therapy sessions and conversations in the home. Second, we will test out the effects on language recovery of the use of tablet-based teletherapy lessons. Finally, we will use the Speech Kitchen methodology noted above to analyze the productions of people with aphasia and people with apraxia of speech (AoS) when reciting a scripted passage. The advantage of this method for speech recognition is that the words that must be recognized are restricted to those in the scripted passage.

3.5 Other Clinical Banks

Following the lead of AphasiaBank, we have developed protocols for data collection from four other varieties of language disorder. DementiaBank already includes a fairly large sample from earlier projects on language in dementia. We will formulate a data collection protocol for this area. RHDBank examines the language and problem-solving abilities of people who have suffered from right hemisphere damage. TBIBank examines language from people suffering from traumatic brain lesions. Both RHDBank and TBIBank use a protocol close to that of AphasiaBank. Finally, ASDBank includes data from both children and adults with autism spectrum disorder.

3.6 FluencyBank

The other most recently funded TalkBank component is FluencyBank, based on a collaboration between Nan Bernstein Ratner (University of Maryland) and Brian MacWhinney (CMU). The development of FluencyBank is supported by two separate federal grants. The grant from NIDCD seeks to characterize the development pathway of fluency and disfluency in children between the ages of 3 and 7. During this period, many of the children that show signs of early disfluency end up as normally fluent, with only a fraction of this population developing stuttering. How and why this occurs developmentally remains a mystery, largely because data from this period are incomplete. To address this, we are using TalkBank methods to conduct a longitudinal study across this period. To supplement this work, NSF has provided support for incorporating data from earlier studies of disfluency from a variety of laboratories, much of it coded in SALT format.

Work in speech technology is centrally important for the development of FluencyBank. We need to not only analyze transcripts for lexicon, morphology, and syntax, but also carefully track word and segment repetitions, retraces, drawls, and overall durations. Ideally, these data should be linked to the audio records through a process of automatic diarization. Our initial work with this method indicates that this is feasible.

3.7 SLABank and BilingBank

SLABank currently includes 31 corpora from second language learners, and BilingBank includes 10 corpora from bilinguals. Nearly all of these corpora are accompanied by audio, although only a few have been linked to the audio at the utterance level. In addition to these corpora from adult learners and bilinguals, the CHILDES database has 32 corpora tracing the development of childhood bilingualism. To facilitate the analysis of grammatical development, we have developed a method for tagging multilingual corpora using a combination of unilingual taggers. This system is based on the taggers and parsers we have developed for Cantonese, Danish, Dutch, English, French, German, Hebrew, Japanese, Italian, Mandarin, and Spanish (MacWhinney, 2008). For bilingual corpora that use any combination of these languages, we use marks to encode the language source of each word. To minimize the actual marks being used, we establish the notion of a matrix (Myers-Scotton, 2005) language, so that only intrusions into the matrix are marked. This form of coding not only allows efficient tagging, but also provides a good profile of code-switching behavior.

We hope to be able to link this growing corpus collection with data from experimental and tutorial approaches to second language learning as characterized in a recent proposal for establishment of an SLAWeb (MacWhinney, in press).

3.8 CABank and SCOTUS

Conversation Analysis (CA) is a methodological and intellectual tradition stimulated by the ethnographic work of Garfinkel (1967) and systematized by Sacks, Schegloff, and Jefferson (1974) among others. With support from the Danish BG Bank Foundation, Johannes Wagner (Southern Denmark University) and Brian MacWhinney (CMU) developed methods for producing Jeffersonian CA transcription within CHILDES. We then collected and formatted a database of CA materials, including such classics as Jefferson's Newport Beach transcripts and the Watergate Tapes. There are currently 20 other corpora in CABank. One particularly large corpus that is not yet in CA format is the SCOTUS corpus developed in collaboration with Jerry Goldman (University of Illinois). This corpus – the largest in TalkBank – includes 50 years of oral arguments from the US Supreme Court linked on the utterance level to the audio. We also have a CHAT-encoded versions of the Santa Barbara Corpus of Spoken American English (SBCSAE) and the Michigan Corpus of Academic Spoken English (MICASE). CHAT/CA is being used in a variety of labs internationally that are planning to contribute additional data.

3.9 ClassBank

ClassBank includes 15 corpora of transcripts linked to video from classroom interactions. The largest of these are the Curtis corpus from a year-long study of instruction in Geometry in fourth grade (Lehrer & Curtis, 2000) and the seven-nation TIMMS study of teaching in Math and Science (Stigler, Gallimore, & Hiebert, 2000).

3.10 SamtaleBank

The creation of SamtaleBank was supported by a DK-CLARIN grant to Bente Maegaard (University of Copenhagen) and Johannes Wagner (Southern Denmark University). This bank includes the conversational component of the current DK-CLARIN corpus for Danish. All materials are carefully transcribed in CA format and linked to either the audio or video media. This collection serves as a model for the further construction of well-prepared materials for Conversation Analysis.

3.11 GestureBank

Creating a database of videotaped, transcribed, and coded interactions for the study of gestures during speaking has proven to be one of the most difficult challenges facing TalkBank. Coding and transcribing gestures is much more difficult than coding and transcribing spoken language. Unlike spoken language, there is no accepted method for gesture coding or transcription. Even if one tries to implement one of the dozens of proposed methods, it can take as long as a week to code one hour of gesture. Partly as a result of these problems, data-sharing has not taken hold as a norm in this community. Faced with these challenges, our work in this area has focused on the construction of a coding system that can be deployed more simply within the framework of the CLAN editor and programs. Our initial proposals along these lines are included along with other tutorial screencasts at http://talkbank.org/screencasts .

3.12 LangBank

With support from an NEH/DFG binational grant, Anke Lüdeling (Humboldt University, Berlin), Detmar Meurers (Tübingen), and Brian MacWhinney (CMU) are developing methods based on TalkBank, SLAWeb, and ANNIS (http://corpus-tools.org). In this project, we are creating systematized and aligned JSON versions of corpora for both Classical Latin and Historical German. This language bank represents an exception to the TalkBank focus on spoken language, because neither of these classical languages is actively spoken in a language community today. Instead, the focus here is on the development of these corpora in the SLAWeb framework (MacWhinney, in press) to support effective language learning. Moreover, this collaboration allows us to make contact with CLARIN-related groups studying issues such as corpus analysis (Berlin) (Lüdeling, Walter, Kroymann, & Adolphs, 2005) readability (Tübingen) (Meurers, 2005, 2012; Meurers et al., 2010), and the learning of classical languages (Leipzig).

3.13 Usage

To monitor the usage of the various components of TalkBank, we can track indices such as articles published and web hits. We are able to rely on http://scholar.google.com to track usage, because we have requested that people using these data include a reference to (MacWhinney, 2000) in their reference list. Table 1 summarizes these indices for the six major funded TalkBank components.

	CHILDES	Talk Bank	Aphasia Bank	Phon Bank	Fluency Bank	Home Bank
Age (years)	28	12	8	6	0.5	1
Words (millions)	59	47	1.8	0.8	0.5	audio
Linked Media (TB)	2.8	1.1	0.4	0.7	0.3	3.5
Languages	41	22	6	18	4	2

Publications	7000+	320	256	480	5	5
Users	2950	930	390	182	50	18
Web hits (millions)	4.3	1.3	0.3	0.1	-	0.2

Table 1: TalkBank Usage

4 Principles

TalkBank relies on a series of principles for data sharing, formatting, access, analysis, and user involvement. In this section, we will review these principles. Many of these principles adhere closely to CLARIN standards. Others seek to expand on these standards.

4.1 Data-sharing

The most fundamental TalkBank principle is the idea that the results of scientific investigations should be shared with the scientific community. This principle may seem like a platitude. We all know that scientists are supposed to open their ideas to further testing and development. However, as we noted earlier, data-sharing has not been adopted as the norm in many areas of the social sciences. The core goal of TalkBank is to correct this situation by building easy methods for data-sharing that will lead to important results for scientific investigation. CLARIN subscribes to similar principles.

Data-sharing can be encouraged through either the carrot or the stick. However, the only stick that carries much weight is one wielded by a funding agency. Agencies such as NIH and NSF now stipulate that, at the end of a project, the results of the project should be fully shared and archived. Funding agencies in Europe have also moved increasingly toward this standard. However, there remain large gaps in the enforcement of these standards. This is beginning to change, as granting agencies have begun to require that proposals must document the effective sharing of data from earlier funded research.

Researchers often claim that they cannot share data because of IRB (Human Subjects) restrictions. However, if there is proper planning and administration of informed consent at the beginning of a study, IRB problems can all be resolved. Similarly, investigators often complain that, if they contribute their data to a database like TalkBank, other researchers could publish analyses that might preempt or "scoop" their own plans for publication. This concern can be addressed by contributing data along with the specification of an embargo period, after which data will become publicly available.

The other approach to data-sharing involves carrots. In past decades, carrots have been more effective than sticks. Researchers have learned that contributed data will be cited, thereby increasing their citation index. To facilitate citation, we associate DOI (Digital Object Identifier) numbers with each corpus. Also, researchers find that by working with TalkBank they become members of a community of interest that furthers their communication with researchers having similar interests. In addition, by contributing data to TalkBank, researchers can use the increasingly powerful TalkBank tools to perform new analyses on their own corpora. This could be done without data contribution, because the programs are all open access. However, if we know that corpora are to be contributed to TalkBank, we will devote special attention to customizing analytic programs for the needs of particular projects.

4.2 Open access

Data-sharing implies open access. If a researcher contributes a corpus to a database, but refuses to permit open access, this is not real data-sharing. Corpora can be protected by passwords if necessary, but these passwords should be readily granted to qualified researchers. Provision of open access to corpora has been a problem for other database efforts, including some of those in CLARIN. Some archives only permit access to data through a search interface. This may work for certain types of queries, but it places restrictions on the types of questions that a researcher may pose regarding a dataset. In other cases, corpora are really not available at all. For example, many of the materials in The Language Archive (tla.mpi.nl) are not available for access. Limits on access also make it difficult for projects such as Linked Open Data or Federated Content Search. Hopefully many of these restrictions on access to corpora will be removed in the future.

4.3 Consistent format

A third important TalkBank principle is that all the data in TalkBank are transcribed in a single consistent format. This is the CHAT format which is compatible with the CLAN programs. This format has been developed over the years to accommodate the needs of a wide range of research communities and disciplinary perspectives. The format is described discursively in the CHAT manual, which is available from http://talkbank.org/manuals/CHAT.html. The full computational description is provided in the XML Schema viewable and browsable from http://talkbank.org/software/xsddoc/ . This XSD description includes links between the XML characterizations of CHAT elements and their description in the MS-Word manual.

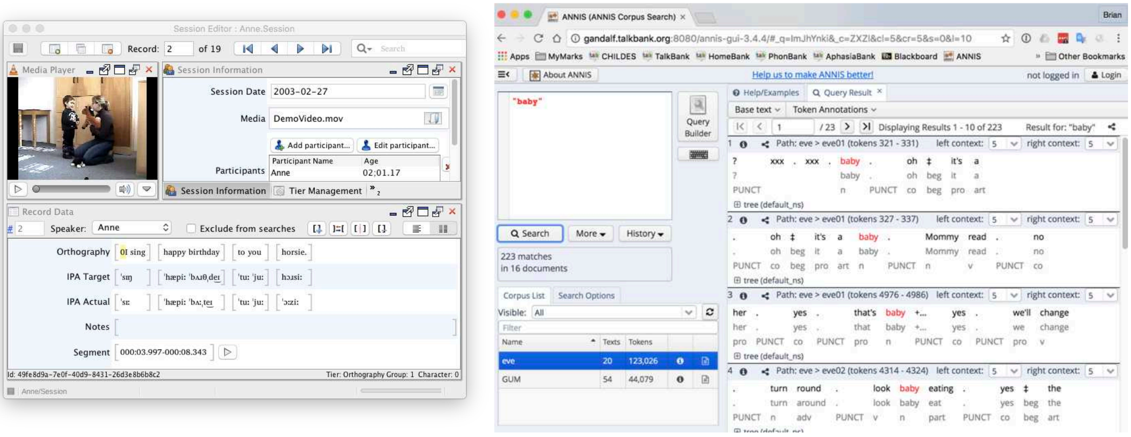
Before data are entered into one of the TalkBank databases, they must first pass through two levels of format checking. The first level relies on the CHECK program built into CLAN. Because this checker is built right into the CLAN Editor, it is easy for users to check their work frequently to make sure they are following the requirements of CHAT. This checker is able to catch most potential errors in the use of CHAT format. However, the fullest checking is done through the Chatter XML formatter and validator that can be downloaded from http://talkbank.org/software/chatter.html. Chatter is able to convert files in CHAT format into XML and vice versa. It can also output PHON format.

4.4 Interoperability

Using conversion programs available inside CLAN, transcripts in CHAT format can be automatically converted both to and from the formats required for Praat (praat.org), PHON (childes.talkbank.org/phon), ELAN (tla.mpi.nl/tools/elan), CoNLL, ANVIL (anvil-software.org), EXMARaLDA (exmaralda.org), LIPP (ihsys.com), SALT (saltsoftware.com), LENA (lenafoundation.org), and Transcriber (trans.sourceforge.net). To provide fuller database and corpus facilities, we created a Pepper importer (Zipser & Romary, 2010) from CHAT data to ANNIS (http://corpus-tools.org) as well as a local ANNIS server (http://gandalf.talkbank.org:8080/annis-gui-3.4.4/).

Because CHAT recognizes such a wide variety of information types (dates, speaker roles, intonational patterns, retrace markings, etc.), when data are converted into the other formats, there must be methods for protecting CHAT data types not recognized by these other programs against loss. This is done in two ways. First, we can often "hide" CHAT data in special comment fields that are not processed by the program, but which will be available later for export. Second, when using the other programs, users are warned to be careful not to alter codes in CHAT format that mark aspects not recognized by the other programs. There are no cases in which information created in the other programs cannot be represented in CHAT, because CHAT is a superset of the information represented in these other programs.

In some cases, these conversions between CHAT and other formats involve the minimalist level of interoperability characterized by annotation graphs (Bird & Liberman, 2001). This level simply marks the beginning and end of some annotation in terms of its time from the beginning of the media. This is the type of remapping achieved for imports and exports to ELAN, ANVIL, Transcriber, and EXMARaLDA. However, other forms of conversion, such as those involving LIPP, LENA, SALT, ANNIS, and PHON include a full remapping of the semantics of the codes used in each format in their corresponding values in CHAT. The two screenshots in Figure 1 give example of the results of these types of transfer. The screenshot on the left shows data from a CHAT transcript that has been exported to and opened in PHON; the one on the right shows CHAT data has been exported to and opened in ANNIS.



Session Date 2003-02-27

Media DemoVideo.mov

Participants

Participant Name	Age
Anne	02;01.17

Speaker: Anne

Orthography	0I sing	happy birthday	to you	horsie.
IPA Target	ˈsɪŋ	ˈhæpiː ˈbʌɹθˌdeɪ	ˈtuː ˈjuː	hɔɹsiː
IPA Actual	ˈsɪː	ˈhæpiː ˈbʌːˌteɪ	ˈtuː ˈjuː	ˈɔːziː
Notes				
Segment	000:03.997-000:08.343			

Id: 49fe8d9a-7e0f-40d9-8431-26d3e8b6b8c2 Tier: Orthography Group: 1 Character: 0

gandalf.talkbank.org:8080/annis-gui-3.4.4/#_q=ImJhYnki&_c=ZXZI&cl=5&cr=5&s=0&l=10

"baby"

223 matches
in 16 documents

Visible: All

Name	Texts	Tokens
eve	20	123,026
GUM	54	44,079

1 / 23 Displaying Results 1 - 10 of 223 Result for: "baby"

1 Path: eve > eve01 (tokens 321 - 331) left context: 5 right context: 5

?	xxx	.	xxx	.	baby	.	oh	‡	it's	a
?					baby	.	oh	beg	it	a
PUNCT					n	PUNCT	co	beg	pro	art

tree (default_ns)

2 Path: eve > eve01 (tokens 327 - 337) left context: 5 right context: 5

.	oh	‡	it's	a	baby	.	Mommy	read	.	no
.	oh	beg	it	a	baby	.	Mommy	read	.	no
PUNCT	co	beg	pro	art	n	PUNCT	n	v	PUNCT	co

tree (default_ns)

3 Path: eve > eve01 (tokens 4976 - 4986) left context: 5 right context: 5

her	.	yes	.	that's	baby	+...	yes	.	we'll	change
her	.	yes	.	that	baby	+...	yes	.	we	change
pro	PUNCT	co	PUNCT	pro	n	PUNCT	co	PUNCT	pro	v

tree (default_ns)

4 Path: eve > eve02 (tokens 4314 - 4324) left context: 5 right context: 5

.	turn	round	.	look	baby	eating	.	yes	‡	the
.	turn	around	.	look	baby	eat	.	yes	beg	the
PUNCT	n	adv	PUNCT	v	n	part	PUNCT	co	beg	art

Figure 1: CHAT data that have been exported to PHON (left) and ANNIS (right)

During the process of building the database, we often needed to reformat files from still other, less documented formats. However, now most of the material we receive is already in CHAT format.

4.5 Media Linkage, Diarization

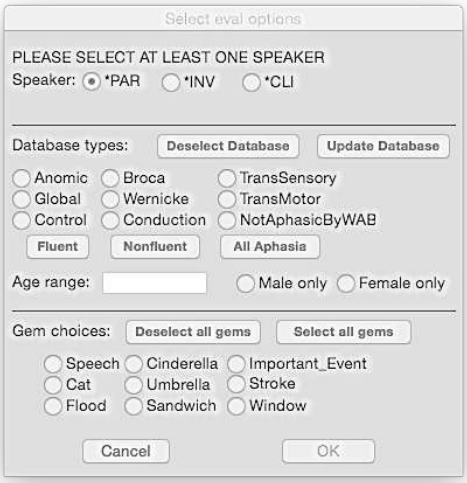
The majority of corpora in TalkBank have transcripts linked to either audio or video at the utterance level. By linking transcripts to the original recordings, we have lifted a burden off of the shoulders of the transcriber. Without linkage, transcription is forced to fully represent all of the important details of the original interaction. With linkage, transcription serves as a key into the original recording that allows each researcher to add or modify codes as needed. If a phonetician does not agree with the transcription of a segment of babbling, then it is easy to provide an alternative transcription.

The linkage of transcripts to recordings opens up a new way of thinking about corpora and the process of data sharing. In the previous model, we could only share the computerized transcripts themselves. For some important child language corpora, such as the Brown corpus, the original recordings have been lost. For others, however, we have been able to locate the original reel-to-reel recordings and convert them to digital files that we then link to the transcripts. Now, when corpora are contributed to TalkBank, we make sure that contributors provide both the transcripts in CHAT and the media.

Linkage to media on the utterance level is valuable for many aspects of language analysis from CA to child language. However, diarization through automatic speech recognition (ASR) methods can provide a more precise characterization of the temporal profiles of words and utterances. Diarization marks the time values for each word, allowing us to also find the values of intra-sentential and inter-sentential pauses. This type of analysis is important for work on language disorders and studies of turn-taking. One of our goals for the future is to increase the diarization of TalkBank corpora.

4.6 Protocol Formulation

Projects such as AphasiaBank and FluencyBank rely heavily on the construction of a data elicitation protocol to maximize the comparability of results across participants. The composition of these protocols is determined by an Advisory Board composed of members of each research community. The goal here is to be systematically compare data from speakers at different ages, speaking different languages, in different tasks and situations, at different stages of learning, and with different clinical profiles. To facilitate these comparisons, we have developed a series of programs for each relevant database. For aphasia, the program is called EVAL. Using this program, we can extract group means for individual aphasia types (Broca's, Wernicke's, anomia, global, transcortical motor, and transcortical sensory) which we then use as comparisons for the results from individual participants with aphasia. The screenshot in Figure 2 shows some of the options which can be used when comparing a given participant with the larger database.



Select eval options

PLEASE SELECT AT LEAST ONE SPEAKER

Speaker: ● *PAR ○ *INV ○ *CLI

Database types: Deselect Database | Update Database

○ Anomic ○ Broca ○ TransSensory
○ Global ○ Wernicke ○ TransMotor
○ Control ○ Conduction ○ NotAphasicByWAB

Fluent | Nonfluent | All Aphasia

Age range: ○ Male only ○ Female only

Gem choices: Deselect all gems | Select all gems

○ Speech ○ Cinderella ○ Important_Event
○ Cat ○ Umbrella ○ Stroke
○ Flood ○ Sandwich ○ Window

Cancel | OK

Figure2: Options for comparing a transcript with a database in EVAL

For child language data, the parallel program is called KIDEVAL and it uses mother-child play sessions in the full CHILDES database as its comparison sample. The comparison database for FluencyBank is under construction. Comparisons of this type are fundamental to the process of clinical assessment, as well as the study of basic developmental processes.

4.7 Analysis Tools

For ten of the languages in the database, we provide automatic morphosyntactic analysis using the MOR, POST, and MEGRASP programs built into CLAN. These languages are Cantonese, Chinese, Dutch, English, French, German, Hebrew, Japanese, Italian, and Spanish. Tagging is done by MOR, disambiguation by POST, and dependency analysis by MEGRASP. MOR was written by Mitzi Morris, based on specifications for a left-associative morphology (LA-MORPH) from Roland Hausser (Hausser, 1999). POST was developed by Christophe Parisse (Parisse & Le Normand, 2000) and MEGRASP was developed by Kenji Sagae (Sagae, Davis, Lavie, MacWhinney, & Wintner, 2007). Details regarding the operation of the taggers, disambiguators, and dependency analyzers for these languages can be found in MacWhinney (2008). In each of these languages processing involves unique computational challenges. The complexity and linguistic detail required for analysis of Hebrew forms is perhaps the most extensive. In German, special methods are used for achieving tight analysis of the elements of the noun phrase. In French, it is important to mark various patterns of suppletion. Japanese requires quite different codes for parts of speech and dependency relations. Eventually, the codes produced by these programs will be harmonized with the GOLD ontology (Farrar & Langendoen, 2010). In addition, we compute a dependency grammar analysis for each of these 10 languages, which we are harmonizing with the Universal Dependency tagset (http://universaldependencies.org). It is also possible to use other dependency taggers rather than MEGRASP by reformatting a CHAT into CONLL format using the CHAT2CONLL and CONLL2CHAT programs in CLAN. The results of the morphological analysis by MOR and POST are stored in the %mor lines of a CHAT files and the results of the grammatical dependency analysis produced by MEGRASP are stored in the %gra lines. Triple clicking on a %gra line in a CHAT files invokes the GraphViz web service that produces a graph of the utterances for display on the user's screen, such as the one in Figure 3 for the first sentence from Julius Caesar's *De Bello Gallico*.

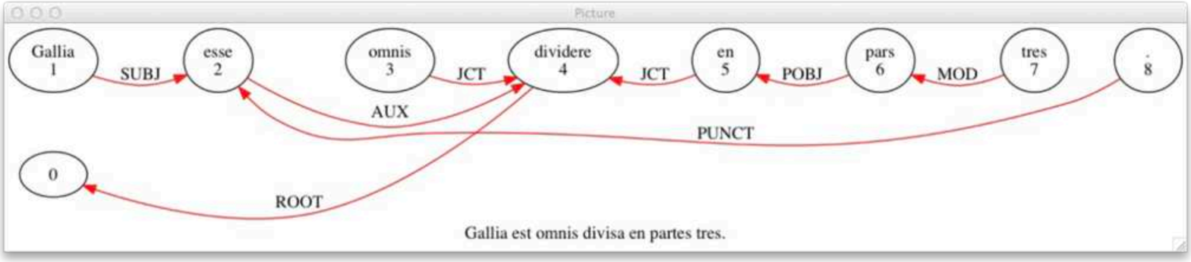


Figure 3: A dependency graph produced by GraphViz from a CHAT %syn line

Because these morphosyntactic analyzers use a parallel technology and output format, CLAN commands can be applied to each of these 10 languages for uniform computation of indices such as MLU (mean length of utterance), VOCD (vocabulary diversity) (Malvern, Richards, Chipere, & Purán, 2004), pause duration, and various measures of disfluency. In addition, we have automated language-specific measures such as DSS (Lee, 1974) (for English and Japanese) and IPSyn (Scarborough, 1990). Following the method of Lubetich and Sagae (2014), we are now developing language-general measures based on classifier analysis with SVN that can be applied to all 10 languages using the codes in the morphological and grammatical dependency analyses. However, there are many other languages in the database for which we do not yet have morphosyntactic taggers. This means that it is a priority to construct MOR systems for languages with large amounts of CHILDES and TalkBank data, such as Catalan, Indonesian, Polish, Portuguese, and Thai.

4.8 Metadata Publication

Metadata for the transcripts and media in the TalkBank databases are included in the two major systems for accessing linguistic data: OLAC, and CMDI/TLA. Each transcript and media file has been assigned a PID (permanent ID) using the Handle System (www.handle.net). In addition, each corpus has received an ISBN number and a DOI (digital object identifier) code. PID numbers are encoded in the header lines of each transcript file and the ISBN and DOI numbers are entered in 0metadata.cdc files included in each corpus as well as in HTML web pages that include extensive documentation for each corpus, photos and contact information for the contributors, and articles to be cited when using the data. All these resources are periodically checked and synchronized using the SCONS program that relies on the fact that there is a completely isomorphic hierarchical structure for the CHAT data, the XML versions of the CHAT data, the HTML web pages, and the media files. If information is missing for any item within this parallel set of structures, the updating program reports the error and it is fixed. All this information is then published using an OAI-PMH compatible method for harvesting through systems such as the Virtual Linguistic Observatory (VLO) developed through the CLARIN. Currently 13% of the records in the VLO come from TalkBank.

4.9 Community Support and Sustainability

Corpus creators may believe that making a database easily available will lead to its general usage. The idea is that you "build it and they will come". However, a fuller version of this motto would be "build it, curate it, publicize it, make usage easy, construct clear documentation, and provide workshops and free snacks, and they will eventually come." In practice, all these things are necessary, and we have worked continually on all these fronts to incorporate the use of TalkBank data and methods into training and research practice.

Making the system easily available is closely linked to the goal of sustainability. TalkBank's approach to sustainability focuses on integrating our corpora and tools with the basic research agenda of each of our participating language research communities. To the degree that we achieve such integration, funding for our work is tied to ongoing funding for basic research. For example, when developing tools for the study of child language development, we focus on methods for automatic morphosyntactic coding, because of the importance of grammatical analysis in language acquisition theory. For aphasia, we focus on morphosyntax, lexical access, error analysis, and aspects of fluency. For the projects on disfluency and stuttering, we work on the application of tools for automatic speech recognition (ASR), including diarization and word-level alignment to characterize the linguistic environment and distribution of disfluencies. We also seek to achieve sustainability and survivability by using open-source software tools with full documentation and by linking to tool chains in the CLARIN infrastructure.

5 Integration with CLARIN

It is our goal to make TalkBank materials fully accessible and discoverable for CLARIN users, and to integrate CLARIN tools into the TalkBank analysis chains. The award of CLARIN-B Centre status indicates that much of this integration has already been achieved. We have implemented all the requirements for this status both for CLARIN and for Data Seal of Approval recognition. We achieved further integration in 2016, through the recognition of TalkBank as a CLARIN-K Centre for Knowledge distribution. In this role, TalkBank will provide information for researchers interested in working with spoken language corpora, using either CLAN or any of the other software analysis system with which CLAN and CHAT are compatible. We can offer support through email, mailing lists, and phone with extremely quick turnaround. We have been creating online screencasts demonstrating the use of TalkBank tools, and we welcome suggestions for the creation of additional methods. These resources can become particularly important if CLARIN seeks to provide a higher level of support for the study of spoken language interactions.

The major challenge currently facing TalkBank integration into CLARIN is a fiscal one. Because the United States is not a member of the European Union, it has no clear mechanism for providing financial support for CLARIN membership. In 2017, the CMU University Library agreed to provide modest support for integration with CLARIN. We hope to extend this first step by creating a CLARIN Infrastructure with multiple research sites in the United States, such as Brandeis, the University of Pennylvania, the University of Illinois, or Columbia. How we can secure long-term funding across these sites remains to be seen.

The process of integration of TalkBank with CLARIN can also be viewed from a slightly different perspective. In addition to making sure that TalkBank aligns with CLARIN standards, we can consider how CLARIN could benefit more fully from TalkBank as a model. First, if CLARIN could adopt the CHAT coding system as the default for data representation for spoken language, it would greatly enhance the value of its resources. Such a step would require buy-in from many parties and additional work in reformatting, but it would be a major step forward. Second, adoption of TalkBank methods for promoting open access and data-sharing would be of great value to CLARIN. Finally, CLARIN could benefit from developing ways of linking sustainability to the development of specific corpora and tools that are crucially relevant to individual research groups. By making its tools a fundamental part of the infrastructure of research communities, CLARIN could guarantee its long-term survival.

6 Conclusion

TalkBank seeks to provide data that can help us integrate insights about language from across all the human sciences. To achieve this goal, it has developed a series of component data banks focusing on specific aspects of human language, all made comparable through a uniform transcription standard and principles for data-sharing.

TalkBank plays an important role within the larger CLARIN infrastructure in terms of providing resources for the analysis of spoken language interactions. Unlike many other resources in this area, TalkBank resources are available through completely open access and rely on a consistent data format. The individual components of TalkBank are each responsive to the practices and agenda of individual research communities. These features of TalkBank may serve as a model for parallel developments in CLARIN.

Acknowledgements

The development of TalkBank is currently supported by NICHD grant HD082736 to Brian MacWhinney for CHILDES, NICHD grant HD051698 To Yvan Rose and Brian MacWhinney for PhonBank, NIDCD grant DC008524 to Brian MacWhinney for AphasiaBank, NSF SBE RIDIR Grants 1539129, 1539133, and 1539010 to Anne Warlaumont, Mark VanDam, and Brian MacWhinney for HomeBank, NEH/DFG grant to Brian MacWhinney, Anke Lüdeling, and Detmar Meurers for LangBank, NIDCD Grant DC015494 to Brian MacWhinney and Nan Ratner for new FluencyBank data, and NSF SBE Grant to Brian MacWhinney and Nan Ratner for FluencyBank archiving.

References

Baroni, M., & Kilgarriff, A. (2006). *Large linguistically-processed web corpora for multiple languages.* Paper presented at the Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations.

Bernstein Ratner, N., & MacWhinney, B. (2016). Your laptop to the rescue: Using the Child Language Data Exchange System archive and CLAN utilities to improve child language sample analysis. *Seminars in Speech and Language, 37*, 74-84.

Biber, D. (1991). *Variation across speech and writing*: Cambridge University Press.

Bird, S., & Liberman, M. (2001). A formal framework for linguistic annotation. *Speech Communication, 33*, 23-60.

Clahsen, H., & Rothweiler, M. (1992). Inflectional rules in children's grammars: Evidence from German participles. In G. Booij & J. Van Marle (Eds.), *Yearbook of Morphology*. Dordrecht: Kluwer.

Farrar, S., & Langendoen, D. T. (2010). An owl-dl implementation of gold *Linguistic Modeling of Information and Markup Languages* (pp. 45-66): Springer.

Freudenthal, D., Pine, J., & Gobet, F. (2010). Explaining quantitative variation in the rate of Optional Infinitive errors across languages: A comparison of MOSAIC and the Variational Learning Model. *Journal of Child Language, 37*, 643-669.

Givon, T. (2005). *Context as other minds: The pragmatics of sociality, cognition, and communication*. Philadelphia, PA: John Benjamins.

Hausser, R. (1999). *Foundations of computational linguistics*. Berlin: Springer.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., . . . Sainath, T. N. (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.

Hymes, D. (1962). The ethnography of speaking. *Anthropology and human behavior, 13*(53), 11-74.

Lee, L. (1974). *Developmental Sentence Analysis*. Evanston, IL: Northwestern University Press.

Lehrer, R., & Curtis, C. L. (2000). Why are some solids perfect? *Teaching Children Mathematics, 6*(5), 324.

Lüdeling, A., Walter, M., Kroymann, E., & Adolphs, P. (2005). Multi-level error annotation in learner corpora. *Proceedings of corpus linguistics 2005*, 15-17.

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

MacWhinney, B. (2008). Enriching CHILDES for morphosyntactic analysis. In H. Behrens (Ed.), *Trends in corpus research: Finding structure in data* (pp. 165-198). Amsterdam: John Benjamins.

MacWhinney, B. (2015a). Introduction: Language Emergence. In B. MacWhinney & W. O'Grady (Eds.), *Handbook of Language Emergence* (pp. 1-32). New York, NY: Wiley.

MacWhinney, B. (2015b). Multidimensional SLA. In S. Eskilde & T. Cadierno (Eds.), *Usage-based perspectives on second language learning* (pp. 22-45). New York, NY: Oxford University Press.

MacWhinney, B. (in press). A shared platform for studying second language acquisition. *Language Learning*.

MacWhinney, B., & Fromm, D. (2015). AphasiaBank as Big Data. *Seminars in Speech and Language, 37*, 10-22.

MacWhinney, B., & Leinbach, J. (1991). Implementations are not conceptualizations: Revising the verb learning model. *Cognition, 29*, 121-157.

MacWhinney, B., & Wagner, J. (2010). Transcribing, searching and data sharing: The CLAN software and the TalkBank data repository. *Gesprächsforschung, 11*, 154-173.

Malvern, D., Richards, B., Chipere, N., & Purán, P. (2004). *Lexical diversity and language development*. New York, NY: Palgrave Macmillan.

Marcus, G. F., Pinker, S., Ullman, M., Hollander, M., Rosen, T. J., Xu, F., & Clahsen, H. (1992). Overregularization in language acquisition. *Monographs of the Society for Research in Child Development*, i-178.

Metze, F., Riebling, E., Warlaumont, A. S., & Bergelson, E. (2016). *Virtual machines and containers as a platform for experimentation*. Paper presented at the Cognitive Science, Philadelphia, PA.

Meurers, D. (2005). On the use of electronic corpora for theoretical linguistics: Case studies from the syntax of German. *Lingua, 115*(11), 1619-1639.

Meurers, D. (2012). Natural language processing and language learning. *The Encyclopedia of Applied Linguistics*. doi:10.1002/9781405198431.wbeal0858

Meurers, D., Ziai, R., Amaral, L., Boyd, A., Dimitrov, A., Metcalf, V., & Ott, N. (2010). *Enhancing authentic web pages for language learners.* Paper presented at the Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications.

Myers-Scotton, J. (2005). Supporting a differential access hypothesis: Code switching and other contact data. In J. F. Kroll & A. M. B. DeGroot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 326-348). New York, NY: Oxford University Press.

Nathani, S., & Oller, D. K. (2001). Beyond ba-ba and gu-gu: Challenges and strategies in coding infant vocalizations. *Behavior Research Methods, Instruments, & Computers, 33*(3), 321-330.

Parisse, C., & Le Normand, M.-T. (2000). Automatic disambiguation of the morphosyntax in spoken language corpora. *Behavior Research Methods, Instruments, and Computers, 32*, 468-481.

Pennebaker, J. W. (2012). *Opening up: The healing power of expressing emotions*: Guilford Press.

Pine, J. M., & Lieven, E. V. M. (1997). Slot and frame patterns and the development of the determiner category. *Applied Psycholinguistics, 18*, 123-138.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a Parallel Distributed Processing Model of language acquisition. *Cognition, 29*, 73-193.

Redeker, G. (1984). On differences between spoken and written language*. *Discourse Processes, 7*(1), 43-55.

Rose, Y., & MacWhinney, B. (2014). The PhonBank Project: Data and software-assisted methods for the study of phonology and phonological development. In J. Durand, U. Gut, & G. Kristoffersen (Eds.), *The Oxford handbook of corpus phonology* (pp. 380-401). Oxford: Oxford University Press.

Sagae, K., Davis, E., Lavie, A., MacWhinney, B., & Wintner, S. (2007). High-accuracy annotation and parsing of CHILDES transcripts *Proceedings of the 45th Meeting of the Association for Computational Linguistics* (pp. 1044-1050). Prague: ACL.

Scarborough, H. S. (1990). Index of productive syntax. *Applied Psycholinguistics, 11*, 1-22. doi:10.1017/S0142716400008262

Stigler, J., Gallimore, R., & Hiebert, J. (2000). Using video surveys to compare classrooms and teaching across cultures: Examples and lessons from the TIMSS video studies. *Educational Psychologist, 35*(2), 87-100.

Valian, V., Solt, S., & Stewart, J. (2009). Abstract categories or limited-scope formulae? The case of children's determiners. *Journal of Child Language, 36*(04), 743-778.

VanDam, M., Warlaumont, A. S., Bergelson, E., Cristia, A., Soderstrom, M., Palma, P. D., & MacWhinney, B. (2016). HomeBank: An online repository of daylong child-centered audio recordings. *Seminars in Speech and Language, 37*(2), 128-142. doi:10.1055/s-0036-1580745

Wexler, K. (1998). Very early parameter setting and the unique checking constraint: A new explanation of the optional infinitive stage. *Lingua, 106*, 23-79.

Zipser, F., & Romary, L. (2010). *A model oriented approach to the mapping of annotation formats using standards.* Paper presented at the Workshop on Language Resource and Language Technology Standards, LREC 2010.

The Curation Module and Statistical Analysis on VLO Metadata Quality

Davor Ostojic	**Go Sugimoto**	**Matej Ďurčo**
ACDH-OEAW	ACDH-OEAW	ACDH-OEAW
Vienna, Austria	Vienna, Austria	Vienna, Austria
davor.ostojic@oeaw.ac.at	go.sugimoto@oeaw.ac.at	matej.durco@oeaw.ac.at

Abstract

The Curation Module is developed to facilitate the metadata ingestion and curation process of the Virtual Language Observatory (VLO) by providing a systematic method to measure metadata quality and a user-friendly interface to inspect profiles, records, and collections of the Component MetaData Infrastructure (CMDI) used for the VLO. A large amount of useful statistics generate a comprehensive data matrix including information about the quality score, publication status, facet coverage, and metadata header, as well as the number of records and concepts. The module helps various stakeholders to automatically and systematically identify the metadata problems. Whilst metadata modellers can evaluate the quality of shared profiles, data creators assess the validity of newly created records. Data providers can use it for the improvement of their metadata for better discoverability and accessibility of valuable linguistic contents, whereas working groups could examine the actual use of profiles and records to define the next version of CMDI and VLO. Thus, the Curation Module supports all stages of metadata management and fosters the analysis and improvement of metadata quality to enhance the CLARIN services. In this article, we present a selection of statistical information on the metadata quality made possible by the Curation Module.

1 Background

Metadata quality is central to resource discovery. It determines the discoverability and accessibility of resources for the users and metadata curation plays an essential role to control the quality. CLARIN is not an exception. Its main metadata catalogue of language resources, Virtual Language Observatory (VLO)[1] suffers from a backlash of the flexibility of Component MetaData Infrastructure (CMDI)[2], which is a standardised metadata framework underlying VLO. In fact, metadata curation has been a long standing issue in CLARIN, hence the Metadata Curation Task Force was founded to tackle it. Most recently, we investigated the variability issues of metadata in VLO (King et al., 2016) and the idea of a Curation Module was formalised to provide a solution to assess the quality of the ingested metadata. By now we know how many CLARIN centres are registered (centre registry[3]), some of which are the data providers of VLO, how many records are ingested into VLO (its home page), how many collections we

[1] https://vlo.clarin.eu

[2] https://www.clarin.eu/content/component-metadata

[3] https://centres.clarin.eu/

Davor Ostojic, Go Sugimoto and Matej Ďurčo 2017. The curation module and statistical analysis on VLO metadata quality. *Selected papers from the CLARIN Annual Conference 2016.* Linköping Electronic Conference Proceedings 136: 90–101.

have received (CMDI harvester web view[4]), and how many metadata concepts (CLARIN Concept Registry[5] hereafter CCR) and profiles (Component Registry[6]) are created to define and semantically bind different types of resource descriptions. In addition, extra efforts brought us such valuable information as to the structure of CMD profiles and the reuse of CMD components and concepts (SMC Browser[7])(Ďurčo 2013) and what percentage of VLO facets are covered (King et al., 2016; Odijk, 2014). However, it was not possible to systematically and automatically collect statistics about the quality of the CMDI metadata. In 2015, we presented the general functional concept of the Curation Module in the context of overall VLO data ingestion workflow (King et al., 2016) in accordance with some previous works (Kemps-Snijders, 2014; Trippel et al., 2014). The Curation Module then became one of the deliverables of CLARIN-PLUS project[8]. This paper will outline the ongoing development of the module and demonstrate the first findings on the metadata quality made possible by this module, as well as other relevant statistics.

2 The Curation Module

2.1 Overview

The Curation Module is a software tool developed as a component of the CLARIN metadata infrastructure for curation and quality assessment / benchmarking of CMD records, collections and profiles. It is intended as technical support for human curation work to monitor and improve the metadata quality. The design of the module was guided by the following four use cases[9]:

1. The metadata editor checks (on-the-fly) the quality and validity of a newly created record.
2. The metadata modeller evaluates the quality of profiles (especially facet coverage), when selecting an existing profile or creating a new profile for new resources.
3. The data provider, repository administrator, or collection manager checks the overall quality of metadata in his/her repository, including the facet coverage.
4. All records ingested into the VLO undergo a systematic process of curation, validation, normalisation and quality assessment (benchmarking).

The Curation Module consists of two parts: a core Java application that works standalone or can be used in other software as library, and a web application which provides a web-based interface as well as a RESTful API. The module can process web resources via URL as well as locally stored CMD records and collections. In addition to the interface for assessing own data, the user can explore pre-processed assessments of public profiles[10] (figure 1) and the collections harvested by the CLARIN aggregator. The Curation Module heavily depends on other CLARIN infrastructure services such as the Component Registry from where it fetches the XSD schema files of the CMD profiles and the Concept Registry from where it retrieves information about concepts.

[4] https://vlo.clarin.eu/data/

[5] https://openskos.meertens.knaw.nl/ccr/browser/

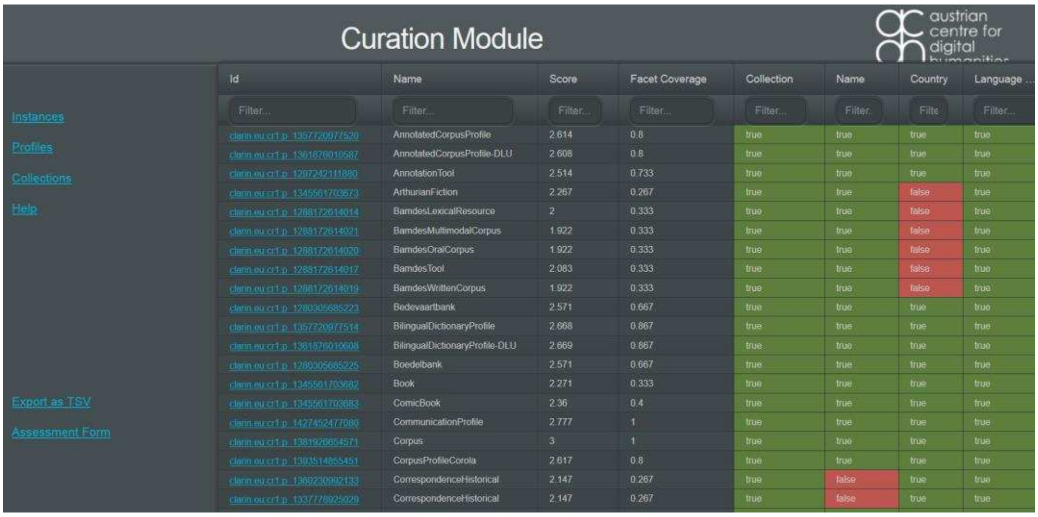
[6] http://catalog.clarin.eu/ds/ComponentRegistry/

[7] https://clarin.oeaw.ac.at/ /smc-browser/

[8] https://www.clarin.eu/node/4213

[9] https://office.clarin.eu/v/CE-2016-0742-CLARINPLUS-D2_1.pdf

[10] Public profiles are the profiles publicly shared in the Component Registry, as opposed to the private profiles (or non-public profiles) which are only visible to the creator.



Curation Module

Id	Name	Score	Facet Coverage	Collection	Name	Country	Language ...
Filter...	Filter...	Filter...	Filter...	Filter...	Filter..	Filte	Filter...
clarin.eu:cr1:p_1357720977520	AnnotatedCorpusProfile	2.614	0.8	true	true	true	true
clarin.eu:cr1:p_1361876010587	AnnotatedCorpusProfile-DLU	2.608	0.8	true	true	true	true
clarin.eu:cr1:p_1297242111880	AnnotationTool	2.514	0.733	true	true	true	true
clarin.eu:cr1:p_1345561703673	ArthurianFiction	2.267	0.267	true	true	false	true
clarin.eu:cr1:p_1288172614014	BamdesLexicalResource	2	0.333	true	true	false	true
clarin.eu:cr1:p_1288172614021	BamdesMultimodalCorpus	1.922	0.333	true	true	false	true
clarin.eu:cr1:p_1288172614020	BamdesOralCorpus	1.922	0.333	true	true	false	true
clarin.eu:cr1:p_1288172614017	BamdesTool	2.083	0.333	true	true	false	true
clarin.eu:cr1:p_1288172614019	BamdesWrittenCorpus	1.922	0.333	true	true	false	true
clarin.eu:cr1:p_1280305685223	Bedevaartbank	2.571	0.667	true	true	true	true
clarin.eu:cr1:p_1357720977514	BilingualDictionaryProfile	2.668	0.867	true	true	true	true
clarin.eu:cr1:p_1361876010608	BilingualDictionaryProfile-DLU	2.669	0.867	true	true	true	true
clarin.eu:cr1:p_1280305685225	Boedelbank	2.571	0.667	true	true	true	true
clarin.eu:cr1:p_1345561703682	Book	2.271	0.333	true	true	true	true
clarin.eu:cr1:p_1345561703683	ComicBook	2.36	0.4	true	true	true	true
clarin.eu:cr1:p_1427452477080	CommunicationProfile	2.777	1	true	true	true	true
clarin.eu:cr1:p_1381926654571	Corpus	3	1	true	true	true	true
clarin.eu:cr1:p_1393514855451	CorpusProfileCorola	2.617	0.8	true	true	true	true
clarin.eu:cr1:p_1360230992133	CorrespondenceHistorical	2.147	0.267	true	false	true	true
clarin.eu:cr1:p_1337778925029	CorrespondenceHistorical	2.147	0.267	true	false	true	true

Figure 1 Curation Module lists the assessment of public profiles

For each input type (CMD record, profile, or collection) the Curation Module defines a distinct workflow of phases each of which collects statistics from different VLO components (the Component Registry and CCR) and generates a corresponding XML report (described in the following subsections). There is also a special type of report which is generated in case of non-recoverable errors during the data assembling workflow. This type of report contains only error messages. If the algorithm of a phase runs successfully, statistics are generated out of the gathered information and the quality assessment score for that phase is calculated. The overall score is calculated by summing up individual scores from each phase in the workflow. Finally, a report is created by combining these statistics, scores, and eventual issues. Each issue has information about the phase in which it occurred, the severity (warnings and errors), and a verbose message. The primary output format for this assessment report is XML, but it is also rendered in HTML for a user-friendly view in the web application. In the next subsections we describe the main features of the three different types of reports.

2.2 Profile report

Profile assessment workflow is divided into three phases: header, components/concepts and facet mapping assessment. In terms of profile scoring (table 1), points are added for the publication status of the profile (public or private), the percentage of elements annotated with concepts, and the VLO facet coverage. The maximum score is 3.0.

Criteria	**Score**
Publication status	0 or 1
The percentage of elements annotated with concepts	[0.0 .. 1.0]
Facet coverage	[0.0 .. 1.0]
Total	**0.0 .. 3.0**

Table 1. Scoring criteria for profiles

A report is structured with the following sections (matching the different phases of the workflow):

1. Meta information about the profile with name, id, description, link to schema, CMDI version and lifecycle status of the profile. The score of this phase is based on the publication status: if the profile is public, one point is given.

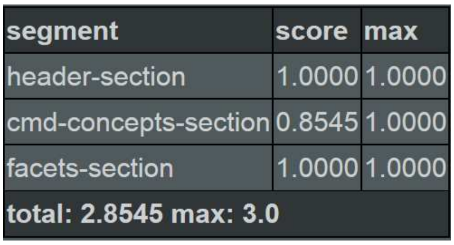
2. Score section presents the summary of the scores from each phase and total score in form of a table (figure 2).

3. Facet mapping section provides information about covered facets. The score represents the facet coverage of the profile.

4. Component section lists the components used in the profile in a table with the following columns: component name, id and count.

5. Concept section firstly shows statistics about elements. Following table lists information about the concept names which links to the corresponding CCR page, CCR status and count. The percentage of annotated elements with the CCR concepts represents the score from this section.

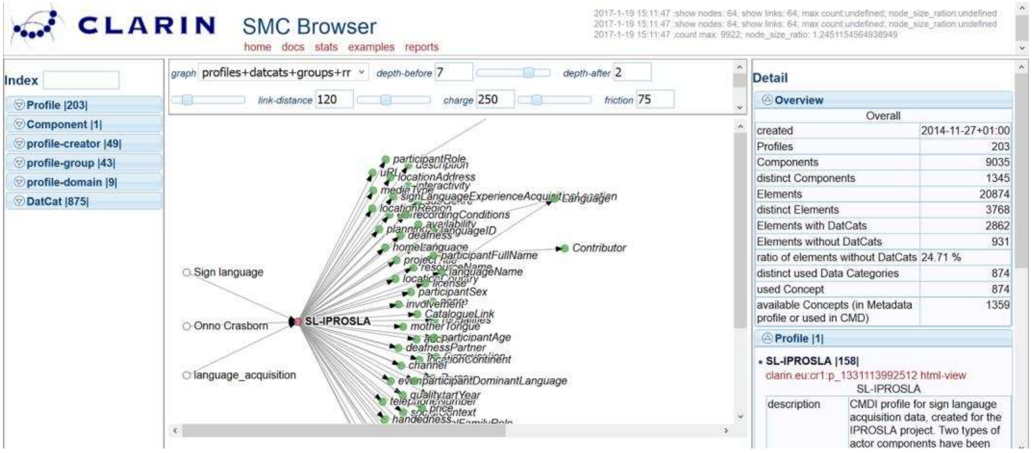
6. Last part of the report shows the issues encountered during assessment workflow. The user can see in which phase they occurred, severity level, and message about the problem.



segment	score	max
header-section	1.0000	1.0000
cmd-concepts-section	0.8545	1.0000
facets-section	1.0000	1.0000
total: 2.8545 max: 3.0		

Figure 2. Score summary for profile assessment

In the pre-processed profile assessment, a direct link is given to the SMC Browser, where the users can explore the complex network of the CMD profile within an interactive application for exploring graph data (figure 3).



CLARIN SMC Browser

2017-1-19 15:11:47 :show nodes: 64; show links: 64; max count:undefined; node_size_ration:undefined
2017-1-19 15:11:47 :show nodes: 64; show links: 64; max count:undefined; node_size_ration:undefined
2017-1-19 15:11:47 :count max: 9922; node_size_ratio: 1.2451154564938949

Index

- Profile |203|
- Component |1|
- profile-creator |49|
- profile-group |43|
- profile-domain |9|
- DatCat |875|

graph profiles+datcats+groups+rr depth-before 7 depth-after 2
link-distance 120 charge 250 friction 75

Detail

Overview

Overall	
created	2014-11-27+01:00
Profiles	203
Components	9035
distinct Components	1345
Elements	20874
distinct Elements	3768
Elements with DatCats	2862
Elements without DatCats	931
ratio of elements without DatCats	24.71 %
distinct used Data Categories	874
used Concept	874
available Concepts (in Metadata profile or used in CMD)	1359

Profile |1|

- SL-IPROSLA |158|
clarin.eu:cr1:p_1331113992512 html-view
SL-IPROSLA

description	CMDI profile for sign langauge acquisition data, created for the IPROSLA project. Two types of actor components have been

Figure 3. A profile in Curation Module directly links to the profile view in the SMC Browser

2.3 Instance report

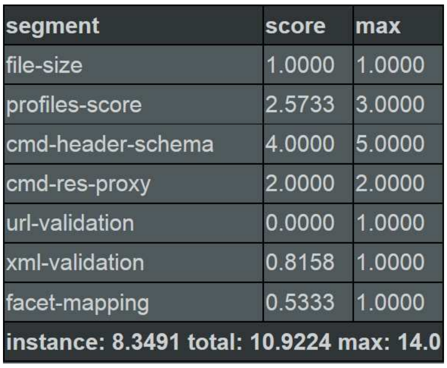
The workflow for instance assessment consists of the following seven processing phases: file, header, profile, resource references, XML validation, facet mapping and URL validation. The score for an instance is based on the score of the used profile, the record size, the completeness of the CMD header, the presence of references to resources, XML syntax validation, the resolvability of links, and facet coverage (table 2). The maximum score is 14.0.

Criteria	**Score**
Profile's score	[0.0 .. 3.0]
File size is less than 10 Mb	0 or 1
Schema location is present	0 or 1
Schema resides in Component Registry	0 or 1
MdProfile element contains a valid value	0 or 1
MdCollectionDisplayName is not empty	0 or 1
MdSelfLink is not empty	0 or 1
Rate of resources with MIME type	[0.0 .. 1.0]
Rate of resource proxy elements with references	[0.0 .. 1.0]
Rate of non-empty XML elements	[0.0 .. 1.0]
Rate of accessible URLs	[0.0 .. 1.0]
Facet coverage	[0.0 .. 1.0]
Total	**0.0 .. 14.0**

Table 2. Scoring criteria for collections

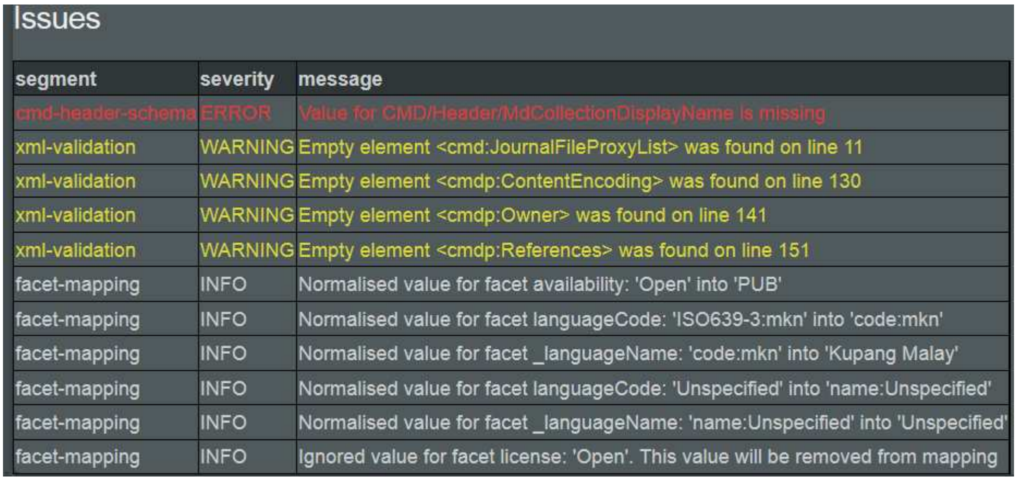
An instance report presents information in the following order:

1. The first section shows combined information from the first three phases of assessment. The user can see information about the file name, profile id linked to the respective profile report, and the size of the record in bytes.
2. Score section (figure 4) displays the summary of individual scores and the overall score with and without the underlying profile score.
3. Facet section tells how the VLO "sees" the record. Beside facets and normalised values, XPath and concept information are available. Eventual missing facets will be listed at the bottom of the table. The score from this phase is used for facet coverage.
4. Resource proxy section presents statistics about the resources described by the record. Score is the summary of percentage of the resource with MIME type and with references.
5. XML validation section shows statistics, gathered during XML validation against the schema, about the elements in the record. The rate of populated elements gives the score from this phase.
6. URL validation section delivers statistics about HTTP links from the record and their resolvability (or persistency). The score is represented with the rate of resolvable links.
7. The last part of the report shows human-readable messages about the errors, warnings and other potentially useful information issued in each assessment phase (figure 5).



segment	score	max
file-size	1.0000	1.0000
profiles-score	2.5733	3.0000
cmd-header-schema	4.0000	5.0000
cmd-res-proxy	2.0000	2.0000
url-validation	0.0000	1.0000
xml-validation	0.8158	1.0000
facet-mapping	0.5333	1.0000
instance: 8.3491 total: 10.9224 max: 14.0		

Figure 4. Example of the score section in the report



Issues

segment	severity	message
cmd-header-schema	ERROR	Value for CMD/Header/MdCollectionDisplayName is missing
xml-validation	WARNING	Empty element <cmd:JournalFileProxyList> was found on line 11
xml-validation	WARNING	Empty element <cmdp:ContentEncoding> was found on line 130
xml-validation	WARNING	Empty element <cmdp:Owner> was found on line 141
xml-validation	WARNING	Empty element <cmdp:References> was found on line 151
facet-mapping	INFO	Normalised value for facet availability: 'Open' into 'PUB'
facet-mapping	INFO	Normalised value for facet languageCode: 'ISO639-3:mkn' into 'code:mkn'
facet-mapping	INFO	Normalised value for facet _languageName: 'code:mkn' into 'Kupang Malay'
facet-mapping	INFO	Normalised value for facet languageCode: 'Unspecified' into 'name:Unspecified'
facet-mapping	INFO	Normalised value for facet _languageName: 'name:Unspecified' into 'Unspecified'
facet-mapping	INFO	Ignored value for facet license: 'Open'. This value will be removed from mapping

Figure 5. Example of the issue section for instance assessment

2.4 Collection report

Workflow for collection assessment consists of assessment of all the contained records and aggregation of collected statistics from each individual record. Currently collections can be only processed in command line mode with a file system path as an input. A collection report contains the following sections:

1. Overview with the name of the collection, the total score (the sum of scores from all instances within), the average score, the minimal and maximal score in the collection.
2. File section supplies statistics about the number of records as well as the total, minimal, maximal and average size of them.
3. Header section lists the profiles referenced by the records in the collection, with the respective records count, score, and links to the corresponding reports.
4. Facet section lists average facet coverage for each individual facet.
5. Resource Proxy and XML validation sections display aggregated statistics as the total and average figures that come from the corresponding sections of the instance reports.

3 Preliminary quality analysis of existing data[11]

In this section we present the main results of the analyses of the two pre-processed datasets, namely profiles and collections, giving some idea about the status quo of the CMDI data as available in the VLO. In addition to the reports of the Curation Module, a few other statistics about the VLO data were collected, visualised, analysed, and interpreted to complement our analyses.

[11] All the numbers are updated as of January 2017.

3.1 Profile analysis

Examining all the profiles referenced by harvested CMD records, the highest score is 2.87 out of 3.0, and 0.79 is the lowest (figure 6). The score distribution is generally good at the moment, suggesting a fairly good initial setup of the score calculation. There is a clear gap between public profiles and private ones, because all the private profiles are below average, implying an easy improvement potential for the private profiles, when being converted to public. This scoring system, thus, attempts to give incentive to make profiles public, because the essence of CMDI is the collective effort of interoperability by publicly sharing common profiles to be reused.

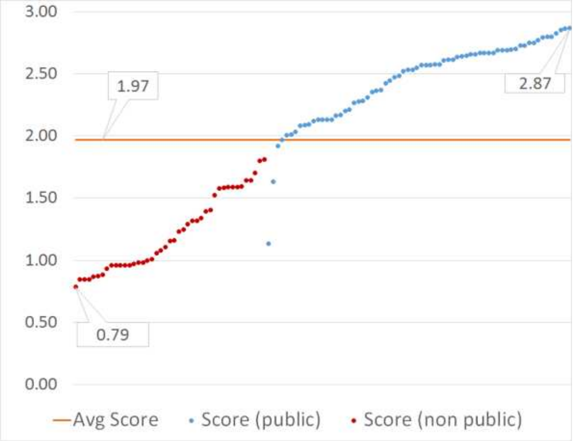


Figure 6. Profile score distribution

Figure 7 demonstrates the top 10 profiles with the highest score, which is a valuable source of information for the CMDI guidance. Data providers (or their data modellers) can go down this list ordered by the score, to find the most suitable profile to use to describe their resources, ensuring the best possible discoverability of their resources. It is also possible that the CMDI working group can learn the usage of CMDI profiles to discuss and develop the future version of the metadata model/framework. In addition, it is our recommendation that CMDI and VLO data ingestion guidelines can be produced and describe the recommendations of profiles according to the outcome of the Curation Module.

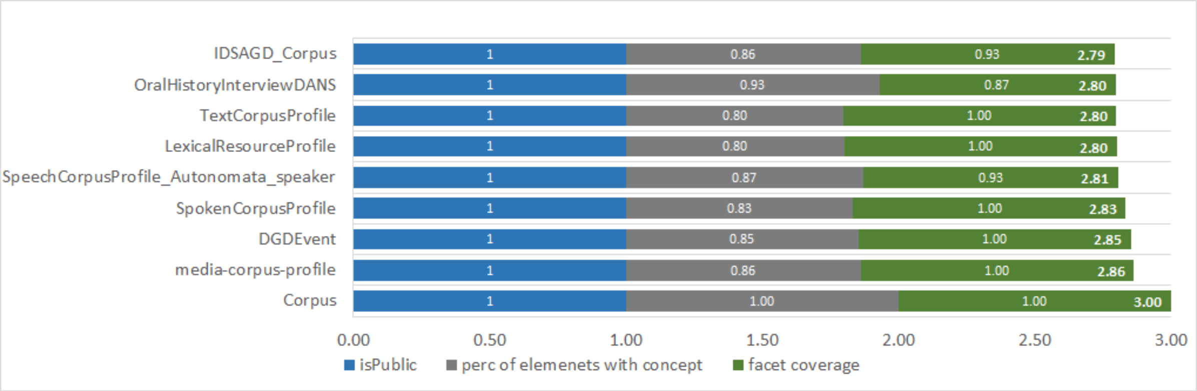


Figure 7. Public profiles with the highest score

Figure 8 shows the ten most referenced profiles by CMD instances/records. Interestingly, four out of ten are not public. In fact, as also discovered above, if the private profiles are changed to public, the scores would jump substantially with very little effort. For example, the Song profile, is associated with over 155.000 instances, albeit relatively low facet coverage (40% or 6 of 15). It contains the following facets: format, name, genre, nationalProject, collection, and languageCode.

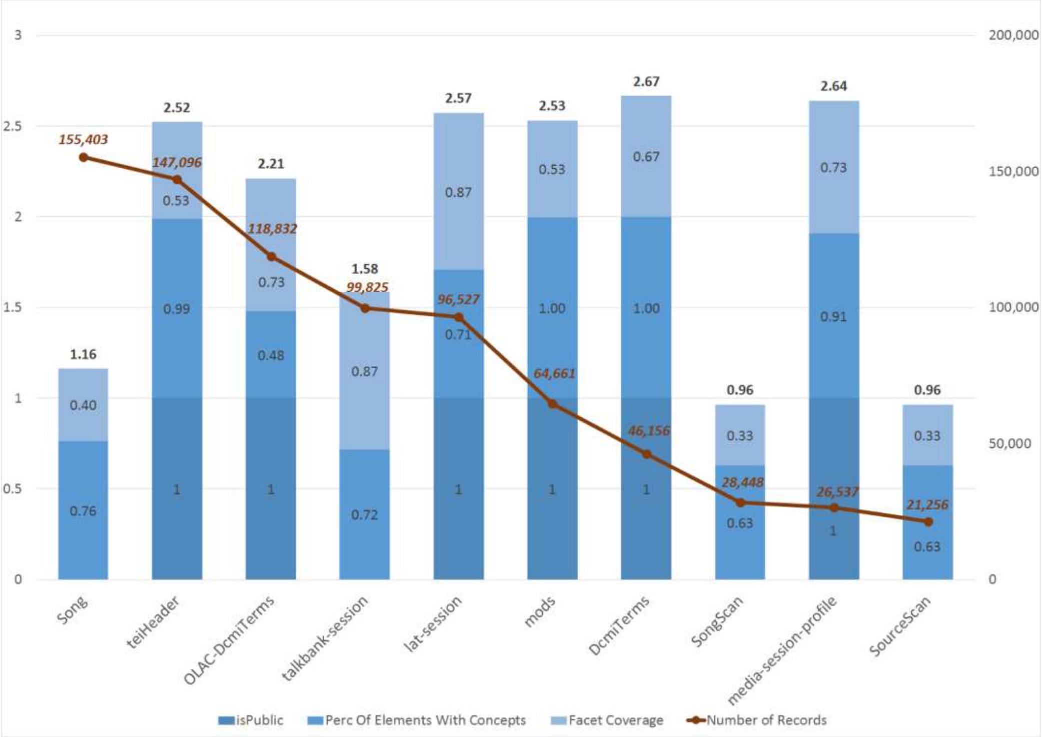


Figure 8. Profiles referenced in VLO with the highest number of records

In terms of the proportion of profiles, figures 9 suggests that approximately two thirds of the profiles are public. In addition, we counted the number of public and used profiles (figure 10). It turned out that about two thirds of public profiles are unused (or, more precisely, there are no records published via VLO). It is likely that when a (published) profile needs to be updated, the profile modeller/creator would abandon the old profile and create a new one, leading to a large number of unused public profiles. However, with CMDI 1.2 this issue has been tackled by establishing a lifecycle for profiles and components, including deprecation that should prevent further proliferation.

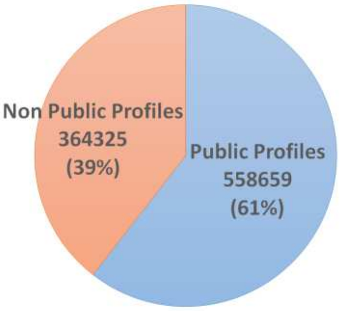


Figure 9. The number of instances using public and non-public profiles

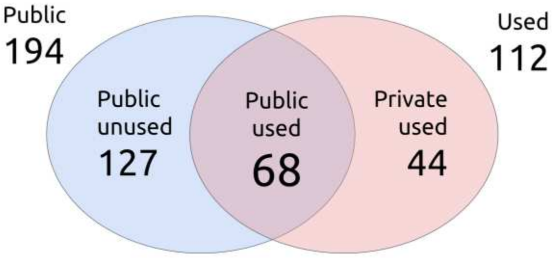


Figure 10. The proportion of public and used profiles

3.2 Collection analysis

The collections, as a set of the CMDI records gathered from the individual data providers, represent a primary discerning principle/dimension for the large body of CMD records harvested and indexed regularly by VLO. The reports for the collections, computed by aggregating the reports of instance assessment, are the most comprehensive report type, primarily relevant for the use case of a repository administrator checking the records that the repository exports to the VLO. Additionally, the contrasting juxtaposition of the metrics of all the individual collections reveals the overview on the statistics of the VLO records. For example, we can easily find that the number of records varies greatly from collection to collection, ranging from 1 to 249,659, and the size in bytes from 1 kB to 5 GB. While most collections use only a single profile, there is a collection using 36 distinct profiles.

The average score for all collections is 10.6 (out of 14.0) suggests a good overall quality (figure 11). However, the overall score distribution is rather concentrated in the area between 10.0 and 13.0, making it hard to differentiate the qualities of each collection. We have to examine the distribution more in detail to know whether the scores are too optimistic, thus not reflecting the intended results of reality, or not. The scoring criteria is always an area of discussion (Kemps-Snijders, 2014; Trippel et al., 2014), requiring a continuous reviewing and calibration. As such, the statistics may be used as rough indicators. The important point is that we now have a tool to automatically measure the quality of all data in a consistent and transparent manner.

Figure 12 illustrates the top 10 highly scored collections with reference to the data volume. It is of particular value to recognise that there are some large collections with low scores. By consulting the data providers and improving the metadata descriptions, we will be able to increase the overall quality of the data in the most efficient way. It is also the advantage of the Curation Module that it can indicate the best practice of (re-)defining profiles. The statistical overview available in the module is especially informative for CLARIN curation team, as they can determine what priorities and strategies should be taken in order to maintain and balance the quality of VLO in a short and long term.

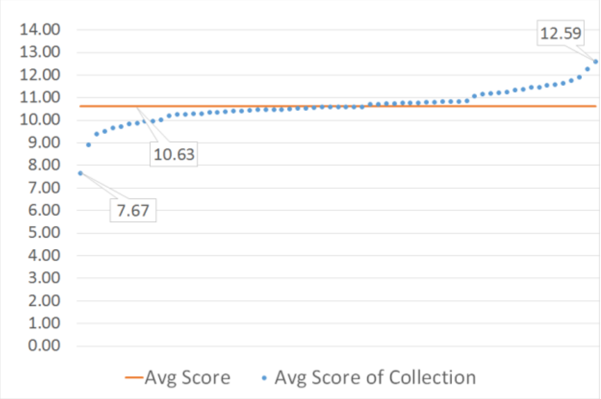


Figure 11. Collections score distribution

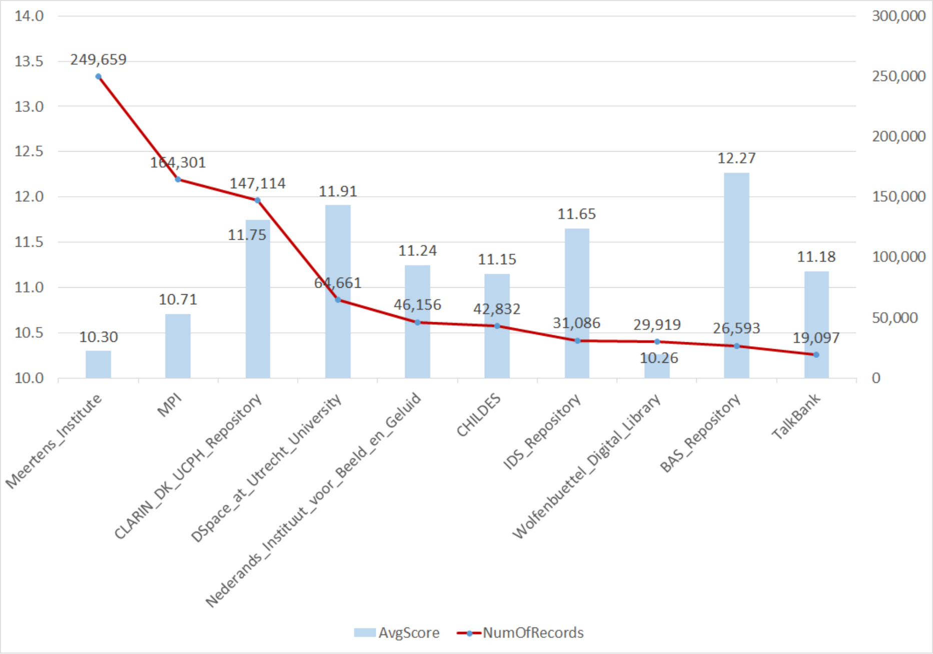


Figure 12. Top 10 highly scored collections with reference to the number of records

3.3 Combined analysis (facet coverage)

Probably the biggest headache of VLO is facet coverage. Echoed with Odijk (2014), King et al. (2016) argued that the extremely low percentage of facet coverage severely hampers the discoverability of language resources in the VLO. In this section, a closer examination on the distribution of covered and uncovered facets is executed in order to identify problematic metadata records. Firstly, our basic statistics of the Curation Module indicate the spread of facet coverage between 7.2% and 94.4%. The facets with the lowest coverage are keyword (7.2%), modality (13.6%), and country (13.6 %). Our analysis goes further to compare the current statistics with those in 2016 (figure 13). It has to be noted that there are a number of changes affecting the direct comparison. The data mapping of VLO (i.e. concept to facet, and value normalisation mapping) has been constantly modified and the VLO has received a large amount of records. Despite such changes, it is still useful to track the statistics over time. Good news is that the coverage of many facets is improved. For instance, languageCode facet (116.1%) and national project (90.5%) have improved dramatically, and the other facets such as availability, subject and resourceClass are in the range of 20%. There are only two facets whose coverage were deteriorated. Continent facet became obsolete in the meantime. The number of records without using format facet has increased from 9.9% to 40.6%, while collection facet has risen from 0% to 8.7%. Most likely reasons of those are the ingestion of a number of records which do not comply with the requirements of the VLO facets.

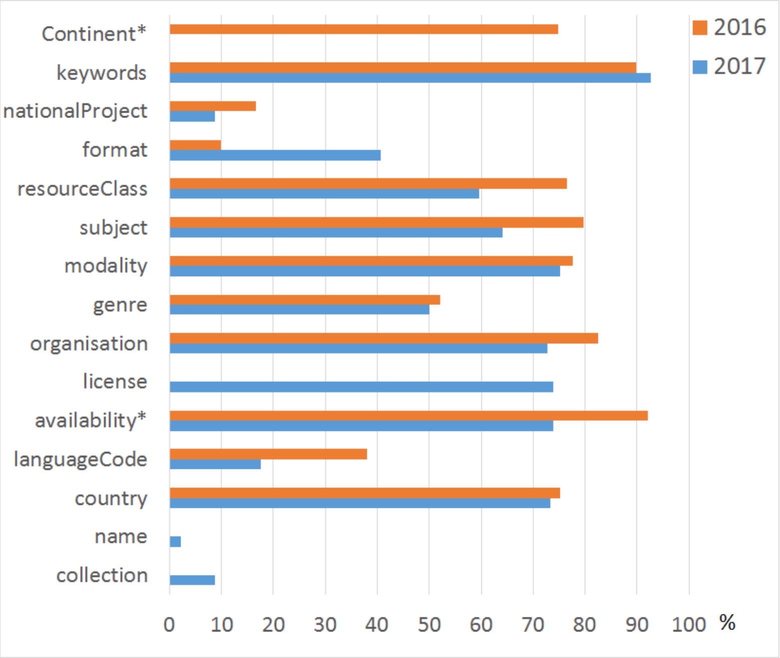


Figure 13. Comparison of uncovered facet of VLO (with data from King et al. (2016))

More interestingly, the collection and profile matrix helps us to generate new types of view on the data. Two figures were produced in order to investigate the relationship between the volume of collections (figure 14) and facets (figure 15) and the average facet coverage. It is clear that there are several collections and profiles which have a significant number of records with relatively low facet coverage (pink areas in the charts). Those can be regarded as the highest priority of improvement with least effort to increase the overall metadata quality of VLO. All in all, we think that the Curation Module is capable of providing a wide spectrum of feedback on the metadata quality.

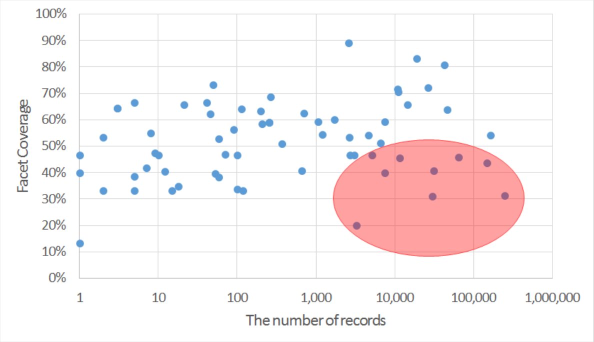


Figure 14. Facet coverage and size of collections

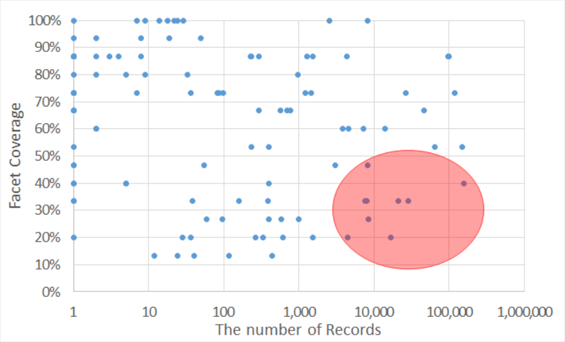


Figure 15. Facet coverage and size of profiles

4 Further developments

We have just finished the second phase of the development of the Curation Module. In the first phase (2015-2016), the Curation Module was provided with baseline functionalities, including the display of statistics with simple tables, the evaluation of a profile or instance by URL, and a simple link to the SMC browser. In the second phase (2016-2017), a series of small improvements were conducted according to internal reviews and feedback from CMDI team and early adopters. First of all, there are more functions to help curation tasks. For instance, comparison of original and normalised values for facets are now included, and links between values and concepts become traceable. Secondly, the output report was adapted with instructive information while web application presents it in a more user-friendly graphical interface. SMC Browser was also more integrated in the level of profiles and concepts. Thirdly, support for filtering and sorting the data in the overview tables for profiles and collections was added, allowing for more efficient exploration of large datasets. Moreover, while URL was the only input method for CMD records before, it becomes possible to upload a single local file. Finally, the table report of profiles and collections is exportable in a tabular format. It enables the users to further process and analyse the statistics with third-party applications. All such small improvement should not be underestimated, because they all contribute to the adaptation of the module and CMDI at large.

Although the two phases successfully delivered a stable service, there is still a wish list for the future, and some ideas are highlighted here. For example, batch upload for assessment can be developed so that the users can aggregate the statistics for given files and/or compare a set of uploaded files. It is also interesting to add more visualisation such as graphs and charts to the assessment report. In addition, automatic email notification with an attachment of (or link to) curation reports could be sent to the data providers. Such active effort would raise the awareness of the quality issue and hopefully encourage the providers to work on improving the data they delivered. Moreover, the calibration of the score should be considered (Kemps-Snijders, 2014). After careful manual revision of the quality reports, the curation team should be able to suggest a new weighting of the individual criteria to achieve a fair scoring of the metadata quality. Furthermore, the modularisation of the module and other VLO components could streamline the CLARIN service infrastructure for efficient maintainability, as the whole technical infrastructure gets bigger and more complicated.

Another major planned enhancement of the functionality includes the addition of time dimension and a facet-centred view. The former will store history of (primarily collection) assessments and thus enable us to monitor the data quality over time, introducing also a possibility to automatically identify sudden drop in any of the metrics. The latter will basically invert the current overviews of profiles and collections and allow us to explore how well (or badly) specific facets are covered relative to the collections and profiles. It will also feature information about the value variability, delivering much needed systematic input for the value normalisation efforts. If developed, those new features actually transform the Curation Module into a CMDI Analytics, which is similar to the concept of web analytics, to collect, monitor, and analyse the statistics of all stages of metadata management. It naturally allows the users to see a good visualisation with charts and to flexibly and seamlessly (re)generate the data and its matrices by manipulating different parameters of values and dimensions. Indeed, the Curation Module is also a part of a long-term vision of VLO backend development. As we suggested (King et al., 2016), we aim for the implementation of an integrated dashboard application which manages the whole processing of data within the aggregation infrastructure, ranging from data harvesting, converting, validating, mapping, normalising, to indexing. All those developments will be of added value for the VLO development team as well as the user evaluation initiatives and the assessment committee when evaluating new centres.

5 Conclusion

The Curation Module is clearly a big step forward. It does not only inform about the metadata quality, but also the level of collaboration. The idea of CMDI, one of CLARIN's pillar achievements, is to collectively develop a standard framework to aggregate heterogeneous metadata for language resources and tools. Therefore, the module objectively answers the question of how much CLARIN community has achieved together in terms of metadata aggregation. As we pointed out that various factors contribute to a number of problems in VLO (King et al., 2016), the module successfully demonstrates and supports them with detailed statistics and visualisations. In this sense, our first set of analyses outlined unprecedented views on the quality of CMD metadata. Although several issues and challenges are identified over time including the user interface, usability, input methods, data workflow, and the calibration of scoring algorithm, it is our mission to develop and maintain the Curation Module continuously, also in relation to a broader framework, the Dashboard, to reinforce the CMDI.

The Curation Module delivers a myriad of statistical facts about CMD instances, collections and profiles. That means they can be informative for various stakeholders. In the beginning, we considered different use cases. Most notably, the curation team and data providers would benefit from the detailed report on the delivered metadata. They can inform them of exactly what happened with the datasets during the process of metadata ingestion. It has a precaution and treatment function. It, on the one hand, can prevent the data providers from supplying low quality metadata, if they check it beforehand. On the other hand, it can report what went well or wrong after ingesting the metadata. In addition, the Curation Module can be used more widely from the very beginning of metadata creation to the future use of metadata. It helps the metadata modellers to compare different possibilities and select the right profile for the sake of metadata quality and accessibility to their valuable content.

There can be more use cases than initially defined. The module can be used for the research and development of the CMDI framework, because it gives the CMDI community a practical feedback on the actual use of CMDI, creating room for consideration for the future update of the CMDI. Moreover, the CLARIN community may want to have an annual report on the progress of data ingestion. In conclusion, the Curation Module supports all stages of metadata management that CLARIN has worked on, therefore, showing a potential to be transformed into a CMDI Analytics. It should, however, not be forgotten that the Curation Module itself does not do anything to improve the metadata. It has to trigger human actions. Nevertheless, we strongly believe that it fosters the analysis and improvement of metadata quality to support CMDI and VLO.

Reference

[Ďurčo 2013] M. Ďurčo. 2013. *SMC4LRT - Semantic Mapping Component for Language Resources and Technology*. (masters)Technical University, Vienna, Austria. http://permalink.obvsg.at/AC11178534

[Ďurčo and Mörth 2014] M. Ďurčo, and K. Mörth. 2014. Towards a DH Knowledge Hub - Step 1: Vocabularies. In *CLARIN Annual Conference* Soesterberg, Netherlands.

[Kemps-Snijders 2014] Kemps-Snijders, M. 2014. *Metadata quality assurance for CLARIN*. .

[King, Ostojic, Ďurčo, and Sugimoto 2016] M. King, D. Ostojic, M. Ďurčo, and G. Sugimoto. 2016. Variability of the Facet Values in the VLO–a Case for Metadata Curation. In *Selected Papers from the CLARIN Annual Conference 2015, October 14–16, 2015, Wroclaw, Poland* (pp. 25–44) Linköping University Electronic Press. http://www.ep.liu.se/ecp/123/003/ecp15123003.pdf

[Odijk 2014] J. Odijk. 2014. Discovering Resources in CLARIN: Problems and Suggestions for Solutions. http://dspace.library.uu.nl/handle/1874/303788

[Trippel, Broeder, Ďurčo, and Ohren 2014] T. Trippel, D. Broeder, M. Ďurčo, and O. Ohren. 2014. Towards automatic quality assessment of component metadata. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (pp. 3851–3856) Reykjavik, Iceland: European Language Resources Association (ELRA). http://lrec2014.lrec-conf.org/en/

ORTOLANG: a French Infrastructure for Open Resources and TOols for LANGuage

Jean-Marie Pierrel
University of Lorraine
CNRS
UMR ATILF
jean-marie.pierrel@atilf.fr

Christophe Parisse
INSERM
University of Paris-Ouest
Nanterre, UMR MoDyCo
cparisse@u-paris10.fr

Jérôme Blanchard
CNRS
University of Lorraine
UMR ATILF
jerome.blanchard@atilf.fr

Etienne Petitjean
CNRS
University of Lorraine
UMR ATILF
etienne.petitjean@atilf.fr

Frédéric Pierre
CNRS
University of Lorraine
UMR ATILF
frederic.pierre@atilf.fr

Abstract

ORTOLANG[1] (Open Resources and Tools for Language: www.ortolang.fr) is a French infrastructure which aims to ensure the management, pooling and sharing, dissemination and long-term preservation of language resources such as corpora, lexicons, terminologies and language processing tools, with particular focus on the languages of France. It will be used as a technical language platform for written and oral language forms. The ORTOLANG software platform is based on a new Digital Object Repository service. By combining a Service Oriented Architecture for high level services and a Software Component Architecture for its Repository Service, the platform seeks to build a robust and reliable Digital Object Repository that provides rich functionalities and a modern interface delivering excellent performances and the best optimization strategies. Thanks to its hardware and software architecture choices, the ORTOLANG platform ensures very flexible evolution possibilities to guarantee long-time support for the hosted resources.

1 Main characteristics of ORTOLANG

The ORTOLANG project is a French infrastructure implemented as part of the "Programme d'Investissement d'Avenir" (Investment program for the future) funded by the French Government.

This infrastructure aims to construct a network including a repository of language data (corpora, lexicons, dictionaries etc.) and readily available, well-documented tools for language processing. The repository was built following the guidelines of CLARIN repository centres, so that it could become a CLARIN centre if the opportunity arose and join the European effort to make language resources available. Following the decision of France to join CLARIN as an observer, we would like to rise to

[1] This research was funded by the French State program "Investissements d'Avenir" ORTOLANG managed by the Agence Nationale de la Recherche (grant reference: ANR-11-EQPX-0032).

Jean-Marie Pierrel, Christophe Parisse, Jérôme Blanchard, Etienne Petitjean and Frédéric Pierre 2017. ORTOLANG: a French infrastructure for Open Resources and TOols for LANGuage. *Selected papers from the CLARIN Annual Conference 2016*. Linköping Electronic Conference Proceedings 136: 102–112.

the task and become a CLARIN B-Centre. The current paper presents the status of ORTOLANG as it is, before finalizing this process.

1.1 Strong emphasis on multidisciplinary openness

The ORTOLANG project is underpinned by a consortium of laboratories and resource centres with complementary expertise in the following fields:

- linguistics with ATILF (*Analyse et Traitement Informatique de la Langue Française* – Computer Processing and Analysis of the French Language: www.atilf.fr), LPL (*Laboratoire Parole et Langage* – Speech and Language Laboratory: http://www.lpl-aix.fr), MoDyCo (*Modèles, Dynamiques, Corpus* - Models, Dynamics, Corpora : http://www.modyco.fr) and LLL (*Laboratoire Ligérien de Linguistique* – Loire Valley Linguistics Laboratory : www.lll.cnrs.fr);
- information technology with LORIA (Laboratoire lorrain de Recherche en Informatique et ses Applications - Lorraine Research Laboratory in Computer Science and its Applications: www.loria.fr) and INIST (Institut de l'information scientifique et technique - Institute for scientific and technical information www.inist.fr), but also partly with ATILF and LPL;
- data base management and management of access to scientific information, through INIST, and to linguistic resources, through CNRTL (*Centre National de Ressources Textuelles et Lexicales* - French National Centre for Textual and Lexical Resources: www.cnrtl.fr) [Pierrel and Petitjean 2007] and SLDR (Speech & Language Data Repository: http://www.sldr.org/) [Bel and Blache 2006].

Our aim is not only to combine expertise from different disciplines, but also to bring together – within this infrastructure for the sharing of language resources and tools – partners who represent the diversity of approaches to language study: constructing linguistic models, experimental and/or applied linguistics, language production and perception, diachronic studies, sociolinguistics, and the automatic processing of languages (written, oral and multimodal).

ORTOLANG draws on the wealth of experience gained by the teams supporting the infrastructure:

- the existing means of partners, resource centres (CNRTL and SLDR) and laboratories who offer a set of available resources and tools, and whose expertise covers the three main aspects targeted: oral language, written language and the preservation of the French heritage of languages;
- involvement in and coherence with TGIR Huma-Num (*Très Grande Infrastructure de Recherche* - Very Large Facility in Humanities and Social Sciences: www.huma-num.fr);
- coherence with the European infrastructure CLARIN (we worked as part of CLARIN during the preliminary phase);
- and finally, coherence with the efforts led by DGLFLF (*Délégation à la Langue Française et aux Langues de France* - General Delegation for the French language and languages of France: www.dglf.culture.gouv.fr) and BNF (*Bibliothèque Nationale de France* – National French Library: www.bnf.fr) concerning the heritage aspects of the languages of France.

1.2 An infrastructure that manages resources for the whole scientific community

The ORTOLANG platform is intended to be an infrastructure for the management, pooling and sharing, long-term preservation and dissemination of language corpora, lexicons, terminologies and tools, which of course remain the property of the depositors (researchers or laboratories). Access rights to these resources thus continue to be defined by their owners. On this point, however, ORTOLANG has made the following strong recommendations:

- compliance with the *Ethics & Big Data Charter*,[2] drawn up through the collective efforts of several players engaged in the creation, dissemination and use of data;
- freedom of use for research, provided there is no commercial utilization;
- prior negotiation with the resource owners, whenever there is a desire for commercial exploitation.

[2] http://wiki.ethique-big-data.org.

All the data deposited in ORTOLANG must be made available to the whole public research community. This is mandatory as ORTOLANG is a free public service for research and visibility of the data is requested in exchange for this free service. However, data can be deposited and stored in private workspaces (see part 3.4) for the duration of a project, for example a PhD Thesis or any funded project. This duration is limited in time and cannot be extended more than one year after the end of the thesis or the project. Exceptions to this rule can be made in very special situations, following the guidelines of the linguistic consortia from Huma-Num. In this case, the data will be made public according to the principles of the French public archiving system.

With these points in mind, several operations have been set up with partners outside the ORTOLANG consortium who have deposited, or wish to deposit, their resources on ORTOLANG. They include:

- linguistics consortia (Huma-Num) – 'Corpus Ecrits', IRCOM (Corpus Oral and Multimodal**)** and more recently CORLI (Corpus Languages and Interactions: https://corli.huma-num.fr/) - through common calls for projects for the finalization and standardization of corpora;
- the French linguistics research federations ILF (*Institut de la Langue Française* : www.ilf.cnrs.fr) and TUL (*Typologie et Universaux Linguistiques* : www.typologie.cnrs.fr). ORTOLANG is thus being used as a medium for the "French reference corpus"[3] initiative of ILF.

2 Objectives and missions of the infrastructure

The objectives and missions of ORTOLANG can be split into three complementary aspects: identification and preparation of data, long-term preservation of the resources and dissemination.

2.1 Identification and preparation of data

At present, one of the difficulties faced in identifying and accessing resources (corpora, lexicons, terminologies and processing tools) stems from their considerable dispersion and the great disparities between them, particularly in terms of coding. Furthermore, over the last twenty years, many language resources of high quality, developed for research projects or theses, have been lost because of a failure to rigorously manage this heritage. This is why the primary objectives are:

- the finalization and standardization of existing resources and tools, with a view to their pooling and sharing. This action is being carried out in close collaboration with the consortia Corpus Ecrits, IRCOM and now CORLI of TGIR Huma-Num. To generate this kind of sharing momentum and extend it to teams outside the consortium, we have set up funding through calls for common projects with the linguistic consortia of Huma-Num so as to support the necessary work on the standardization of resources that teams outside the consortium wish to deposit on the ORTOLANG platform;
- the control and validation of resources and tools, including in particular support for the authors of resources about current standards, metadata, norms and international recommendations, such as XML (Extensible Markup Language), TEI (Text encoding Initiative: www.tei-c.org), LMF (Lexical Markup Framework: www.lexicalmarkupframework.org).

2.2 Long-term preservation of resources

To ensure the long-term preservation of resources, we have implemented three types of actions:

- the curating of resources and tools;
- secure storage and maintenance of resources;
- long-term archiving, using the solution set up by TGIR Huma-Num[4] in conjunction with CINES.[5]

[3] http://www.ilf.cnrs.fr/spip.php?rubrique95.

[4] http://www.huma-num.fr/services-et-outils/archiver

[5] Centre Informatique National de l'Enseignement Supérieur : https://www.cines.fr/

2.3 Dissemination

Finally, to ensure the necessary dissemination and exploitation of resources, we offer aid and support to users for setting up procedures enabling platform users to exploit the shared resources and tools by drawing on the prior experience of the resource centres CNRTL and SLDR (which are set to be fully merged into ORTOLANG).

3 Hardware and software architecture

3.1 Hardware

ORTOLANG has a hardware architecture specifically designed for the purposes of this project (cf. https://dev.ortolang.fr/doc.infrastructure.html) and recently upgraded. The ORTOLANG hardware architecture is based on a dedicated cluster of computing servers, a SAN (Storage Area Network) and an automated tape backup system (LTO6). The entire software platform is hosted on a virtualized environment solution (VMWare) in order to provide complete flexibility (CPU, RAM and storage dynamic allocations) to better suit each service need. Internet connectivity is ensured using two redundant connections (10Gb/s and 1Gb/s) and two firewalls. The internal network connectivity between servers, Storage Area Network and backup system uses up to 12 Fiber Channel links. The whole infrastructure is hosted and operated by INIST, one of the members of the ORTOLANG consortium.

Our current hardware equipment is composed of:

- a cluster of six servers: three DELL R620 servers (48 cores – 768 GB of RAM) and three DELL R630 servers (60 cores – 1152 GB of RAM);
- 165 useful TB of disks in Raid 6;
- a back-up system based on a Quantum library with two LTO6 readers and fifty 300TB slots.

3.2 Software architecture

The goal of the ORTOLANG Diffusion Service is to build a robust and reliable Digital Object Repository. It is based on a Service Oriented Architecture for high level services and a Software Component Architecture for its repository service. This diffusion service will fully comply with the recommendations of CLARIN for the resource centres in the near future. The service is connected directly to the website www.ortolang.fr, enabling users to browse through resources or to select resources via metadata requests. The software architecture of this platform is described below in section 4. ORTOLANG is accessible via various Application Programing Interfaces (APIs): REST [Richardson & Ruby 2007], OAI-PMH,[6] Handle Persistent Identifier, FTP.[7] Some components are accessible via multiple interfaces. We provide a REST interface for most of the operations on workspaces and other components of the platform. We provide more specific interfaces such as an FTP connection on workspaces in order to upload very large files or numerous files at once. We also manage an OAI-PMH interface of published resources and the Handle Persistent Identifier on each file that is published. Our implementation is free and open-source (LGPLv3[8]). The source code is available online from an open source software repository (see https://www.openhub.net/p/ortolang).

3.3 A CLARIN-compatible dissemination centre

The lower layer of the ORTOLANG software architecture (the dissemination centre) complies with the constraints of quality of service (maximum availability) and document management meeting Data Seal of Approval (DSA) requirements. The infrastructure, which is largely invisible to users, is a reliable data warehouse (corpora, lexicons, terminologies and language processing tools) incorporating the following functions:

[6] Open Archives Initiative - Protocol for Metadata Harvesting: https://www.openarchives.org

[7] File Transfer Protocol: https://www.w3.org/Protocols/rfc959/

[8] https://www.gnu.org/licenses/lgpl-3.0.fr.html

- identification of each resource by means of a Handle Persistent Identifier;
- proof of integrity of the data associated with a Handle by means of a checksum linked to the Handle;
- metadata: OAI-PMH, OAI Dublin-Core, OLAC,[9] CMDI,[10] RDF;[11]
- version management: any modification of data leads to a new version;
- authentication of users via a Single Sign On mechanism, using the Education-Research federation of RENATER (Réseau National de télécommunications pour la Technologie l'Enseignement et la Recherche: National telecommunication network for technology, teaching, and research) to authenticate users requesting access to restricted data.

3.4 A user-friendly interface for depositing and consulting resources

Special efforts have been made to offer an interface and workspaces that provide depositors with a flexible procedure that is as user-friendly as possible, to enable non-IT specialists to easily deposit their resources and draw attention to them.

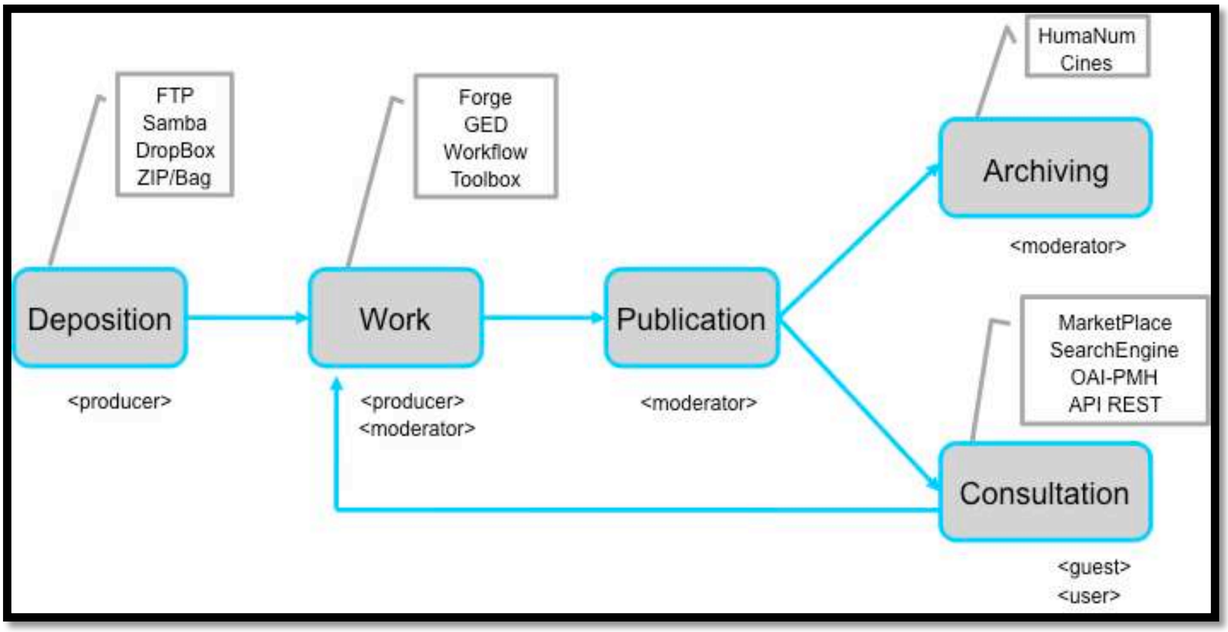


Figure 1: ORTOLANG deposition workflow chart

With this objective in mind, we propose a 5-stage workflow (see Figure 1):

- Depositing: After opening an online workspace, the producer is provided with a simple means of depositing the data, even if they are not yet ready for publication. Various methods are proposed for deposition or uploading: via FTP, via a Web interface, or by uploading compressed files. As soon as resources are deposited, they are secured by the use of reliable media (redundancy) and by daily back-up copies on tape.
- Working in a secure workspace: The producer is provided with specific online tools so that metadata can be edited in a user-friendly way and the work can be enriched (alignment, annotation, etc.). Metadata can be edited at the general level (information about the whole project) and can also be further specified at the document level (information and rights about specific documents, for example). Metadata can be described in French and in English. During this work phase, access to data is controlled, and data are only visible to the workspace members and platform administrators. Furthermore, resource producers can take

[9] http://www.language-archives.org/

[10] https://www.clarin.eu/content/component-metadata

[11] https://www.w3.org/RDF/

advantage of support from three expertcentres specializing in written (ATILF/CNRTL), oral (SLDR & Modyco), and multi-modal (SLDR & Modyco) data.

- Publishing: once the data are ready, the producer can submit the work for publication. The producer can then monitor the status of his/her requests, and – in collaboration with the administrators – achieve a stable version of the resource.
- Archiving: ORTOLANG is not responsible for the archiving process itself. This is handled by Huma-Num and CINES. ORTOLANG will submit the published data for archiving. Automatic data enrichment during earlier phases means that the data are "clean" and the archiving format has been checked. All the data that conform to the archiving conditions of the CINES will be forwarded for public archiving to Huma-Num and CINES.
- Consulting and reusing: Data can be consulted in various ways: via a Web interface that lists all the published resources, which are split into categories and described with detailed metadata. Online browsing through the content of resources is also possible. References can be added to published data in a new workspace.

4 Software organisation

4.1 Initial requirements

Building on the experimental work we carried out in 2011 in CLARIN, we analyzed the characteristics of the CLARIN network of centres, [12] looked more specifically at specific centres such as LINDAT/CLARIN,[13] and at the CLARIN B Centre checklist.[14] Besides, we analyzed the needs of the French research community and defined use-case scenarios for such a platform, as described in the previous section. Our requirements also included using open source and free third-party software, performing data de-duplication, setting fine-grained access control rules, giving scalability to millions of objects and ensuring a fast response time.

In 2011, when we first started work on a digital object repository, we took the time to analyze and test existing solutions to find a platform that would meet our requirements. Back then, we tested two different software solutions. The first one, Fedora,[15] was an interesting platform and we started developing a proof of concept demo that would lay the groundwork for a larger digital repository. Unfortunately, we encountered many technical challenges that were hard to overcome. The underlying software lacked flexibility and would not scale up to the requirements we had set ourselves. The second platform we analyzed was D-Space.[16] But at the time, no serious work appeared to be ongoing and the software development stalled.

These reasons drove us to the conclusion that to meet our requirements, we needed to start developing our own software architecture that would create a solid Digital Object Repository.

4.2 Software architecture

In order to ensure maximum flexibility and maintenance, we chose a Service Oriented Architecture pattern to design the software architecture. The application relies on six Top Level Services and a farm of dedicated business specific services. The main repository service (ORTOLANG Diffusion) is designed using a Software Component Architecture and implemented using JEE (Java Enterprise Edition) Technologies.

4.3 High level services

All six top level services and the tools farm are hosted independently in the cluster and are based on different kinds of software. These services are connected using very simple protocols (mostly HTTP[17]) to ensure loose coupling, easy deployment, load balancing and scalability.

[12] https://www.clarin.eu/content/overview-clarin-centres

[13] https://lindat.mff.cuni.cz/en

[14] hdl:1839/00-DOCS.CLARIN.EU-78

[15] http://fedorarepository.org/

[16] https://www.dspace.com

[17] Hypertext Transfer Protocol: https://www.w3.org/Protocols/

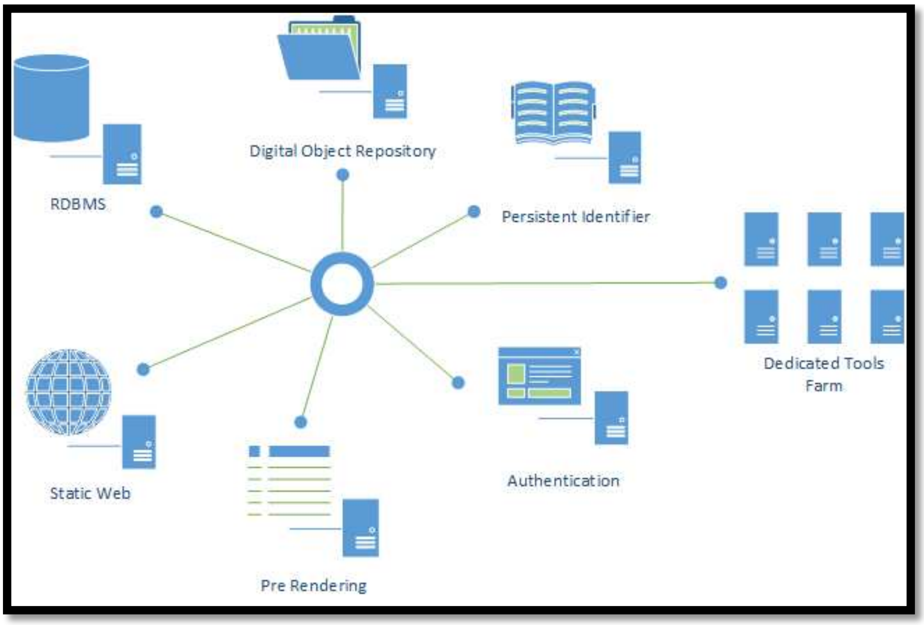


Figure 2: ORTOLANG High Level Services

Relational database management system: We use a relational database management system (PostgreSQL[18]) to store all the service data that require transaction isolation. Because of the very large volume of binary data, we avoid storing binary content in the database but only critical data that require ACID (Atomicity, Consistency, Isolation, Durability) properties [Haerder and Reuter 1983]. This service is hosted in its own virtual machine and can be scaled either by increasing VM capacity or by using clustering strategy.

Authentication: As we need an external authentication mechanism, we chose the OAuth protocol.[19] This protocol allows authentication (Direct Grant) without the need for a user interface.

To do so, we use Keycloak,[20] an open source identity management component developed by Jboss.[21] Keycloak provides connectors for most of Java Application Servers but also for Javascript applications. To handle RENATER integration, we developed a small application that plays the role of the Shibboleth Discovery Service.[22]

For users who are not registered in the French authentication system (or in the future European system), it is possible to access the data that are free (which are a large part of the ORTOLANG data) or to use a classical identification with username and password. This identification does not provide the same rights as the institutional authentication.

Repository: The digital repository service aims to manage user resources from the first deposit to the final publication providing specific features for each phase of the resource life cycle. This service has been developed internally in order to meet our requirements but relies on some low-level software components that have been embedded in the repository application to propose functionalities such as File Storage, Version Configuration, Content Management.

Persistent identifier: We provide persistent resource identifiers based on the Handle system. We have packaged the Handle.net server software as a platform service but in read-only mode. Thus, it is up to the repository application to create and update handle entries in the database to ensure consistency. This is achieved by including Handle write operations in a global transaction in the repository service.

[18] https://www.postgresql.org/

[19] https://oauth.net/

[20] http://www.keycloak.org

[21] http://www.jboss.org/

[22] http://shibboleth.net/

Web server: Client side applications (repository browser and administration) are developed using HTML5/Javascript and use the AngularJS framework.[23] These applications only need to be served as static web content. We use Nginx[24] to serve these applications but also to do a specific routing of requests coming from search engine indexers to a dedicated service.

Pre-rendering: In order to ensure the best search engine indexation, this service provides static versions of the website pages. It acts like a Client Web Application browser and stores the rendered pages in order to serve them directly to the search engine indexer bot, thus avoiding Javascript interpretation side effects.

Tools farm: Specific applications can interact with the repository using its REST API. This allows anybody to develop their own application (for example a specific file format conversion application) and submit this application as a tool for ORTOLANG. These applications are self-contained and must provide their own user interface. Authentication is done using OAuth and applications can access user information and data only if authorized by the user. Using this mechanism ensures that each application has permissions granted by a user to be able to access data in the repository.

4.4 Repository service architecture

The digital object repository service (ORTOLANG Diffusion) business logic is complex and is the result of merging the logic of many existing components. It provides a virtual online versioned file-system (like DropBox) for each resource: a workspace. Around this workspace, we provide the ability to enrich content by setting some metadata using the format provided for all types of content (files, folders, and resources). These metadata will be indexed to populate an internal search engine and to give visibility to the published workspace in a kind of market place that will present all the published resources. Collaborative functionalities and publication processes allow a group of people to work on the same resource before and after its publication. All the business logic is defined in dedicated components that are exposed through interfaces providing a consistent repository service.

Implementation is done in a Java JEE application using EJB (Enterprise JavaBeans), JPA (Java Persistence API) and JMS (Java Message Service) to ensure the robustness and stability of the platform. Some EJB components wrap subsystems that are completely embedded in the platform such as a BPEL (Business Process Execution Language) Engine,[25] a NoSQL Database[26] or a Lucene index base.[27] This component wrapping avoids coupling between platform components and embedded ones making it possible to switch to any other implementation. That's the key point of an SCA Architecture. Each component is also testable independently using mock strategy.

Main principles: In order to avoid coupling between software components, we have based the identification of all objects in the repository on a unique key managed by a registry. All operations are performed using the key of an object. The registry maintains the association between a key and the concrete object in the database (group, workspace, collection, process) by mapping an object identifier to its registry key. An object identifier is composed of its service name, its type name and its internal id. The registry manages some common aspects of all concrete objects e.g. the state, lock, author, properties, history, etc.

[23] https://angularjs.org/

[24] http://nginx.org/en/

[25] http://www.activiti.org

[26] http://orientdb.com

[27] https://lucene.apache.org/

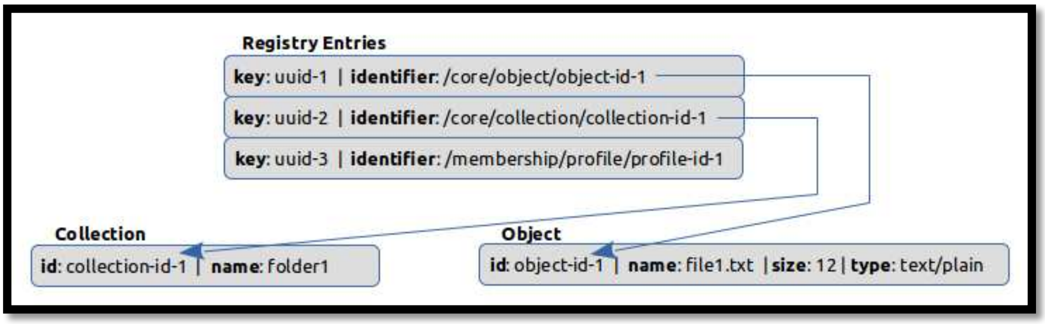

Registry Entries
key: uuid-1 | identifier: /core/object/object-id-1
key: uuid-2 | identifier: /core/collection/collection-id-1
key: uuid-3 | identifier: /membership/profile/profile-id-1
Collection
id: collection-id-1 | name: folder1
Object
id: object-id-1 | name: file1.txt | size: 12 | type: text/plain

Figure 3: Registry Service Mapping

Data granularity choices: We have opted for fine-grained objects, thus the smallest object in the system is a simple file. Objects can also be larger, and as big as a complete part of a file system. This has a very large side effect because each object that is referenced in the registry can have its own history, owner, security rules, index entry, publication state, etc. We chose to place granularity at the lowest level possible but we also defined a structure to organize files into folders (collections) and folders into workspaces using a one to many relation.

Workspace and version control: Most of the repository business logic is organized around a structure that we call a workspace and that holds the content of a complete resource. A workspace, like a versioned file-system, manages folders and files and keeps track of their modifications; it also manages metadata to describe the resource and its content. Once a producer assumes that his workspace content has reached a sufficient level of maturity or needs to be visible by external users, the content can be submitted to the publication process. Multiple versions of the same resource can be published to follow the resource's life cycle.

Binary content aspects: Binary content associated with files is not stored as is. A dedicated component is used to generate a hash (SHA-1: Secure Hash Algorithm 1[28]) for each binary stream stored and organizes the physical storage on an underlying file system that can be shared in multiple volumes. Using a hash SHA-1 as identifier for stored streams means that de-duplication of identical binary content can be performed.

Metadata: It is possible to set metadata on any object in a workspace using a one to many named relation. Metadata are typed, and for some particular types, content is structured using a schema.

We use the JSON (JavaScript Object Notation) data format[29] for structured metadata. These structured metadata are indexed in a dedicated NoSQL database (OrientDB[30]) in order to produce an enriched database that allows to search for objects based on their characteristics and using a rich query language. The metadata stored in the system can be converted to an adequate format for metadata harvesting such as OLAC Dublin Core used by Isidore (a French search platform allowing access to digital data in the Humanities and Social Sciences: https://www.rechercheisidore.fr/) or CMDI used by CLARIN.

Security concerns: A set of security rules is defined for each registry key and enables specific permissions to be stored on each object. Security is enforced by a dedicated component that can be queried for a specific permission (read, update, delete) on a particular key.

Asynchronous treatments: We use Java Messaging Service in order to handle asynchronous jobs. We have dedicated a topic in which all platform events are fired. Some listeners are in charge of triggering actions on event reception (indexing a file, extracting metadata, notifying, and logging).

Process management: In order to manage the publication and review processes, we have defined a runtime component that can handle Business Process Model and Notation scripts.[31] This component is a wrapper over a well-known BPEL Engine (Business Process Execution Language): Activiti. BPEL

[28] DOI: http://dx.doi.org/10.6028/NIST.FIPS.180-4

[29] http://www.json.org/json-fr.html

[30] http://orientdb.com/orientdb/

[31] http://www.bpmn.org/

processes are injected into the runtime component which allows transactional processing and human tasks management.

Search engine: We use asynchronous indexers for search functionalities. A key/value base (Apache Lucene[32]) is designed for full text searches whereas the NoSQL base allows more specific queries and faceted results. All search results are security filtered to ensure privacy.

API (REST, OAI-PMH, Handle, FTP): Some components are accessible via multiple interfaces. We provide a REST interface for most of the operations on workspaces and other components of the platform. We provide more specific interfaces such as an FTP connection on workspaces in order to upload very large files or numerous files at once. We also manage the OAI-PMH interface of published resources and Handle Persistent Identifier on each key that is published.

Performances: We have carried out performance tests on the repository that show that we are able to store more than 25 million objects without significant response time modification.

5 Tools farm ecosystem

We plan to open a tools farm to provide specific processing for resources. Some specific language treatment tools (tokenizer, text extraction, speech and text alignment, etc.) will be integrated as external applications and, relying on the OAuth grant permission mechanism, will access ORTOLANG resource files in a secure way, even before publication of the content. These tools will help resource providers to work on their files before publication but will also allow users to apply some treatments on a selected set of ORTOLANG resources.

We have already produced a proof of concept by integrating TreeTagger [Schmid 1994] as an external tool for ORTOLANG. We are about to release a file conversion tool (avconv[33]) and a concordancer allowing indexation of a specific set of files for dedicated search.

At the same time, we are working on a sample web application that will make new tool integration easy by customizing this application connected to the ORTOLANG Repository. In the best case, the customization will consist in writing a simple HTML form and a mapping between this form's values and a shell command line.

All the tools will be hosted in their own server, either on the ORTOLANG cluster or on an external server and will use the ORTOLANG REST API.

6 Achievement of the project

As of January 2017, the first phase of the project is finished. Improvements to the infrastructure, especially the user interface, are still ongoing, to take into account the actual use of the infrastructure by the researchers. Some software developments are still in the final phase and will be included in the project in 2017.

New developments are scheduled and should be finalized during the second phase of the EQUIPEX project. Our target concerns mostly the enrichment of resources and tools. The goals include the development of a concordancer that processes large volumes and can be used on any written language corpus, the enrichment of a French morphosyntactic lexicon, the development of an oral corpus transcription aid tool, the development of plugins to enable interoperability between the various editing and annotation tools, and the standardization of various corpora including COLAJE [Morgenstern and Parisse 2012], L'Est Républicain [ATILF 2011], ESLO [Eshkol-Taravella et al. 2012], PFC [Durand et al. 2009], and TCOF [ATILF 2017]. Some tools are already available as independent tools in the "Tools" section of the ORTOLANG main repository.

At the time of writing, the website (www.ortolang.fr) offers access to a constantly growing set of resources with possibilities of searching for a resource based on standardized metadata (resource type, language, rights of use, source, coding format and annotation types). At the end of March 2017, the platform hosted almost 189 corpora, 14 lexicons, 21 terminologies, 27 processing tools and several integrated projects, such as the CNRTL lexical portal (http://www.cnrtl.fr/portail/) serving more than 600,000 queries a day (http://www.cnrtl.fr/aide/stat/). This corresponds to a total 4.9 To of data and more than 300,000 files.

[32] https://lucene.apache.org/

[33] https://libav.org/avconv.html

7 Conclusion

After more than two years of effort, we have deployed a dedicated hardware cluster that hosts the ORTOLANG platform, and have developed a new Digital Object Repository in accordance with our needs and requirements. ORTOLANG already acts as a robust choice to deposit resources and metadata.

The final objectives of our work are to comply with the guidelines of the Data Seal of Approval and to comply with the technical requirements defined for CLARIN centres.

As France has joined the CLARIN ERIC with observer status, our current goal is to become a CLARIN B-Centre (http://clarin.eu/content/centres) by the end of 2017.

References

[ATILF 2011] ATILF 2011. Corpus journalistique issu de l'Est Républicain [Corpus]. ORTOLANG (Open Resources and TOols for LANGuage) - www.ortolang.fr, https://hdl.handle.net/11403/est_republicain/v1, https://hdl.handle.net/11403/est_republicain/v1

[ATILF2017] ATILF 2017. TCOF: Traitement de Corpus Oraux en Français [Corpus]. ORTOLANG (Open Resources and TOols for LANGuage) - www.ortolang.fr, https://hdl.handle.net/11403/tcof/v1.

[Bel and Blache 2006] Bernard Bel and Philippe Blache. 2006. Le Centre de Ressources pour la Description de l'Oral (CRDO). Travaux interdisciplinaires du Laboratoire parole et langage d'Aix-en-Provence (TIPA), vol. 25. 2006, p. 13-18.

[Durand and al. 2009] Jacques Durand, Bernard Laks & Chantal Lyche. 2009. Le projet PFC: une source de données primaires structurées. In J. Durand, B. Laks et C. Lyche (eds)(2009) Phonologie, variation et accents du français. Paris: Hermès. pp. 19-61, https://hdl.handle.net/11403/pfc/v1.

[Eshkol-Taravella et al. 2012] Eshkol-Taravella I., Baude O., Maurel D., Hriba L., Dugua C., Tellier I. 2012, Un grand corpus oral « disponible » : le corpus d'Orléans 1968-2012., in Ressources linguistiques libres, Traitement Automatique des Langues. Volume 52 – n° 3/2011, 17-46, https://hdl.handle.net/11403/eslo/v1

[Haerder and Reuter 1983] Theo Haerder, Andreas Reuter, A. (1983). "Principles of transaction-oriented database recovery". ACM Computing Surveys. 15 (4): 287

[Morgenstern and Parisse 2012] Aliyah Morgenstern and Christophe Parisse 2012. The Paris Corpus, *Journal of French Language Studies,* 22/01, 7–12, https://hdl.handle.net/11403/colaje/v1.1.

[Pierrel and Petitjean 2007] Jean-Marie Pierrel et Etienne Petitjean. 2007. Valorisation et exploitation scientifiques de documents numériques pour la recherche en linguistique : l'exemple du CNRTL, *Actes de CIDE 2007 Congrès International sur le Document Numérique*, Nancy, 2-4 juillet 2007, p13-24, Europia 2007, ISBN 978-2-909285-38-2

[Richardson and Ruby 2007] Leonard Richardson and Sam Ruby. 2007. RESTful Web Services, O'Reilly Media, 454 p.

[Schmid 1994] Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of International Conference on New Methods in Language Processing, Manchester, UK.

Conversion and Annotation Web Services for Spoken Language Data in CLARIN

Thomas Schmidt
Institute for the German Language (IDS)
Mannheim
Germany
thomas.schmidt@ids-mannheim.de

Hanna Hedeland
Hamburg Centre for Language Corpora (HZSK)
University of Hamburg
Germany
hanna.hedeland@uni-hamburg.de

Daniel Jettka
Hamburg Centre for Language Corpora (HZSK)
University of Hamburg
Germany
daniel.jettka@uni-hamburg.de

Abstract

We present an approach to making existing CLARIN web services usable for spoken language transcriptions. Our approach is based on a new TEI-based ISO standard for such transcriptions. We show how existing tool formats can be transformed to this standard, how an encoder/decoder pair for the TCF format enables users to feed this type of data through a WebLicht tool chain, and why and how web services operating directly on the standard format would be useful.

1 Introduction

Web services operating on language resources are a central idea for the CLARIN infrastructure. This includes services for the annotation of text data, such as lemmatizers, Part-Of-Speech-Taggers, Named Entity Recognizers, and so forth. WebLicht (Hinrichs et al., 2010) is an application that integrates many such services into a common web-based framework and enables users to build and apply annotation chains for a given set of data. So far, most such services were built with, and are meant to operate on, “canonical” written language data, typically edited texts from newspapers, books, etc., in the standard orthography of a major language. For “non-canonical” types of written data (such as CMC data[1], or historical language data[2]) and for spoken language data, these services are often not directly usable for at least two reasons:

a. The data come in formats which are more complex than the simple “stream of tokens” (see Menke et al., 2015) expected by many annotation services. They can contain transcriptions of simultaneous (“overlapping”) speech or two or more alternative transcriptions for the same stretch of speech (if the transcriber is uncertain), both of which require parallel structures to be encoded in the data format.

b. The text data may require additional processing steps or adaptations of annotation methods in order to yield useful results. Not all “tokens” of a spoken language transcription are simple words or punctuation. We also find descriptions of pauses or non-verbal actions, which may have to undergo additional processing before they can be fed into, say, a Part-Of-Speech-Tagger. The use of non-standardized writing (as in “modified orthography” to represent pronunciation deviating from the standard), the lack or diverging use of punctuation (as in systems which use punctuation to represent prosodic properties of speech), semi-lexical material (like hesitation markers or interjections), or incomplete tokens (as in a repair sequence or an aborted utterance) may cause similar problems.

[1] See: http://de.clarin.eu/en/curation-project-1-3-german-philology.

[2] See: http://www.deutschestextarchiv.de/doku/software.

Thomas Schmidt, Hanna Hedeland and Daniel Jettka 2017. Conversion and annotation web services for spoken language data in CLARIN. *Selected papers from the CLARIN Annual Conference 2016.* Linköping Electronic Conference Proceedings 136: 113–130.

We present an approach to making existing services and service environments in CLARIN usable for spoken language data. That this is possible and useful in principle has already been demonstrated by proof-of-concept implementations in the tools ELAN (Sloetjes, 2014) and EXMARaLDA (Schmidt and Wörner, 2014, see section 6), which both provide interfaces from their respective tool formats to WebLicht and WebMaus (Kisler et al., 2012). The approach described here is potentially more flexible because it is based on a recently published TEI-based ISO standard for spoken language transcriptions and thus not directly tied to a specific tool or tool format.

The general architecture we envision and have started to implement is depicted in Figure 1. The point of departure for most users will be one of a few established formats of tools for multimedia annotation. This needs to be converted to the ISO/TEI standard format which then constitutes the basis for all further processing steps. Existing services in WebLicht can be made usable by providing an encoder/decoder pair to/from WebLicht's TCF format. Additional services specializing on spoken language data can operate directly on the ISO/TEI data.

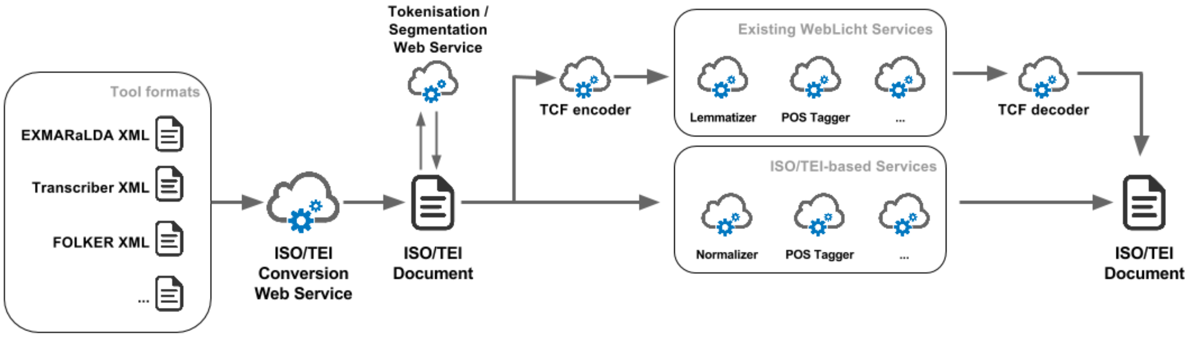


Figure 1: Architecture

We briefly sketch the most important characteristics of the ISO/TEI format in section 2. Sections 3 (conversion from tool to ISO/TEI) and 4 (TCF encoder/decoder) describe the conversion steps needed to connect existing tools with WebLicht via the ISO standard. Section 5 addresses some details of the CLARIN integration of these services, while section 6 deals with the way users can interact with them. In section 7, we briefly introduce the idea of web services directly operating on the ISO/TEI standard.

2 A TEI-Based ISO Standard for Spoken Language Transcriptions

The Text Encoding Initiative's Guidelines (TEI, chapter 8)[3], have always contained suggestions for representing spoken language transcriptions within the TEI framework. However, this portion of the guidelines has never been sufficiently established in the respective research communities to really work as an interchange format or even a standard. As Bird and Liberman (2001, p. 26) put it:

> The TEI guidelines for 'Transcriptions of Speech' offer access to a very broad range of representational techniques drawn from other aspects of the TEI specification. The TEI report sketches or alludes to a correspondingly wide range of possible issues in speech annotation. All of these seem to be encompassed within [the author's AG framework], but it does not seem appropriate to speculate at much greater length about this, given that this portion of the TEI guidelines does not seem to have been used in any published transcriptions to date. [emphasis added]

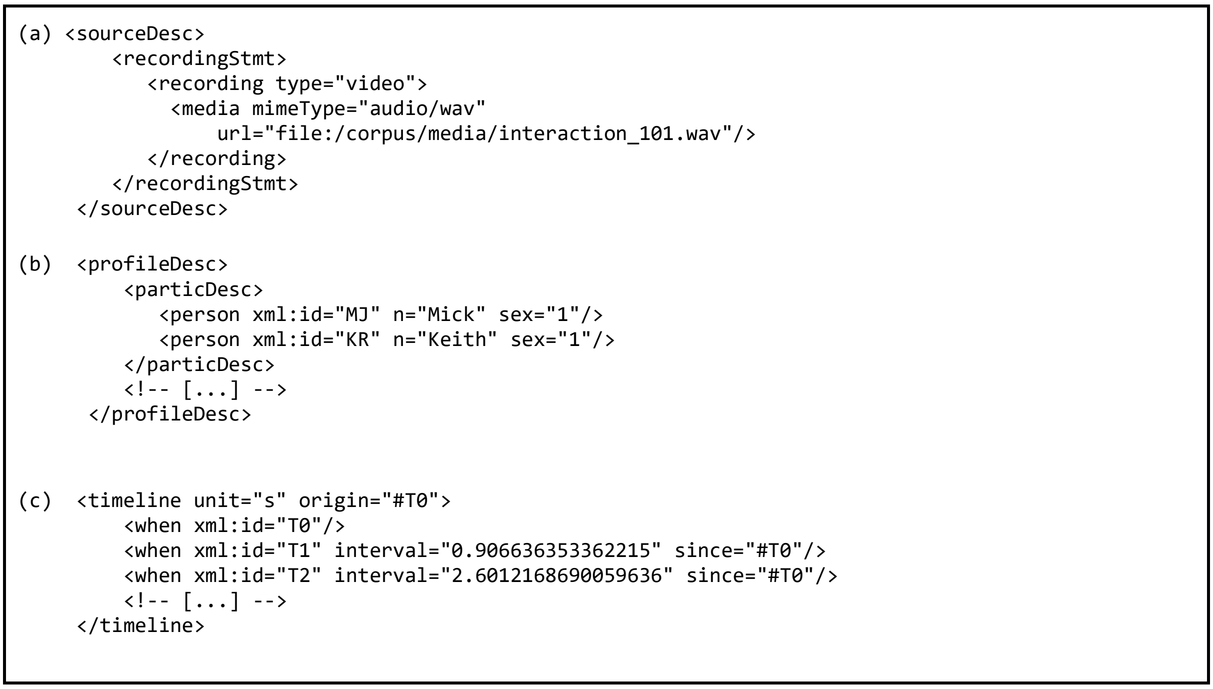
One reason for this lack of acceptance is that work with spoken language data relies on specialized tools for efficient transcription, and the relation of TEI to these tools was never sufficiently clarified. Schmidt (2005) therefore made a first suggestion on how to reconcile the time-based data models, which most tools are based on, with the hierarchy-based data model underlying the TEI proposal (see also Parisse and Morgenstern, 2010, for a similar approach). Taking into account findings from a tool interoperability study (Schmidt et al., 2009), the proposal for representing spoken language transcriptions on the basis of TEI was further refined in Schmidt (2011), and from 2012 on, became the topic of an ISO standardization project (ISO/ TC 37/SC 4/WG 6), concluded in summer 2016 with the official publication of "Language resource management - Transcription of Spoken Language (24642)" (see ISO,

[3] See: http://www.tei-c.org/Vault/P5/1.0.0/doc/tei-p5-doc/en/html/TS.html.

2016). We will briefly outline some characteristics of this standard that are important to the subject of this paper.

The general focus of the standard is on orthography-based (i.e. not: IPA-based) transcription of recordings of authentic interaction (i.e. not: monologic "speech" or experiment data). Guiding design principles were the maxim to reuse as many elements as possible from chapter 8 of the existing TEI guidelines[4] and to orient their use towards interoperability with established tools. In particular, this meant a conscious limitation of choices in the numerous cases where the guidelines offer more than one concept for representing one and the same phenomenon.

As exemplified in Figure 2, basic building blocks for the document structure are (a) one or more <recording> elements specifying the underlying audio and/or video file(s), (b) a <particDesc> defining the participants of the interaction, and (c) a <timeline> providing offsets into a recording.



```
(a) <sourceDesc>
        <recordingStmt>
           <recording type="video">
             <media mimeType="audio/wav"
                 url="file:/corpus/media/interaction_101.wav"/>
           </recording>
        </recordingStmt>
     </sourceDesc>

(b)  <profileDesc>
         <particDesc>
            <person xml:id="MJ" n="Mick" sex="1"/>
            <person xml:id="KR" n="Keith" sex="1"/>
         </particDesc>
         <!-- [...] -->
      </profileDesc>


(c)  <timeline unit="s" origin="#T0">
         <when xml:id="T0"/>
         <when xml:id="T1" interval="0.906636353362215" since="#T0"/>
         <when xml:id="T2" interval="2.6012168690059636" since="#T0"/>
         <!-- [...] -->
     </timeline>
```

Figure 2: Recording(s), participant(s) and the timeline as basic building blocks

The main part of the document is then made up of a sequence of <u> elements. They correspond to individual speaker contributions and contain the actual transcription text, references to points in the timeline and to the respective speaker (@who). As illustrated in Figure 3, the standard allows different levels of detail for the actual markup of the transcription text. In the simplest case (example 1 in Figure 3), a plain text string can be used, which is temporally aligned via mandatory @start and @end attributes of the <u> element. Intervening temporal alignment can be added in the form of additional <anchor> milestone elements (example 2) whose @synch attribute refers to a point in the timeline.

The microstructure of the speaker contribution can be represented by inserting additional markup, most importantly <w> for word tokens, <pause> for pauses and <vocal> or <kinesic> for non-verbal phenomena like coughing or laughing (example 3). Finally, segmentations of speaker contributions into units above the word level (the "sentence equivalents" of spoken language such as intonation units) can be represented by intervening <seg> elements (example 4). As we will discuss below, the additional markup below <u> is crucial for many automatic annotation methods. The examples in Figure 3 illustrate transcription proper, i.e. the direct representation in written form of what is heard or seen in the primary data.

[4] See: http://www.tei-c.org/Vault/P5/1.0.0/doc/tei-p5-doc/en/html/TS.html.

```
(1) <u who="MJ" start="#T0" end="#T2">
        I ((cough)) see a door. I (0.3) want to paint it (black/blue).
    </u>

(2) <u who="MJ" start="#T0" end="#T2">
        I ((cough)) see a door.
        <anchor synch="#T1"/>
        I (0.3) want to paint it (black/blue).
     </u>
(3) <u who="MJ" start="#T0" end="#T2">
        <w>I</w>
        <vocal><desc>cough</desc></vocal>
        <w>see</w><w>a</w><w>door</w><p>.</p>
        <anchor synch="#T1"/>
        <w>I/w>
        <pause dur="PT0.3S"/>
        <w>want</w><w>to</w><w>paint</w><w>it/w>
        <unclear><choice><w>black</w><w>blue</w></choice></unclear>
        <p>.</p>
     </u>

(4) <u who="MJ" start="#T0" end="#T2">
        <seg type="intonation-phrase" subtype="falling">
                <w>I</w>
                <vocal><desc>cough</desc></vocal>
                <w>see</w><w>a</w><w>door</w>
        </seg>
        <anchor synch="#T1"/>
        <seg type="intonation-phrase" subtype="falling">
                <w>I/w>
                <pause dur="PT0.3S"/>
                <w>want</w><w>to</w><w>paint</w><w>it/w>
                <unclear><choice><w>black</w><w>blue</w></choice></unclear>
        </seg>
    </u>
```

Figure 3: <u> elements with different levels of internal markup

In order to represent additional annotations on that material, the ISO/TEI standard provides the possibility to introduce an arbitrary number of standoff annotation layers in <spanGrp> elements and to group these with the <u> element they belong to, using an <annotationBlock> element (see Banski et al., 2016). This mechanism is crucial also for storing annotations that result from automatic annotation methods in WebLicht. Figure 4 illustrates the annotation of an utterance with lemmas and part-of-speech tags.

```
<annotationBlock who="MJ" start="#T0" end="#T2" xml:id="ab1">
   <u xml:id="u1">
        <seg type="intonation-phrase" subtype="falling" xml:id="seg1">
                <w xml:id="w1">I</w>
                <vocal xml:id="voc1"><desc>cough</desc></vocal>
                <w xml:id="w2">see</w>
                <w xml:id="w3">a</w>
                <w xml:id="w4">door</w>
        </seg>
   </u>
   <spanGrp type="lemma">
        <span from="#w1" to="#w1">I</span>
        <span from="#w2" to="#w2">see</span>
        <span from="#w3" to="#w3">a</span>
        <span from="#w4" to="#w4">door</span>
   </spanGrp>
   <spanGrp type="pos">
        <span from="#w1" to="#w1">PPER</span>
        <span from="#w2" to="#w2">V</span>
        <span from="#w3" to="#w3">DET</span>
        <span from="#w4" to="#w4">NN</span>
   </spanGrp>
</annotationBlock>
```

Figure 4: <annotationBlock> grouping an <u> with standoff annotation (lemmatization and POS tagging) in <spanGrp>

With the same mechanism (figure 5), orthographically normalized forms can be assigned to transcribed forms when the latter do not follow standard orthography. This is the case, for instance, in many conversation analytic transcription systems that use "literary transcription" or "eye dialect" to represent actual pronunciations that deviate from the ones suggested by the orthographic form (as in "gotta" for "got to").

```
<annotationBlock who="CB" start="#T0" end="#T2" xml:id="ab1">
   <u xml:id="u1">
        <w xml:id="w1">sure</w>
        <w xml:id="w2">nuff</w>
        <w xml:id="w3">an</w>
        <w xml:id="w4">yes</w>
        <w xml:id="w5">I</w>
        <w xml:id="w6">do</w>
   </u>
   <spanGrp type="normalized">
        <span from="#w1" to="#w1">sure</span>
        <span from="#w2" to="#w2">enough</span>
        <span from="#w3" to="#w3">and</span>
        <span from="#w4" to="#w4">yes</span>
        <span from="#w5" to="#w5">I</span>
        <span from="#w6" to="#w6">do</span>
   </spanGrp>
</annotationBlock>
```

Figure 5: <annotationBlock> grouping an <u> with standoff annotation (orthographically normalized forms) in <spanGrp>

For the spoken language data hosted at the CLARIN centers in Hamburg (HZSK) and Mannheim (IDS/AGD), we have confirmed that a lossless, fully automatic transformation from existing formats (mostly EXMARaLDA and/or FOLKER) to ISO compliant TEI is possible. Current and future developments at these data centers will make this standard a central element of all workflows.

In the context of the Parthenos project[5], support will be given via INRIA to further develop and document the standard and disseminate it to the scientific community.

3 Converting Common Transcription Formats to ISO/TEI

Unlike other text types (such as manuscripts, dictionaries etc.) addressed by the TEI, spoken language transcription is rarely done by editing an XML document directly. Researchers crucially rely on tools which support the alignment of sound/video and transcription in an ergonomic graphical user interface. In an early CLARIN Deliverable (Hinrichs and Vogel, 2010), ANVIL (Kipp, 2014), CLAN (MacWhinney, 2000), ELAN (Sloetjes, 2014), EXMARaLDA (Schmidt and Wörner, 2014), FOLKER (Schmidt, 2012), Praat (Boersma, 2014), and Transcriber (Barras et al., 2000), have been identified as the annotation tools that are currently most relevant to the CLARIN community for this task. Most of them (CLAN and Praat being the exceptions) do work with XML based formats, but, so far, none of them "natively" operates on a TEI compliant format. In order to make the ISO/TEI standard work in practice, it is therefore essential to furnish users with an easy-to-use way of converting from a given tool format to the ISO/TEI format (ideally also the other way around, but this is more complex and will not be dealt with here[6]). The closer the tool format is to TEI's general structure, the more straightforward this conversion will be.

Transcriber and FOLKER both use a data model which closely resembles the ISO/TEI approach insofar as it organizes transcripts as lists of speaker contributions (marked-up as <Turn> in Transcriber and as <contribution> in FOLKER) which can be seen as directly corresponding to TEI's <u> elements. Whereas FOLKER (example 2a in Figure 6) allows additional markup for transcribed text underneath that unit (most importantly <w> for tokens), Transcriber represents the actual transcription text as plain

[5] See: http://www.parthenos-project.eu/ and http://www.parthenos-project.eu/Download/Deliverables/D4.1_Standardization_Survival_Kit.pdf

[6] The ISO/TEI format is potentially richer in information than any single of the tool formats taken into consideration. A conversion from ISO/TEI to a tool format therefore has to deal with all kinds of possible information loss.

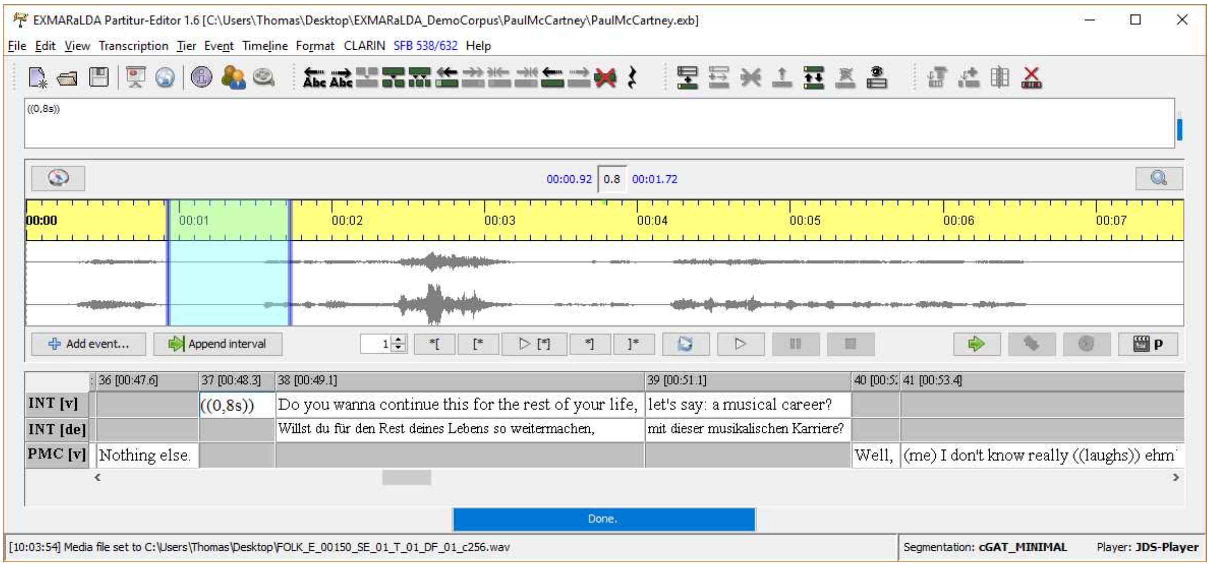
character data and uses additional markup only for non-speech elements (1a). Both formats can be transferred to ISO/TEI (2b and 1b) as a more or less direct mapping of elements, without any fundamental structural changes. This can be achieved by simple XSLT transformations.

```
(1a) <Turn speaker="spk1" startTime="0.511" endTime="7.356">
        <Sync time="0.511"/><Event desc="souffle" type="noise" extent="instantaneous"/>
        <Sync time="1.593"/>hé bien euh bonsoir à tous et merci d'être restés si nombreux
        <Sync time="5.174"/>
        <Event desc="rire" type="noise" extent="instantaneous"/>
     </Turn>

(1b) <u who="#spk1" start="#T1" end="#T4">
         <vocal><desc>souffle</desc></vocal>
         <anchor synch="#T2"/>
         hé bien euh bonsoir à tous et merci d'être restés si nombreux
         <anchor synch="#T3"/>
         <vocal><desc>rire</desc></vocal>
     </u>

(2a) <contribution speaker-reference="IL" start-reference="TLI_33" end-reference="TLI_36">
        <w>zweihundert</w><w>ist</w><w>die</w><w>äh</w>
        <breathe type="in" length="1"/>
        <w>ist</w><w>die</w><w>planung</w>
        <time timepoint-reference="TLI_34"/>
        <pause duration="micro"/>
        <w>wenn</w><w>es</w><w>ausgebaut</w>
        <time timepoint-reference="TLI_35"/>
        <w>wird</w>
     </contribution>

(2b) <u who="#IL" start="#TLI_33" end="#TLI_36">
        <w xml:id="wd1e791">zweihundert</w>
        <w xml:id="wd1e799">ist</w>
        <w xml:id="wd1e801">die</w>
        <w xml:id="wd1e805">äh</w>
        <vocal>
           <desc>short breathe in</desc>
        </vocal>
        <w xml:id="wd1e808">ist</w>
        <w xml:id="wd1e810">die</w>
        <w xml:id="wd1e813">planung</w>
        <anchor synch="#TLI_34"/>
        <pause type="micro"/>
        <w xml:id="wd1e817">wenn</w>
        <w xml:id="wd1e819">es</w>
        <w xml:id="wd1e821">ausgebaut</w>
        <anchor synch="#TLI_35"/>
        <w xml:id="wd1e824">wird</w>
   </u>
```

Figure 6: Transformation of Transcriber (1) and FOLKER (2) file formats to ISO/TEI

CLAN's CHAT format works similarly in principle, with the important difference, however, that it is a plain text format and thus less directly usable as an input to an XSLT transformation. We currently use an existing CHAT-to-EXMARaLDA converter and transform its result to ISO/TEI.

The other formats differ from Transcriber, FOLKER and CHAT in that they are tier-based and thus do not provide a direct equivalent for the <u> element. For conversion of EXMARaLDA files to ISO/TEI, we rely on the concept of a "segment chain" – a maximally long sequence of contiguous annotations in a main tier – as the basic building block of a transcription. Segment chains are mapped to <u> elements and the respective contents of dependent tiers are then integrated in appropriate subordinate <spanGrp> elements. The principle is discussed in more depth in Schmidt (2005). Again, a single XSLT transformation is sufficient to achieve the conversion from EXMARaLDA (1a in Figure 8) to the ISO/TEI format (1b).


EXMARaLDA Partitur-Editor 1.6 [C:\Users\Thomas\Desktop\EXMARaLDA_DemoCorpus\PaulMcCartney\PaulMcCartney.exb]
File Edit View Transcription Tier Event Timeline Format CLARIN SFB 538/632 Help
((0,8s))
00:00.92 0.8 00:01.72
00:00 00:01 00:02 00:03 00:04 00:05 00:06 00:07
Add event...
Append interval
Done.

	36 [00:47.6]	37 [00:48.3]	38 [00:49.1]	39 [00:51.1]	40 [00:5:	41 [00:53.4]
INT [v]		((0,8s))	Do you wanna continue this for the rest of your life,	let's say: a musical career?		
INT [de]			Willst du für den Rest deines Lebens so weitermachen,	mit dieser musikalischen Karriere?		
PMC [v]	Nothing else.				Well,	(me) I don't know really ((laughs)) ehm˙

[10:03:54] Media file set to C:\Users\Thomas\Desktop\FOLK_E_00150_SE_01_T_01_DF_01_c256.wav

Segmentation: cGAT_MINIMAL Player: JDS-Player

Figure 7: Musical score representation of EXMARaLDA's tier-based transcription format

```
(1a) <tier id="TIE0" speaker="SPK1" category="v" type="t" display-name="INT [v]">
         <event start="T37" end="T38">((0,8s))</event>
         <event start="T38" end="T39">Do you wanna continue this for the rest of your life, </event>
         <event start="T39" end="T40">let's say: a musical career? </event>
     </tier>
     <tier id="TIE2" speaker="SPK1" category="de" type="a" display-name="INT [de]">
         <event start="T38" end="T39">Willst du für den Rest deines Lebens so weitermachen, </event>
         <event start="T39" end="T40">mit dieser musikalischen Karriere? </event>
     </tier>
     <tier id="TIE3" speaker="SPK0" category="v" type="t" display-name="PMC [v]">
         <event start="T36" end="T37">Nothing else. </event>
         <event start="T40" end="T41">Well, </event>
         <event start="T41" end="T42">(me) I don't know really ((laughs)) ehm˙ </event>
     </tier>

(1b) <annotationBlock who="#SPK0" start="#T36" end="#T37">
         <u xml:id="u_d1e119">Nothing else. </u>
     </annotationBlock>
     <annotationBlock who="#SPK1" start="#T37" end="#T40">
         <u xml:id="u_d1e102">((0,8s))Do you wanna continue this for the rest of your life,
             <anchor synch="#T39"/>let's say: a musical career? </u>
         <spanGrp type="de">
             <span from="#T38" to="#T39">Willst du für den Rest deines
                   Lebens so weitermachen, </span>
             <span from="#T39" to="#T40">mit dieser musikalischen Karriere? </span>
         </spanGrp>
     </annotationBlock>
     <annotationBlock who="#SPK0" start="#T40" end="#T42">
         <u xml:id="u_d1e122">Well, (me) I don't know really ((laughs)) ehm˙ </u>
     </annotationBlock>
```

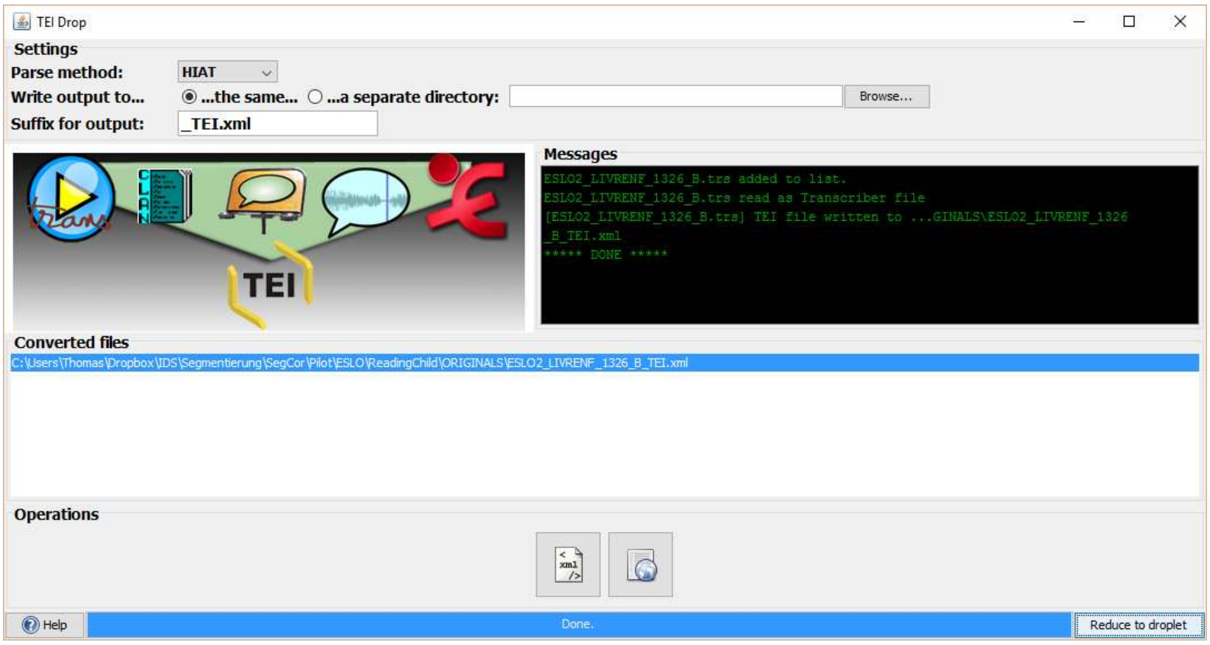
Figure 8: Transformation of EXMARaLDA file format to ISO/TEI

Praat's TextGrid format can be treated in a similar manner with, again, a detour via an existing Praat-to-EXMARaLDA converter because TextGrids are plain text, not XML, files. Since, however, Praat's data model only allows tiers to have a name (in the form of a simple string) and has no place for further structural information about tiers (such as speaker assignment or dependencies between transcription and annotation tiers), some information necessary for a TEI conversion will either have to be derived from ad-hoc tier naming conventions, or be added manually in a tool (such as EXMARaLDA) whose data model is sufficiently specific in this respect.

The situation is a bit more complex for ELAN's EAF format because of its more powerful possibilities of defining tier dependencies. In the context of the extended focus of the HZSK related to the associated CLARIN-D discipline-specific working group on Linguistic Fieldwork, Ethnology, and Language Typology, we are currently experimenting with existing EAF data sets from a language documentation background, and can confirm that, at a minimum, an ELAN to ISO/TEI conversion should be possible as a rule on a per-corpus basis. We have not tackled ANVIL's file format yet.

Depending on the information available in the input format, some results of conversions from a tool format to ISO/TEI will have token information (example 2b in Figure 6) and some will not (examples 1b in Figures 6 and 8). Tokenization information – expressed in the ISO/TEI format most importantly through <w> markup – is, however, crucial for the majority of automatic language tools. Unfortunately, tokenization algorithms developed for written language are of limited use for spoken language transcriptions because they are not aware of the form and meaning of non-speech elements like pauses, descriptions of non-verbal behavior etc. Before a spoken language transcription is fed into a POS tagger or similar tools, a tokenization must therefore be carried out in order to separate (i.e. markup) "real" words from other, non-word elements. In the context of EXMARaLDA, we have developed a mechanism (called segmentation algorithm) which makes use of the regularities defined by transcription systems (such as "pauses are written as decimal numbers between round brackets") in order to achieve such a tokenization (again, see Schmidt, 2005 for further details). So far, EXMARaLDA is able to cleanly tokenize transcripts following the HIAT (Rehbein et al., 2004), GAT (Selting et al., 2009), cGAT (Schmidt et al., 2015), CHAT (MacWhinney, 2000) and DIDA (Klein & Schütte, 2001) transcription conventions. Ideally, these segmentation algorithms would have to be adapted to operate on the ISO/TEI (rather than the EXMARaLDA) format. They could then be offered (for instance, by a web service) as an additional processing step between the tool format conversion and further annotation through, say, a lemmatizer or POS tagger. We plan to address this requirement in the near future. For the time being, EXMARaLDA and its built-in segmentation algorithms for the different transcription systems can be used as an intermediary, for instance to transform a (non-tokenized) Transcriber file via EXMARaLDA, where a HIAT segmentation algorithm is applied, to a tokenized ISO/TEI file.

Export filters from the tool to the ISO/TEI format have been built directly into EXMARaLDA and FOLKER. For other tool formats or batch conversion of entire corpora, the EXMARaLDA distribution provides TEI-Drop (see Figure 9), a droplet desktop application onto which users can drag and drop a number of input files (in Transcriber, FOLKER, EXMARaLDA, CHAT or EAF), specify a few parameters such as which segmentation algorithm to use and where to write the output, and which will then perform the conversions in a single step.



Parse method: HIAT

Write output to... ● ...the same... ○ ...a separate directory:

Suffix for output: _TEI.xml

Messages

```
ESLO2_LIVRENF_1326_B.trs added to list.
ESLO2_LIVRENF_1326_B.trs read as Transcriber file
[ESLO2_LIVRENF_1326_B.trs] TEI file written to ...GINALS\ESLO2_LIVRENF_1326
_B_TEI.xml
***** DONE *****
```

Converted files

C:\Users\Thomas\Dropbox\IDS\Segmentierung\SegCor\Pilot\ESLO\ReadingChild\ORIGINALS\ESLO2_LIVRENF_1326_B_TEI.xml

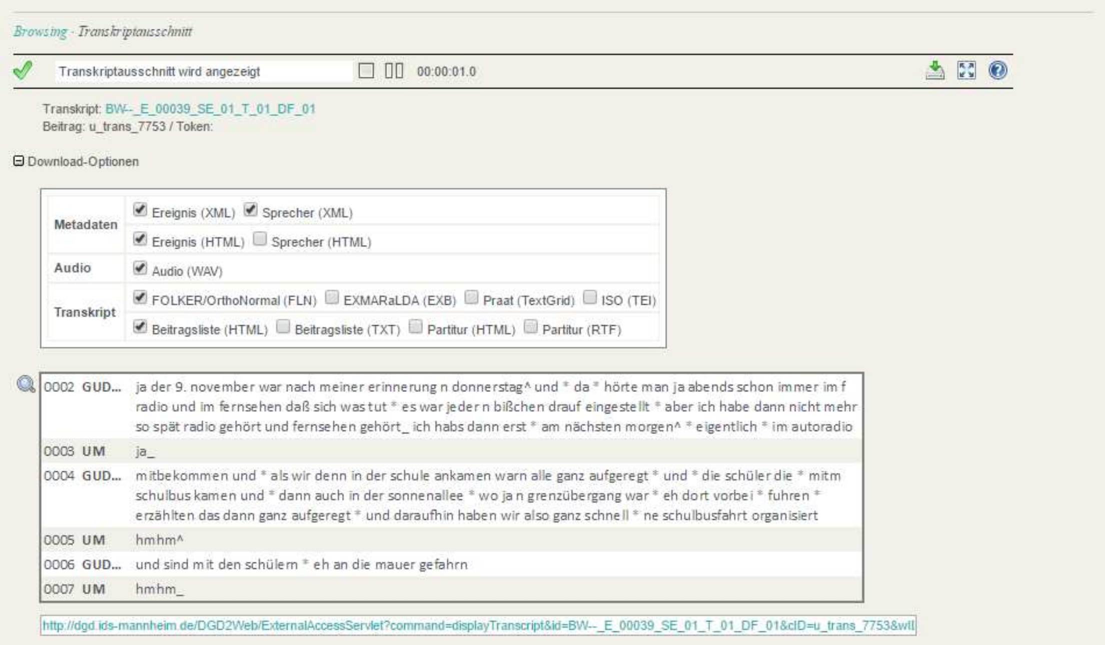
Done.

Figure 9: Screenshot of TEI Drop

In order for the standard to be adopted on a larger scale, it is crucial not only that the possibility of converting existing data is provided, but also that essential existing language resources are made available in that format to the community. The Archive for Spoken German with its DGD platform[7] (Schmidt,

[7] See: http://dgd.ids-mannheim.de

2014) as the central provider of German oral corpora, as well as the Hamburg Centre for Language Corpora with its repository of multilingual spoken data[8] (Jettka and Stein, 2014) are two certified CLARIN centers that are in principle ready to offer their resources (more than 5000 hours of material, altogether) in the ISO/TEI format.



Browsing - Transkriptausschnitt

Transkriptausschnitt wird angezeigt 00:00:01.0

Transkript: BW--_E_00039_SE_01_T_01_DF_01
Beitrag: u_trans_7753 / Token:

⊟ Download-Optionen

Metadaten	☑ Ereignis (XML) ☑ Sprecher (XML)
	☑ Ereignis (HTML) □ Sprecher (HTML)
Audio	☑ Audio (WAV)
Transkript	☑ FOLKER/OrthoNormal (FLN) □ EXMARaLDA (EXB) □ Praat (TextGrid) □ ISO (TEI)
	☑ Beitragsliste (HTML) □ Beitragsliste (TXT) □ Partitur (HTML) □ Partitur (RTF)

0002	GUD...	ja der 9. november war nach meiner erinnerung n donnerstag^ und * da * hörte man ja abends schon immer im f radio und im fernsehen daß sich was tut * es war jeder n bißchen drauf eingestellt * aber ich habe dann nicht mehr so spät radio gehört und fernsehen gehört_ ich habs dann erst * am nächsten morgen^ * eigentlich * im autoradio
0003	UM	ja_
0004	GUD...	mitbekommen und * als wir denn in der schule ankamen warn alle ganz aufgeregt * und * die schüler die * mitm schulbus kamen und * dann auch in der sonnenallee * wo ja n grenzübergang war * eh dort vorbei * fuhren * erzählten das dann ganz aufgeregt * und daraufhin haben wir also ganz schnell * ne schulbusfahrt organisiert
0005	UM	hmhm^
0006	GUD...	und sind mit den schülern * eh an die mauer gefahrn
0007	UM	hmhm_

http://dgd.ids-mannheim.de/DGD2Web/ExternalAccessServlet?command=displayTranscript&id=BW--_E_00039_SE_01_T_01_DF_01&cID=u_trans_7753&wI

Figure 10: Export options for transcript excerpts in the Database for Spoken German (DGD)

For the time being, that format will be an additional option beside the established formats favored by the respective centers (as in Figure 10). In the mid-term, however, we aim at making ISO/TEI the central format for disseminating our resources. Ongoing or recently piloted cooperations with international partners – most importantly the ESLO corpus in Orléans (Eshkol-Taravella et al., 2012), the CLAPI database in Lyon (Groupe ICOR, 2017), the Speech Island Portal in Austin/Texas[9], and the spoken components of the Australian National Corpus[10] – may, we hope, lead to a wider distribution of the standard on an international level.

Another crucial aspect for the acceptance of the standard is the interoperability with existing tools to further annotate, to query, and to analyze transcribed data. Support for ISO/TEI as an import/export format in WebAnno[11] (Eckart de Castilho et. al, 2014) is planned to be developed starting this spring, which would enable web-based manual annotation using annotation layers such as dependency parsing and co-reference resolution. For the search and visualization tool ANNIS (Krause and Zeldes, 2016)[12], a prototype Pepper (Zipser and , 2010) conversion module has been implemented. Finally, the Solr-based Multi Tier Annotation Search (MTAS, Brouwer and Kemps-Snijders, 2016)[13] system developed at the Meertens Instituut has very recently been extended to also index the ISO/TEI format and a version including ISO/TEI transcription samples is being tested. Support for the ISO/TEI format would enable a powerful CQL-based search to be used with the various transcription formats for which conversion methods exist.

[8] See: http://hdl.handle.net/11022/HZSK-0000-0000-2C76-B-REPOSITORY.
[9] See: http://speechislands.org
[10] See, for instance: https://www.ausnc.org.au/corpora/gcsause
[11] See: https://webanno.github.io/
[12] See: http://corpus-tools.org/annis/
[13] See: https://meertensinstituut.github.io/mtas/

4 Using TCF-based Web Services in WebLicht

Currently, all services integrated in WebLicht operate on version 0.4 of the Text Corpus Format (TCF)[14]. In order for an ISO/TEI document to be sent through a WebLicht tool chain, it therefore needs to be encoded in TCF before the first step of the chain is applied, and the result of the chain needs to be decoded from TCF to ISO/TEI after the last step of the chain. TCF is a format based on the "stream of tokens" idea and assumes that the basic structure of any document is a linear sequence of token elements:

> The tokens layer is [thus] the main anchor layer among TextCorpus layers, i.e. all other layers (with the exception of the text layer) directly or indirectly (via other layers) reference tokens by referencing token identifiers. [Introduction to section 3 of the TCF specification]

TCF thus does not have in its basic structure any means of representing information that is crucial to spoken language transcription, such as time alignment and speaker assignment, and it also does not provide the possibility to distinguish different types of tokens (such as words vs. non-speech descriptions) on its basic layer. ISO/TEI-to-TCF conversion is therefore bound to be lossy – not all information contained in a transcription can be meaningfully mapped onto some TCF element.

Our general approach for the TCF encoding of ISO/TEI files is therefore:

1. To only map those elements of an ISO/TEI file which have a more or less direct equivalent in TCF. Basically, this boils down to mapping <w> elements (and, as the case may be <pc> elements containing punctuation) in ISO/TEI to <token> elements in TCF.[15]
2. To assume that most services in WebLicht will work fine with this reduced set of information, i.e. will produce useful results, although some types of information have been removed.
3. To keep on the <tokens> layer the @xml:id attributes of the original document (in an @ID attribute of the individual <token> elements).
4. To keep in the <textSource> element the entire original ISO/TEI document.

When such a TCF document is fed through a tool chain in WebLicht, any resulting additional annotation layers can then be remapped to a suitable ISO/TEI form via the memorized @xml:id attributes. Since the original document in the <textSource> element is also fed through the tool chain unchanged, this decoding step can be stateless as required by the SOA architecture of CLARIN in general.

Many tools in WebLicht require information about sentence boundaries. Since spoken language transcriptions, as a rule, do not operate with the concept of a "sentence", a further important question in the encoding process is therefore which elements of the source document can be meaningfully mapped onto the <sentences> layer in TCF. Our approach to this question is very straightforward: if a segmentation of the transcript in the form of <seg> elements directly underneath the <u> elements exists, we use that segmentation as a sentence equivalent. Although, semantically, these entities (such as intonation phrases or speech acts) are usually not sentences in the written-language sense of the word, they are usually sufficiently similar to sentences in order to function as an adequate entity in their stead. In the absence of <seg> elements, we use the entire <u> element as a sentence equivalent.

Figure 11 illustrates how an excerpt of an ISO/TEI transcription (1) is encoded in TCF (2), supplemented with the result of a POS tagger (3) and then redecoded to ISO/TEI (4).

[14] See: http://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/The_TCF_Format

[15] Alternatively, one might map onto TCF's <token> element not the character data of <w> elements, but the content of <span> elements in <spanGrp> which represent orthographic normalised forms (see figure 5 above), if such an annotation level exists. This has a potential of improving annotation results in WebLicht, because individual tools in the chain would then operate on the standard orthography they usually expect. This could be implemented as a parameter passed to the TEI to TCF converter or even be decided autonomously by the web service. We currently refrain from such an implementation because our focus here is on establishing a general mechanism for CLARIN integration of spoken language data.

```
(1) <u who="MJ" start="#T0" end="#T2">
   <seg type="intonation-phrase" subtype="falling">
        <w xml:id="w1">I</w>
        <vocal><desc>cough</desc></vocal>
        <w xml:id="w2">see</w>
        <w xml:id="w3">a</w>
        <w xml:id="w4">door/w>
    </seg>
    <anchor synch="#T1"/>
    <seg type="intonation-phrase" subtype="falling">
        <w xml:id="w5">I</w>
        <pause dur="PT0.3S"/>
        <w xml:id="w6">want</w>
        <w xml:id="w7">to</w>
        <w xml:id="w8">paint/w>
        <w xml:id="w9">it</w>
        <unclear>
            <choice>
                <w xml:id="w10">black</w>
                <w xml:id="w11">blue</w>
            </choice>
        </unclear>
    </seg>
</u>

(2) <TextCorpus>
    <text>I see a door I want to paint it black blue</text>
    <tokens>
       <token ID="w1">I</token>
       <token ID="w2">see</token>
       <token ID="w3">a</token>
       <token ID="w4">door</token>
       <token ID="w5">I</token>
       <!-- [...] -->
       <token ID="w9">it</token>
       <token ID="w10">black</token>
       <token ID="w11">blue</token>
    </tokens>
    <sentences>
       <sentence ID="s_1" tokenIDs="w1 w2 w3 w4"/>
       <sentence ID="s_2" tokenIDs="w5 w6 w7 w8 w9 w10 w11"/>
    </sentences>
    <textSource type="application/tei+xml;format-variant=tei-iso-spoken;tokenized=1">
        <![CDATA[
            <TEI xmlns="http://www.tei-c.org/ns/1.0">
               [...]
               <u who="MJ" start="#T0" end="#T2">
               [...]
            </TEI>
        ]]>
    </textSource>
<TextCorpus>

(3) <TextCorpus>
    <!-- [...] -->
    <POStags tagset="stts">
            <tag ID="pt_0" tokenIDs="w1">PPER</tag>
            <tag ID="pt_1" tokenIDs="w2">V</tag>
            <tag ID="pt_2" tokenIDs="w3">DET</tag>
            <tag ID="pt_3" tokenIDs="w4">NN</tag>
            <!-- [...] -->
            <tag ID="pt_11" tokenIDs="w11">ADJ</tag>
        </POStags>
    <!-- [...] -->
</TextCorpus>

(4) <annotationBlock who="MJ" start="#T0" end="#T2" xml:id="ab1">
   <u xml:id="u1">
        <seg type="intonation-phrase" subtype="falling" xml:id="seg1">
                <w xml:id="w1">I</w>
                <vocal xml:id="voc1"><desc>cough</desc></vocal>
                <w xml:id="w2">see/w>
                <w xml:id="w3">a/w>
                <w xml:id="w4">door/w>
        </seg>
   </u>
   <spanGrp type="pos">
        <span from="#w1" to="#w1">PPER</span>
        <span from="#w2" to="#w2">V</span>
        <span from="#w3" to="#w3">DET</span>
        <span from="#w4" to="#w4">NN</span>
   </spanGrp>
</annotationBlock>
```

Figure 11: Encoding, WebLicht processing and decoding of an ISO/TEI file

5 CLARIN integration

An important practical question concerning the integration of the conversion services into the WebLicht context (and, ultimately, into the CLARIN infrastructure in general) is how to inform other services of the format of the admissible input (what the service “consumes”) and the expected output (what the service “produces”) of any given single converter. MIME type[16] in CMDI specifications had been used for that purpose before, but without a clear, commonly agreed guideline, resulting in several unsolved challenges:

1. MIME types have to be sufficiently specific for the application that is meant to handle the respective file format. For example, a MIME type like “application/tei+xml”, although perfectly valid, will not be sufficient for the converters described here, because those expect a specific type of TEI file (i.e. one that conforms to the ISO standard) and will fail when confronted with other TEI variants.
2. Only one MIME type should be used consistently per given file format. For instance, it would be possible to describe an ELAN file either as “text/x-eaf+xml” (to be found in several existing CLARIN records) or as “application/xml” (more in line with current recommendations for XML files). A web service must at least know all possible variants or, better still, be sure that exactly one variant is used in the infrastructure.
3. The use of registered MIME types should be preferred over non-registered MIME types. Non-registered MIME types (e.g. using the “x-” prefix) have been used ad hoc to describe specific annotation formats, but using non-registered MIME types in this context is problematic, since they were “intended exclusively for use in private, local environments”[17].

After discussions involving the CLARIN task force on WebLicht/TEI, the CLARIN-D developer group and the CLARIN standards committee, we settled for the MIME type assignments in Table 1. These MIME types comply with current recommendations and best practices[18] . The first part always consists of a registered MIME type so that even applications unaware of the specific linguistic annotation formats still have a chance to do some useful processing of the data (e.g. a web browser will recognize an XML file as such and display its DOM tree). In the second part, an additional parameter “format-variant” is used to provide the more detailed format information needed by the web services described here.

[16] For reasons of consistency, we use “MIME type” to refer to media types (as they are now officially called) throughout this paper.

[17] See: https://tools.ietf.org/html/rfc6838#section-3.4

[18] See: https://tools.ietf.org/html/rfc2046 and https://www.w3.org/TR/webarch/#xml-media-types

Format	MIME type
Text Corpus Format (*.tcf)	text/tcf+xml or application/xml; format-variant=weblicht-tcf[19]
ISO/TEI for transcriptions of spoken language	application/tei+xml; format-variant=tei-iso-spoken[20]
EXMARaLDA Basic transcription (*.exb)	application/xml; format-variant=exmaralda-exb
Transcriber annotation file (*.trs)	application/xml; format-variant=transcriber-trs
FOLKER transcription (*.flk / *.fln)	application/xml; format-variant=folker-fln
CHAT transcription file (*.cha)	text/plain; format-variant=clan-cha[21]
ELAN Annotation File (*.eaf)	application/xml; format-variant=elan-eaf[22]
Praat TextGrid (*.textGrid)	text/plain; format-variant=praat-textgrid[23]

Table 1: MIME types for different annotation formats

6 Using WebLicht tool chains in practice

The different converters are made available as individual web services, hosted by the CLARIN cener HZSK and exposed via a CMDI description (e.g. http://hdl.handle.net/11022/0000-0000-9ABA-1 for a development version of the EXMARaLDA-to-ISO/TEI converter) to CLARIN services such as the Virtual Language Observatory and WebLicht. Once fully integrated into CLARIN in that way[24], WebLicht will recognize the different annotation tool formats and offer users the ISO/TEI conversion service as a possible first conversion step. By applying the ISO/TEI-to-TCF converter to the result, the rest of the annotation services in WebLicht become available for the data. In a final step, the result of an annotation chain can be decoded to ISO/TEI again.

The WebLicht web interface is a good way of getting acquainted with and testing different annotation tool chains for a given piece of data. However, for many users who have already decided on a tool chain, it may be more convenient to apply it directly inside the tool with which the data are created. For that purpose, we have integrated functionality for calling the SaaS variant of WebLicht – WebLicht as a Service (WaaS[25]) – out of the EXMARaLDA Partitur-Editor desktop tool. A typical workflow that is served by that functionality would proceed as follows:

1. A transcription is created in the EXMARaLDA Partitur-Editor, or imported there from a CLAN, ELAN, FOLKER, Praat or Transcriber file.
2. A tool chain for WebLicht is constructed by:
   a. Exporting a TCF version of the transcription from the Partitur-Editor
   b. Uploading that TCF version to the WebLicht web interface
   c. Constructing and testing a suitable tool chain inside the web interface
   d. Saving the tool chain locally
3. Users obtain a valid WaaS key[26] using Single sign-on (based on Shibboleth/CLARIN SPF)

[19] The latter variant is consistent with the patterns for the other MIME types. For reasons of backward compatibility, however, the WebLicht developers prefer to stick to the first variant

[20] Note that "application/tei+xml" is a registered MIME type, and that the MIME type for the TEI format defined at the Deutsches Textarchiv (DTA) for encoding (historical) written text has been decided upon in this process to be "application/tei+xml; format-variant=tei-dta"

[21] An additional parameter "charset" might be used to specify the encoding, as is common for plain text files

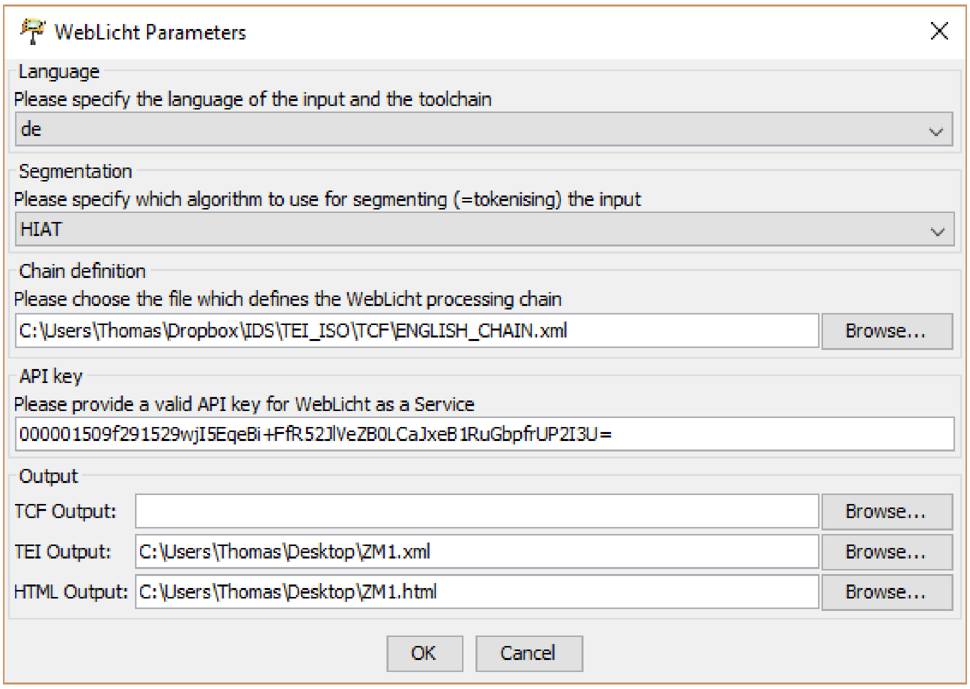
[22] This is a suggestion only - none of the services described here consumes or produces eaf files, so the MIME type is nowhere used

[23] This is a suggestion only - none of the services described here consumes or produces textGrid files, so the MIME type is nowhere used. An additional parameter "charset" might be used to specify the encoding, as is common for plain text files

[24] All conversion services described here are now online in a beta version (see Appendix). Full WebLicht integration is still pending, but we expect it to be complete by the time this paper is due to be published.

[25] See: https://weblicht.sfs.uni-tuebingen.de/WaaS/

[26] See: https://weblicht.sfs.uni-tuebingen.de/WaaS/apikey



WebLicht Parameters

Language

Please specify the language of the input and the toolchain

de

Segmentation

Please specify which algorithm to use for segmenting (=tokenising) the input

HIAT

Chain definition

Please choose the file which defines the WebLicht processing chain

C:\Users\Thomas\Dropbox\IDS\TEI_ISO\TCF\ENGLISH_CHAIN.xml

API key

Please provide a valid API key for WebLicht as a Service

000001509f291529wjI5EqeBi+FfR52JlVeZB0LCaJxeB1RuGbpfrUP2I3U=

Output

TCF Output:

TEI Output: C:\Users\Thomas\Desktop\ZM1.xml

HTML Output: C:\Users\Thomas\Desktop\ZM1.html

Figure 12: Parameter Dialog for WebLicht in the EXMARaLDA Partitur-Editor

The tool chain can then be used on the original data and further datasets from the same collection by calling “WebLicht…” from the CLARIN menu in the Partitur-Editor. This will bring up the dialog from Figure 12, in which parameters for WaaS can be specified.

After clicking on “OK”, EXMARaLDA will perform the necessary conversions to TCF, send the resulting file to WaaS alongside the specified parameters, receive the reply from the service and store it locally, either as the TCF obtained from WaaS, as an ISO/TEI file (after applying the decoding converter), or as an HTML file representing visually the different annotations added in WebLicht.

7 ISO/TEI-based Web Services

While TCF encoding and decoding is a workable solution for using existing services in WebLicht on spoken language data, it is, in the long run, not an optimal way of dealing with such data.

Obviously, the information lost in the TCF encoding process has a potential value for many automatic processing methods. For example, a POS tagger optimized for spoken language might make use of n-gram statistics involving information about pauses, or a lemmatizer may ignore, or treat in a special manner, defect tokens such as aborted words. Likewise, some services may want to use orthographic normalizations (see figure 5 above), time alignment information or even the corresponding section of the audio or video signal data. These pieces of information are available all in the ISO/TEI format, but not in TCF.

Moreover, in some cases, basic assumptions tacitly included in TCF’s data model may turn out to be overly simplified when applied to spoken language data. For instance, TCF specifies “the language of the data [emphasis added]”[27]inside a single attribute @lang on the <textCorpus> element. Multilingual interactions, such as interpreted talk, however, contain by definition data in at least two languages, and there is no way of telling a TCF based tool which part of the document is in which language. In the ISO/TEI format, by contrast, an @xml:lang can optionally be used on every element in the document to provide such information.

Ideally, automatic annotation tools and services optimized for spoken language transcription would therefore operate directly on the ISO/TEI format, and the detour via TCF conversion would then become unnecessary. One existing example for this is an adaptation of TreeTagger and the STTS tagset for use with transcriptions from the FOLK project (see Westpfahl and Schmidt, 2016). We are currently working on turning this mechanism into a CLARIN compliant (but not TCF compatible) web service. Similar tools developed in the context of the AGD and HZSK corpora, such as a tool for automatic orthographic

[27] See: http://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/The_TCF_Format

normalization of transcripts, and a tool for fine-aligning transcripts via MAUS, could be treated in an analogous manner.

8 Conclusion

As the present contribution shows, there are ways of making existing CLARIN components and architectural concepts that were originally or mainly developed for canonical written language data usable also for spoken language data. The issue of standardization is crucial in this task, since, without a widely used and sufficiently specified common basis, the number of processing steps needed to convert a given piece of data from and to a form usable by WebLicht would multiply. The newly published ISO/TEI standard provides such a common basis for various established tool formats and transcription conventions for spoken language data. While being able to represent common time-based annotation formats consistently and without information loss, its hierarchical structure facilitates conversion to traditional formats for canonical written language data.

Since conversion to these formats is however not possible without information loss, we believe that spoken language data should ideally be processed by services aware of and capable of interpreting the complexity and peculiarities inherent to this type of data. In a wider scope, CLARIN as an infrastructure also needs to consider and meet the requirements and expectations of various research communities working with different types of non-canonical and/or non-written language data. Though this is a challenging task to take on, considering the increasing availability of audio and video data and thus its increasing importance in linguistic and language based research, it seems crucial to keep up with this development in order for CLARIN to become and remain equally available and important to researchers from all targeted fields.

Another critical aspect of a successful digital research infrastructure for language resources and technology is a common set of relevant standards recognized by all components of the infrastructure. This requires relevant file formats to be identified and described on a level specific enough to allow for automatic processing. Such descriptions are necessary to be able to inform not only other components within the infrastructure – but also the users – about the technical characteristics of language resources and the suitability and interoperability of existing tools and services. The experiences with processing various flavors of TEI within WebLicht may serve as a useful pilot experiment for future efforts to provide researchers with reliable guidance in choosing tools and services for their workflows and processing pipelines.

Consequently, there is also work to be done regarding the various widely used de facto standard annotation formats in order to allow for a consistent use of MIME types and possible complementary descriptions within the CLARIN infrastructure. As an international organization with a strong network, CLARIN could play an important role in this matter, creating an impact that goes beyond the infrastructure itself, e.g. through coordination of the registration of relevant formats with existing standardization bodies.

References

[Bański et al. 2016] Bański, Piotr, Bertrand Gaiffe, Patrice Lopez, Simon Meoni, Laurent Romary, Thomas Schmidt, Peter Stadler, and Andreas Witt. 2016. In Claudia Resch, Vanessa Hannesschläger, and Tanja Wissik, editors, *TEI Conference and Members' Meeting 2016 – Book of Abstracts*, Vienna, pages 35-37. http://tei2016.acdh.oeaw.ac.at/sites/default/files/TEIconf2016_BookOfAbstracts.pdf

[Barras et al. 2000] Barras, Claude, Edouard Geoffrois, Zhibiao Wu, and Mark Liberman. 2001. Transcriber: development and use of a tool for assisting speech corpora production. *Speech Communicatin – Special issue on Speech Annotation and Corpus Tools*, 33(1-2):5–22. http://languagelog.ldc.upenn.edu/myl/Barras2001.pdf. DOI: 10.1016/S0167-6393(00)00067-4.

[Bird and Liberman 2001] Bird, Steven and Mark Liberman. 2001. A formal framework for linguistic annotation. *Speech Communicatin – Special issue on Speech Annotation and Corpus Tools*, 33(1-2):23–60. http://languagelog.ldc.upenn.edu/myl/BirdLiberman2001.pdf. DOI: 10.1016/S0167-6393(00)00068-6.

[Boersma 2014] Boersma, Paul. 2014. The Use of Praat in Corpus Research. In Jacques Durand, Ulrike Gut, and Gjert Kristoffersen, editors, *The Oxford Handbook of Corpus Phonology*. Oxford University Press, Oxford, UK. DOI: 10.1093/oxfordhb/9780199571932.013.016.

[Brouwer and Kemps-Snijders 2016] Brouwer, Maththijs and Kemps-Snijders, Marc. 2016. A SOLR/Lucene based Multi Tier Annotation Search solution. *Proceedings of the CLARIN Annual Conference* (CAC), 2016, Aix, France. https://www.clarin.eu/sites/default/files/brouwer-kempssnijdersCLARIN2016_paper_21.pdf

[Eckart de Castilho et. al 2014] Eckart de Castilho, Richard, Biemann, Chris, Gurevych, Irina and Yimam, S.M. 2014. WebAnno: a flexible, web-based annotation tool for CLARIN. In *Proceedings of the CLARIN Annual Conference* (CAC) 2014, Soesterberg, Netherlands.

[Eshkol-Taravella et. al 2012] Eshkol-Taravella, Iris, Oliver Baude, Denis Maurel, Linda Hriba, Céline Dugua, and Isabelle Tellier. 2012. Un grand corpus oral «disponible»: le corpus d'Orléans 1968-2012. *Ressources linguistiques libres*, 52(3/2011):17–46. https://www.atala.org/IMG/pdf/Eshkol-TAL52-3.pdf.

[Groupe ICOR 2017] GROUPE ICOR (Heike Baldauf-Quilliatre, Isabel Colón de Carvajal, Carole Etienne, Emilie Jouin-Chardon, Sandra Teston-Bonnard, and Véronique Traverso) (in press). CLAPI, une base de données multimodale pour la parole en interaction : apports et dilemmes. In Mathieu Avanzi, Marie-José Béguelin, and Federica Diémoz, editors, *Corpus de français parlés et français parlés des corpus,* Cahiers Corpus, Université de Neuchâtel.

[Hinrichs and Vogel 2010] Hinrichs, Erhard and Iris Vogel. 2010. CLARIN - Interoperability and Standards. In *CLARIN devliverable D5.C-3*. http://www-sk.let.uu.nl/u/D5C-3.pdf.

[Hinrichs et al. 2010] Hinrichs, Marie, Thomas Zastrow, and Erhard Hinrichs. 2010. WebLicht: Web-based LRT Services in a Distributed eScience Infrastructure. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valetta, Malta. http://www.lrec-conf.org/proceedings/lrec2010/pdf/270_Paper.pdf.

[IETF] Internet Engineering Task Force. 2013. *Media Type Specifications and Registration Procedures – Best Practices*. https://tools.ietf.org/html/rfc6838#section-3.4.

[ISO 2016] ISO 2462:2016. *Language resource management – Transcription of spoken language*. http://www.iso.org/iso/catalogue_detail.htm?csnumber=37338.

[Jettka & Stein 2014] Jettka, Daniel, Daniel Stein. 2014. The HZSK Repository: Implementation, Features, and Use Cases of a Repository for Spoken Language Corpora. *D-Lib Magazine*, 20(9/10). http://www.dlib.org/dlib/september14/jettka/09jettka.html. DOI: 10.1045/september2014-jettka.

[Kipp 2014] Kipp, Michael. 2014. ANVIL: A Universal Video Research Tool. In Jacques Durand, Ulrike Gut, and Gjert Kristoffersen, editors, *The Oxford Handbook of Corpus Phonology*. Oxford University Press, Oxford, UK. DOI: 10.1093/oxfordhb/9780199571932.013.024.

[Kisler et al. 2010] Kisler, Thomas, Florian Schiel, and Han Sloetjes. 2012. Signal processing via web services: the use case WebMAUS. In *Proceedings of Digital Humanities 2012*, Hamburg, pages 30–34. http://clarin-d.de/images/workshops/proceedingssoasforthehumanities.pdf.

[Klein and Schütte 2001] Klein, Wolfgang, Wilfried Schütte. 2001. *Transkriptionsrichtlinien für die Eingabe in DIDA*. Institut für Deutsche Sprache, Mannheim, Germany. http://agd.ids-mannheim.de/download/dida-trl.pdf.

[Krause and Zeldes 2016] Krause, Thomas and Zeldes, Amir. 2016. ANNIS3: A new architecture for generic corpus query and visualization. In *Digital Scholarship in the Humanities 2016 (31).* http://dsh.oxfordjournals.org/content/31/1/118.

[MacWhinney 2000] MacWhinney, Brian. 2000. *The CHILDES Project: Tools for Analyzing Talk*. Psychology Press, New York, USA.

[Menke et al. 2015] Menke, Peter, Farina Freigang, Thomas Kronenberg, Sören Klett, and Kirsten Bergmann. 2015. First Steps towards a Tool Chain for Automatic Processing of Multimodal Corpora. *Journal of Multimodal Communication Studies*, 2:30–43. http://jmcs.home.amu.edu.pl/wp-content/uploads/2015/09/Menke_et_al_2014_JMCS.pdf.

[Parisse and Morgenstern 2010] Parisse, Christophe and Aliyah Morgenstern. 2010. A multi-software integration platform and support for multimedia transcripts of language. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)* [Workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality], Valetta, Malta. http://www.lrec-conf.org/proceedings/lrec2010/.

[Rehbein et al. 2004] Rehbein, Jochen, Thomas Schmidt, Bernd Meyer, Franziska Watzke, and Annette Herkenrath. 2004. Handbuch für das computergestützte Transkribieren nach HIAT. In *Arbeiten zur Mehrsprachigkeit*, volume 56. http://www.exmaralda.org/files/azm_56.pdf.

[Schmidt 2005] Schmidt, Thomas. 2005. Time-based data models and the Text Encoding Initiative's guidelines for transcription of speech. In *Arbeiten zur Mehrsprachigkeit (Folge B)*, volume 62. http://www.exmaralda.org/files/SFB_AzM62.pdf.

[Schmidt et al. 2009] Schmidt, Thomas, Susan Duncan, Oliver Ehmer, Jeffrey Hoyt, Michael Kipp, Dan Loehr, Magnus Magnusson, Travis Rose, and Han Sloetjes. 2009. An Exchange Format for Multimodal Annotations. In Michael Kipp, Jean-Claude Martin, Patrizia Paggio, Dirk Heylen, editors, *Multimodal Corpora*. Springer, Berlin, Germany. http://link.springer.com/chapter/10.1007/978-3-642-04793-0_13. DOI: 10.1007/978-3-642-04793-0_13.

[Schmidt 2011] Schmidt, Thomas. 2011. A TEI-based Approach to Standardising Spoken Language Transcription. *Journal of the Text Encoding Initiative*, 1:1–22. http://jtei.revues.org/142. DOI: 10.4000/jtei.142.

[Schmidt 2012] Schmidt, Thomas. 2012. EXMARaLDA and the FOLK tools. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. http://www.lrec-conf.org/proceedings/lrec2012/pdf/529_Paper.pdf.

[Schmidt 2014] Schmidt, Thomas. 2014. The Database for Spoken German – DGD2. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. http://www.lrec-conf.org/proceedings/lrec2014/pdf/171_Paper.pdf.

[Schmidt and Wörner 2014] Schmidt, Thomas and Kai Wörner. 2014. EXMARaLDA. In Jacques Durand, Ulrike Gut, and Gjert Kristoffersen, editors, *The Oxford Handbook of Corpus Phonology*. Oxford University Press, Oxford, UK. DOI: http://dx.doi.org/10.1093/oxfordhb/9780199571932.013.030.

[Schmidt et al. 2015] Schmidt, Thomas, Wilfried Schütte, and Jenny Winterscheid. 2015. *cGAT. Konventionen für das computergestützte Transkribieren in Anlehnung an das Gesprächsanalytische Transkriptionssystem 2 (GAT2)*. https://ids-pub.bsz-bw.de/frontdoor/index/index/docId/4616.

[Selting et al. 2009] Selting, Margret, Peter Auer, Dagmar Barth-Weingarten, Jörg Bergmann, Pia Bergmann, Karin Birkner, Elizabeth Couper-Kuhlen, Arnulf Deppermann, Peter Gilles, Susanne Günthner, Martin Hartung, Friederike Kern, Christine Mertzlufft, Christian Meyer, Miriam Morek, Frank Oberzaucher, Jörg Peters, Uta Quasthoff, Wilfried Schütte, Anja Stukenbrock, and Susanne Uhmann. 2009. Gesprächsanalytisches Transkriptionssystem 2 (GAT 2). *Gesprächsforschung - Online-Zeitschrift zur verbalen Interaktion*, 10:353–402. http://www.gespraechsforschung-ozs.de/heft2009/px-gat2.pdf.

[Sloetjes 2014] Sloetjes, Han. 2014. ELAN: Multimedia Annotation Application. In Jacques Durand, Ulrike Gut, and Gjert Kristoffersen, editors, *The Oxford Handbook of Corpus Phonology*. Oxford University Press, Oxford, UK. DOI: 10.1093/oxfordhb/9780199571932.013.019.

[TEI] Text Encoding Initiative. 2015. *Guidelines*. http://www.tei-c.org/Guidelines/.

[Westpfahl and Schmidt 2016] Westpfahl, Swantje and Thomas Schmidt. 2016. FOLK-Gold – A GOLD standard for Part-of-Speech-Tagging of Spoken German. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia. http://www.lrec-conf.org/proceedings/lrec2016/pdf/397_Paper.pdf.

[Zipser and Romary 2010] Zipser, Florian and Romary, Laurent. 2010. A model oriented approach to the mapping of annotation formats using standards. In: Proceedings of the Workshop on Language Resource and Language Technology Standards, LREC 2010. Malta. http://hal.archives-ouvertes.fr/inria-00527799/en/.

Appendix A. Details for the web services

Beta versions of the conversion services are available as listed below, the source code for all services is available on GitHub: https://github.com/hzsk/HZSK-CLARIN-Services/.

EXMARaLDA to ISO/TEI conversion

PID:	http://hdl.handle.net/11022/0000-0000-9A6C-A
Consumes:	application/xml; format-variant=exmaralda-exb
Produces:	application/tei+xml; format-variant=tei-iso-spoken
Parameters:	seg - the segmentation algorithm to be used, one of (HIAT|cGAT|CHAT)
	lang - the language of the document, two letter ISO language code

FOLKER to ISO/TEI conversion

PID:	http://hdl.handle.net/11022/0000-0001-B538-4
Consumes:	application/xml; format-variant=folker-fln
Produces:	application/tei+xml; format-variant=tei-iso-spoken
Parameters:	lang - the language of the document, two letter ISO language code

Transcriber to ISO/TEI conversion

PID:	http://hdl.handle.net/11022/0000-0001-B539-3
Consumes:	application/xml; format-variant=transcriber-trs
Produces:	application/tei+xml; format-variant=tei-iso-spoken
Parameters:	seg - the segmentation algorithm to be used, one of (HIAT|cGAT|CHAT)
	lang - the language of the document, two letter ISO language code

CHAT to ISO/TEI conversion

PID:	http://hdl.handle.net/11022/0000-0001-B53A-2
Consumes:	text/plain;format-variant=clan-cha
Produces:	application/tei+xml; format-variant=tei-iso-spoken
Parameters:	seg - the segmentation algorithm to be used, one of (HIAT|cGAT|CHAT)
	lang - the language of the document, two letter ISO language code

ISO/TEI to TCF conversion

PID:	http://hdl.handle.net/11022/0000-0001-B53B-1
Consumes:	application/tei+xml; format-variant=tei-iso-spoken
Produces:	application/xml; format-variant=weblicht-tcf

TCF to ISO/TEI conversion

PID:	http://hdl.handle.net/11022/0000-0001-B53C-0
Consumes:	application/xml; format-variant=weblicht-tcf
Produces:	application/tei+xml; format-variant=tei-iso-spoken

Researcher Hands-On Training in the Digital Humanities: The ACDH Tool Gallery as an Austrian Case Study

Tanja Wissik
Austrian Centre for Digital Humanities
Austrian Academy of Sciences, Austria
tanja.wissik@oeaw.ac.at

Claudia Resch
Austrian Centre for Digital Humanities
Austrian Academy of Sciences, Austria
claudia.resch@oeaw.ac.at

Abstract

In this paper we discuss practical experiences with hands-on training in the Digital Humanities based on an Austrian case study. We will present the “ACDH Tool Galleries”, an initiative organised by the Austrian Centre for Digital Humanities (ACDH) of the Austrian Academy of Sciences. This series of educational events aims to create a platform for developers and professionals to share their expertise and provide education and practical training opportunities for users of Digital Humanities tools. In order to give insight into the ways this initiative has been received by the community, we present survey data collected among the participants of these training courses.

1 Introduction

Although there is now a wide variety of computational tools and digital methods available for humanities scholars to use in their research, it has been observed that not all tools are adopted with equal enthusiasm by the researchers who would benefit most of them (Kemman and Kleppe, 2015). Considering that a lack of familiarity or practical know-how regarding the available options may lie at the root of this hesitance, researcher training can play an important role in promoting a more far-reaching utilisation of computational tools and digital methods in the humanities.

In this paper, we explore the potential of researcher training for spreading information about the available tools to potential users, making particular reference to an Austrian case study. After outlining the DH teaching and training landscape in Austria, we present a recently established hands-on researcher training series and discuss survey data collected among the participants of the training initiative.

2 Researcher training in the Digital Humanities

Given that the digital humanities are a fairly new field of research, one aspect of fostering the advance of the discipline is its inclusion in academic education at various stages and levels, be it BA, MA and PhD courses or summer schools (cf. Sahle, 2013). However, there are considerable disparities in the degree to which DH training programs have been cultivated in different countries. In Austria, for example, the first professorships for DH were appointed as late as 2016 and DH curricula at universities are still in their infancy[1]. As an interim solution, summer schools and workshops are useful and well-established formats for conveying skills in the digital humanities (Rehbein and Fritze, 2012). One example for a provider of summer schools is the international Digital Humanities Training Network, where several summer school organisers collaborate with each other (DH Training Network, 2016). A similar organisation offering workshops is the Digital Scholarship Training Programme of the British Library (McGregor et al., 2016). Summer schools and workshops are not only useful where there is a lack of university level courses, they can also have a more diversified target audience than university courses. For example, they can address established researchers in more advanced academic positions who are now confronted with new technologies and ways to carry out their research. Given the tight time schedule of researchers, one- or two-day workshops are preferable to training programs that require a long-term commitment, as a typical masters course in DH would. In the following section, we present

[1] We have other offerings such as minors etc. For more information see https://registries.clarin-dariah.eu/courses/courses/.

Tanja Wissik and Claudia Resch 2017. Researcher hands-on training in the digital humanities: The ACDH tool gallery as an Austrian case study. *Selected papers from the CLARIN Annual Conference 2016*. Linköping Electronic Conference Proceedings 136: 131–138.

a case study of a hands-on training event series initiated in Austria. We use the term "series" because it is not a unique event and we use "hands-on" to acknowledge the fact that the practical application of digital methods and skills play an important role in DH.

3 The ACDH Tool Gallery – an Austrian case study

The idea of the ACDH Tool Galleries was to allow developers and professionals to share their theoretical knowledge on the tools designed for DH users and provide practical training in their use (Wissik and Resch, 2016). In order to reflect this two-pronged approach, we opted for a format that combined short lectures scheduled in the morning with hands-on training sessions scheduled in the afternoon. During these sessions, the experts lead the group step-by-step through the features and functionalities of various tools. Although one day is usually not sufficient to master the use of the tools or services in question, participants can use the opportunity to get an overview of their options as well as the potential benefits of using particular tools or services in their own field of research.

Considering that the institute is often confronted with very basic questions regarding the use of various tools, the training events offer a good opportunity to establish connections between tool developers and DH users and to initiate discussions regarding the scope of application in the respective fields. The hands-on part of the training session is particularly valuable as it gives the attendees the chance to immediately consult with tool experts if they encounter a problem during the workshop. This guarantees participants a safe and guided start in their exploration of new tools and services. Furthermore, the practical experience helps participants to evaluate the features and abilities of various tools and to become aware of difficulties or limitations.

Since the inception of the program in 2015, two seasons of ACDH Tool Galleries have already been completed and a third is ongoing. Each season included three ACDH Tool Galleries: two ACDH Tool Galleries with morning lectures and a hands-on session in the afternoon and one 'Extended Version' of the ACDH Tool Gallery. The latter was embedded in the Digital Humanities Austria Conferences and featured a whole day of theoretical presentations followed by a day of hands-on sessions where the participants could experiment with different tools in a bazaar-like atmosphere. So far, the ACDH Tool Galleries have covered topics like handwriting recognition, linguistic annotation, semantic technologies, data management, text encoding with TEI as well as network and visualisation tools.

3.1 Promotion and preliminary organisational efforts

While the Tool Galleries were originally conceived as a service for employees of the academy, the format was soon extended to a larger audience. This happened quite organically as the original recipients of the Tool Gallery newsletters forwarded and shared the announcements with their contacts. Additionally, the dates were made public via the academy's event calendar and various mailing lists. Since July 2016, the ACDH has also been using Twitter to promote the Tool Galleries. The institute's website was used both to promote the events and to publish presentation material, exercises and tutorials after the events.

By organising Tool Galleries three times a year, the ACDH hopes to achieve a certain regularity and continuity concerning the initiative. This objective is also reflected in the styling and appearance of the promotion material, where we aimed to establish a "brand" with a certain recognition value by using recurring design elements and a recognisable logo for each new event.

Figure 1: ACDH Tool Gallery Logo

As the number of available places was limited for organisational reasons, prospective participants were asked to register via an online form. To emphasize the educational character of the format, the ACDH also offered an official certificate of participation for those who wanted documentation of their attendance. So far, no ECTS credits have been assigned, but this might be an option up for consideration in the future.

3.2 ACDH Tool Gallery 1.2 on (basic) linguistic annotation

As mentioned above, each of the training sessions in the program was dedicated to a different digital research tool. The ACDH Tool Gallery 1.2, for instance, put its focus on (basic) linguistic annotation and was addressed to both linguists and professionals from all text-based disciplines. The first talk, given by Ulrich Heid (University of Hildesheim), introduced the audience to the relevance of linguistic annotation and was followed by two short project contributions that demonstrated the possibilities and challenges of automatic annotation. After Heid's presentation, annotation examples from two ACDH-based projects, the Austrian Baroque Corpus (*ABaC:us*) and the Austrian Media Corpus, were introduced. *ABaC:us* (☞ https://acdh.oeaw.ac.at/abacus/), which is part of the CLARIN Centre Vienna and its Language Resource Portal, is a historical language resource containing Austrian literary sources from the 17th and 18th century. The Austrian Media Corpus (http://www.oeaw.ac.at/acdh/de/amc) is a large collection of media texts from Austrian newspapers and magazines as well as press releases and transcribed television interviews spanning the last three decades.

The second block of Tool Gallery 1.2 was presented by Marie Hinrichs and Claus Zinn of the University of Tübingen, who introduced the participants to *Weblicht*, a 'web-based linguistic chaining tool' (Hinrichs et al., 2010). We chose this app for being the most suitable research environment for demonstrating the automatic annotation of texts. Hinrichs and Zinn presented its fully functional processing chain, which features linguistic tools such as tokenizers, part of speech taggers, parsers etc., and showed how these services can be customised and combined by the user. While *Weblicht* is well known among linguists, the event was an occasion for those from other text-based disciplines to learn about the benefits and potentials of automatic basic linguistic annotation. The idea that participants should bring their own texts in order annotate them and visualize the results in an appropriate way was set into practice under the guidance of the experts and made the hands-on session quite lively.

The Tool Gallery was concluded with a presentation by researchers from the institute itself. They presented the recently developed *tokenEditor* (http://www.oeaw.ac.at/acdh/de/tokenEditor), a web application for the manual annotation (or the manual review of automatic annotations) of texts.

4 Survey data based on the ACDH Tool Gallery

In order to evaluate the new format and its reception, the ACDH undertook a survey of the participants. The survey takes into account online registration data (365 registered participants) as well as data collected via anonymous questionnaires from 188 participants of the six ACDH Tool Galleries that took place between 2015 and 2016. The questionnaires were handed out on site and were collected at the end of each of these events. We opted for paper questionnaires that were distributed during the events as opposed to online questionnaires. That way, we could ensure that they were filled out immediately after the workshop, when impressions were still fresh and organisers were on hand to provide clarification where needed. We also expected it to be easier to motivate participants to take the survey on site than via email communication. The questionnaires were divided into three sections: one reflected the topic of the given training event, a second concerned the specific format of the training event, and a third was designed to collect basic personal data (e.g. age, occupation). In the following section of the paper, we outline the general results of the survey and discuss, exemplarily, the results of the content section of the survey for the ACDH Tool Gallery 1.2 on (basic) linguistic annotation. We have chosen this specific event to show the relation and synergies between the training events and CLARIN.

4.1 General analysis

Since the ACDH Tool Gallery was initially conceived as an in-house training opportunity, it is not surprising that nearly half of the 365 registered participants, namely 48%, came from various departments of the Austrian Academy of Sciences. In addition to that, 21% came from the University of Vienna and 31% from other universities. Considering that the Tool Galleries are organised as a series of events and take place three times a year, it is also useful to look at the amount of repeat participants. Of the registered participants, 16% registered for two events, 13% registered for three or more events and 71% of the participants participated only once.

The age distribution of the 188 survey respondents was as follows: 67% of participants were between the ages of 20 and 40, 30% were between 40 and 60 and 3% were over 60 years old. For a subset of the questionnaires (154 questionnaires[2]) we could even make a more granular age analysis: Here, 40% of the participants were between the ages of 30 and 40, 27% were between the ages 20 and 30, 20% were between the ages 40 and 50,11% between the ages 50 and 60, and 1% was over 60 years old. This shows that academics of all career stages attended the training events. Regarding their disciplinary background (Fig. 2), most of the participants were scholars in the humanities, followed by information scientists and archivists. The category "other" includes, for example, librarians, lexicographers and IT coordinators. Within the humanities, the participants came from a wide range of disciplines, such as archaeology, history studies, musicology, linguistics, literary studies and theatre studies.

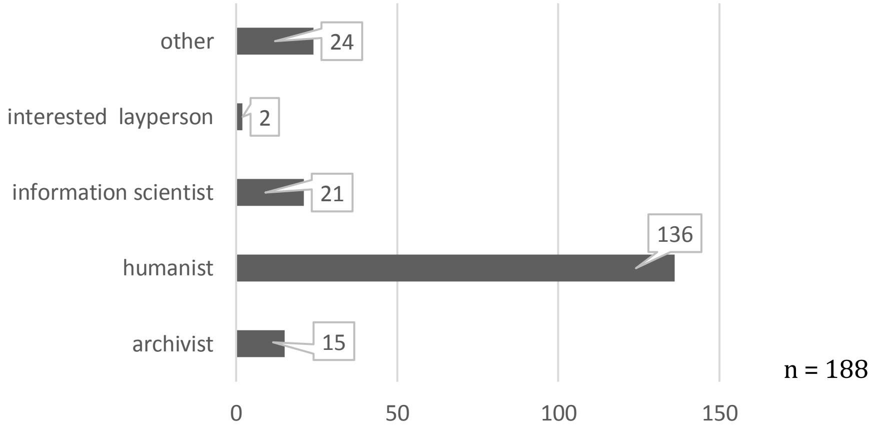


Figure 2: Occupation / disciplinary background

Having put a lot of effort in the advertising of the training events, we also wanted to obtain feedback on the effectiveness of the different communication channels used. As the survey showed, the majority of participants had heard about the event from colleagues or via one of the mailing lists. Other participants indicated that they got the information from the institute's website or the academy's calendar of events. In the category "other", participants specified that they had seen the announcements on the respective conference websites (in cases where the ACDH Tool Gallery took place within the context of a bigger event) and on Twitter. As the ACDH Twitter presence was not launched before July 2016, only the last two ACDH Tool Galleries had been promoted actively via our own Twitter account. Nevertheless, even before that, the ACDH Tool Galleries were mentioned on the private Twitter accounts of staff members and participants.

[2] In this analysis, the 34 questionnaires from the first ACDH Tool Gallery 1.1 were excluded, because they did not contain a more granular age information.

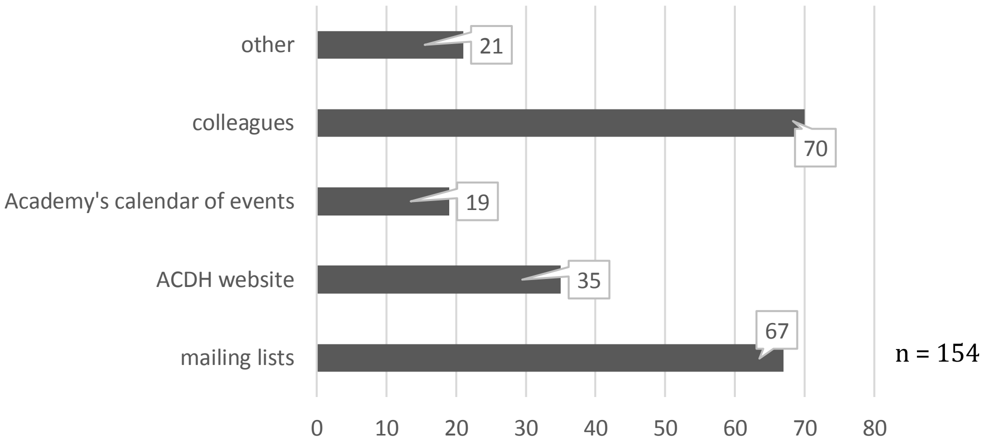


Figure 3: Dissemination channels[3]

One section of the questionnaire inquired how useful the participants found the combination of lecture and hands-on session in these training events. Of the 188 respondents, 185 respondents generally or fully agreed that the combination of lectures and hands-on session was useful; only three participants did not find it useful and one person did not specify (see figure 4).

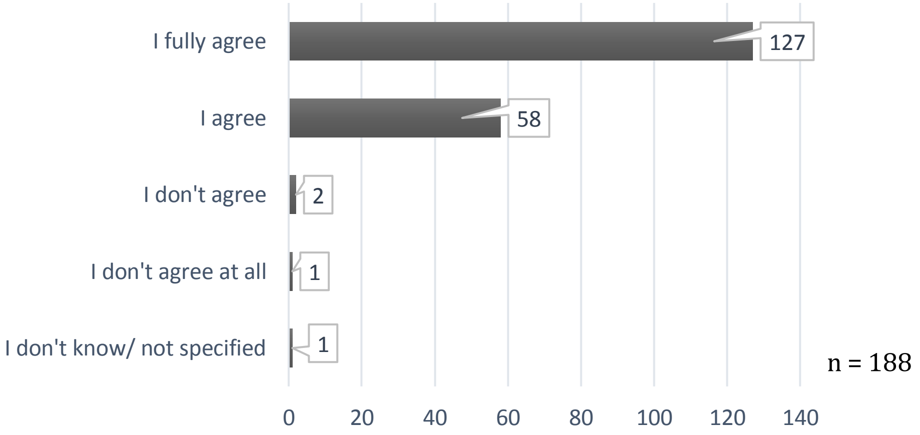
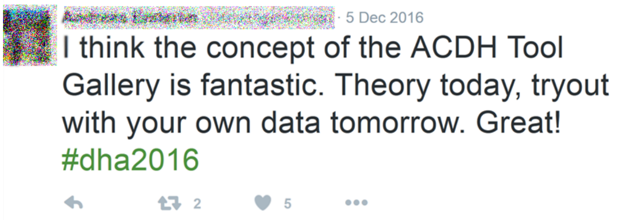


Figure 4: Usefulness of combined lecture & hands-on training

Participants also had the opportunity to leave feedback via the free commentary field in the questionnaire or via Twitter (see figure 5). Feedback received this way was generally laudatory. This positive reception is also reflected in the answers regarding the question "Would you recommend and re-attend an ACDH Tool Gallery event?": 40% agreed or 5% fully agreed with the statement, 4% did not know if they would recommend or re-attend it and 1% stated they would not recommend and re-attend the ACDH Tool Gallery.


[redacted] · 5 Dec 2016

I think the concept of the ACDH Tool Gallery is fantastic. Theory today, tryout with your own data tomorrow. Great! #dha2016

[3] The total number of questionnaires for this question was 154 because in the questionnaires during the first event the question regarding the dissemination channels was not included yet.

Figure 5: Comment via Twitter

For the organisation of the most recent Tool Gallery events, we also aimed to take into account the needs and wishes of our participants. For this purpose, we included a section in the surveys to inquire what kind of tool they would be most interested in exploring. Some suggestions were provided by us, but participants were also given the opportunity to additionally propose tools they had heard about or wanted to get a deeper understanding of. The analysis of these questionnaires has repeatedly shown that there is still a strong demand for tools used for processing and encoding textual sources. With this information in mind, the last two Tool Galleries of the third season will focus on the features of the XML editor Oxygen and text encoding according to the TEI guidelines.

In the course of the almost three seasons of Tool Galleries that have so far been organised, we have also observed that there is continuous interest in topics that nearly every researcher is concerned with when working with original source material: Questions such as how to store, structure, manage and share data. In reaction to this insight, we have also offered a course introducing various tools for creating a data management plan. Finally, a Tool Gallery dedicated to the topic of licensing will complement these items in the program. Its focus will lie on providing information on existing guidelines and directives as well as advice in handling legal issues.

4.2 Specific analysis from ACDH Tool Gallery 1.2 on (basic) linguistic annotation

As has been mentioned, the ACDH Tool Gallery 1.2 was dedicated to the presentation of tools for the support of text annotation, particularly linguistic annotation. De Jong et al. (2011) observed that, “[h]umanities researchers can hardly be indifferent to the promise of innovative tools for the support of content exploration and content annotation. Both are key elements in their daily research practice and as such can be considered the alpha and omega of their analytical and comparative work.” While we have records of some computer scientists attending the workshops, in our case, most of the participants were indeed researchers from the humanities, more precisely from history, history of art, musicology, Indology, literary studies, Slavic studies and English studies.

In the ACDH Tool Gallery 1.2, we counted 43 registered participants and received 22 survey responses. On the basis of these 22 questionnaires it could be observed that the majority of workshop participants (77%) had prior experience with the use of digital tools and methods in their research, while 23% stated they were interested in using them in the future. Going into more detail, 45% of the respondents had already used (linguistic) annotation tools, and 55% had no prior experience with them. A breakdown of these figures according to age groups can be seen in figure 6.

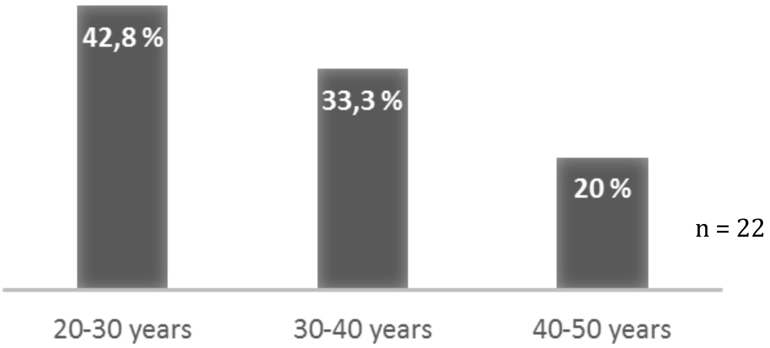


Figure 6: Percentage of respondents from ACDH Tool Gallery 1.2 using digital methods according to age groups

Of those with prior experience, several respondents mentioned *TreeTagger*, but none had any prior experience with *Weblicht*. Hence, we assume that the Tool Gallery was a good opportunity to make the *Weblicht* application better known in Austria and advertise it outside the CLARIN community, especially among historians and literary scholars.

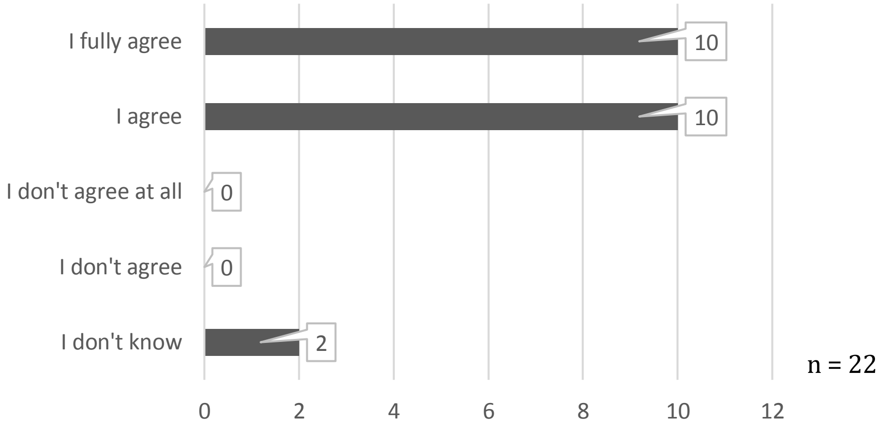


Figure 7: Use of linguistic annotations in further research

Despite the lack of prior familiarity with (linguistic) annotation tools at the beginning of the training event (roughly half of the participants had no prior experience), all of the respondents agreed or fully agreed that they got an overview of the use of linguistic annotation and of what research questions could be answered with the help of linguistic annotation. Furthermore, nearly all of them (20 out of 22) agreed or fully agreed that they would be interested in linguistically annotating their research material in the future (see figure 7). Moreover, nearly all the participants (21 out of 22) agreed or fully agreed that additional linguistic annotation would make their resources more interesting for other research disciplines. However, 19% of the participants agreed or fully agreed that the linguistic annotation of their own resources would be too time-consuming, particularly if the computer-generated annotations needed further manual correction. Participants' opinions were split concerning their faith in their ability to undertake these manual revisions themselves: 47% agreed that they would have to be done by linguists, while 48% disagreed and 5% did not have an opinion.

The ACDH Tool Gallery has demonstrated the importance of linguistic annotation in DH projects, since linguistic annotation can serve as a starting point for the further annotation or processing of texts. It facilitates information extraction and allows for the calculation of frequencies and distributions. For example, when studying historical correspondences, changing power structures may be observed through varying forms of address. In linguistically annotated text, address patterns can be searched systematically (e.g. adjective noun combinations) and their frequencies and variation over time can be measured. Historical texts, in particular, might need additional lemma information in order for full-text searches to turn up all instances of a term despite the existence of orthographic variants.

5 Conclusion

In this paper, we have presented the ACDH Tool Galleries, a new research training event series for the digital humanities. As the survey results show, the events were very well received among members of the Austrian Academy of Sciences but also in other Austrian academic institutions. Our analysis shows that there is active demand for training events for researchers in the humanities at all career stages. Moreover, the format of the Tool Gallery can be used for the dissemination of tools and resources developed by research infrastructure consortia such as CLARIN, which could complement the CLARIN user involvement group's efforts on a national level (Wynne, 2015) and would be in accordance with the User Engagement Handbook (Wynne, 2015a). One of our goals for the future is to apply our experiences to a wider European context and to share our knowledge with other CLARIN members who intend to offer similar courses. To facilitate this exchange we plan to prepare a concluding report based on our experiences. At the same time, we are considering new approaches and strategies for conveying particular elements of the courses, for instance through short video introductions or webinars.

In order to make training events such as the ACDH Tool Galleries successful and effective, careful and anticipative organisation is needed. We agree with Rehbein and Fritze (2012) that the organisational effort is higher than in "traditional" seminars and the technical set up takes longer. Furthermore, we find it very important to foster interaction between persons from different disciplines. For the training event series, it is essential to invite ICT experts and researchers who have experience with the application of the tools in their own research and allow more inexperienced researchers to benefit from their expertise. This is in line with the idea that "[h]umanities scholars, and ICT-developers and students should all learn about the principles, challenges and biases of each other's discipline" (de Jong et al., 2011).

References

[DH Training Network 2016] DH Training Network (2016) http://www.culingtec.uni-leipzig.de/ESU_C_T/node/409 (20.06.2016)

[Hinrichs et al. 2010] Erhard Hinrichs, Marie Hinrichs, Thomas Zastrow. 2010. WebLicht: Web-based LRT services for German. *Proceedings of the ACL 2010 System Demonstrations*, 25-29, Uppsala, Sweden, 13 July 2010. http://www.aclweb.org/anthology/P10-4005 (21.06.2016)

[de Jong et al. 2011] Francisca 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. *Proceedings of the 2nd Conference on Supporting Digital Humanities (SDH 2011)*, Denmark. Centre for Language Technology, Copenhagen.

[Kemmer and Kleppe 2015] Max Kemmer and Martijn Kleppe. 2015. User Required? On the Value of User Research in the Digital Humanities. In: Jan Odijk (ed). *Selected Papers from the CLARIN 2014 Conference*, October 24-25, 2014, Soesterberg, The Netherlands, 63-74.

[McGregor et al. 2016] Nora McGregor, Mia Ridge, Stella Wisdom and Aquiles Alencar-Brayner. (2016). The Digital Scholarship Training Programme at British Library: Concluding Report & Future Developments. In *Digital Humanities 2016: Conference Abstracts*. Jagiellonian University & Pedagogical University, Kraków, 623-625. http://dh2016.adho.org/abstracts/178

[Rehbein und Fritze 2012] Malte Rehbein and Christiane Fritze. 2012. Hands-On Training Digital Humanities: A Didactic Analysis of a Summer School on Digital Editing. In: Brett D. Hirsch (ed.). *Digital Humanities Pedagogy: Practices, Principles and Politics*, 47-78.

[Resch and Czeitschner 2015] Claudia Resch and Ulrike Czeitschner. 2015. ABaC:us – Austrian Baroque Corpus. https://acdh.oeaw.ac.at/abacus/

[Sahle 2013] Patrick Sahle. 2013. "DH Studieren! Auf dem Weg zu einem Kern- und Referenzcurriculum der Digital Humanities". *DARIAH-DE Working Papers Nr. 1*. Göttingen: DARIAH-DE, 2013. URN: urn:nbn:de:gbv:7-dariah-2013-1-5.

[Wissik and Resch 2016] Tanja Wissik and Claudia Resch. 2016. Digitale Tools und Methoden für die geisteswissenschaftliche Forschung praxisnah erklärt: Ein neues Format im Test. In: *Digital Humanities 2016: Conference Abstracts.* Jagiellonian University & Pedagogical University, Kraków, 711-713. http://dh2016.adho.org/abstracts/87

[Wynne 2015] Martin Wynne. 2015. User Involvement. Presentation at the Clarin Annual Conference 2015. https://www.clarin.eu/sites/default/files/20151016-CAC-04-Wynne-User-Involvement-CAC2015-05.pdf

[Wynne 2015a] Martin Wynne. 2015a. User Engagement Handbook. Version 2.1. https://office.clarin.eu/v/CE-2015-0590-UserInvolvementHandbook-v2.1.pdf

Linköping Electronic Conference Proceedings 136
eISSN 1650-3740 (Online) • ISSN 1650-3686 (Print) 2017
ISBN 978-91-7685-499-0

Front cover illustration:
Group photo by Quirijn Backx, picture composition by Karolina Badzmierowska • CLARIN ERIC