CLARIN Annual Conference 2021

PROCEEDINGS

Edited by

Monica Monachini, Maria Eskevich

27 – 29 September 2021
Virtual Edition

Please cite as:
Proceedings of CLARIN Annual Conference 2021. Eds. M. Monachini and M. Eskevich.
Virtual Edition, 2021.

Programme Committee

Chair:

- Monica Monachini, Institute of Computational Linguistics “A. Zampolli" (IT)

Members:

- Lars Borin, University of Gothenburg (SE)
- António Branco, Universidade de Lisboa (PT)
- Tomaž Erjavec, Jožef Stefan Institute (SI)
- Eva Hajičová, Charles University Prague (CZ)
- Erhard Hinrichs, University of Tübingen (DE)
- Marinos Ioannides, Cyprus University of Technology (CY)
- Langa Khumalo, North West University (ZA)
- Nicolas Larrousse, Huma-Num (FR)
- Krister Lindén, University of Helsinki (FI)
- Karlheinz Mörth, Austrian Academy of Sciences (AT)
- Costanza Navarretta, University of Copenhagen (DK)
- Jan Odijk, Utrecht University (NL)
- Maciej Piasecki, Wrocław University of Science and Technology (PL)
- Stelios Piperidis, Institute for Language and Speech Processing (ILSP), Athena Research Center (GR)
- Eirikur Rögnvaldsson, University of Iceland (IS)
- Kiril Simov, IICT, Bulgarian Academy of Sciences (BG)
- Inguna Skadiņa, University of Latvia (LV)
- Koenraad De Smedt, University of Bergen (NO)
- Marko Tadič , University of Zagreb (HR)
- Jurgita Vaičenonienė, Vytautas Magnus University (LT)
- Tamás Váradi, Research Institute for Linguistics, Hungarian Academy of Sciences (HU)
- Kadri Vider, University of Tartu (EE)
- Martin Wynne, University of Oxford (UK)

Reviewers:

- Lars Borin, SE
- António Branco, PT
- Tomaž Erjavec, SI
- Eva Hajičová, CZ
- Martin Hennelly, ZA
- Erhard Hinrichs, DE
- Marinos Ioannides, CY
- Nicolas Larrousse, FR
- Krister Lindén, FI
- Monica Monachini, IT
- Karlheinz Mörth, AT
- Costanza Navarretta, DK
- Jan Odijk, NL
- Stelios Piperidis, GR
- Eirikur Rögnvaldsson, IS
- Kiril Simov, BG
- Inguna Skadiņa, LV
- Koenraad De Smedt, NO
- Marko Tadić, HR
- Jurgita Vaičenonienė, LT
- Tamás Váradi, HU
- Kadri Vider, EE
- Martin Wynne, UK

Subreviewers:

- Federico Boschetti, IT
- Christophe Parisse, FR
- Thorsten Trippel, DE
- Valeria Quochi, IT
- Zijian Győző Yang, HU
- Efstathia Soroli, FR
- Enikő Héja, HU
- Bence Nyéki, HU
- Angelo Mario Del Grosso, IT
- Olivier Baude, FR
- Kinga Jelencsik-Mátyus, HU

CLARIN 2021 submissions, review process and acceptance

- Call for abstracts: 19 January 2021, 1 March 2021
- Submission deadline: 28 April 2021
- In total 40 submissions were received and reviewed (three reviews per submission)
- Virtual PC meeting: 16-17 June 2021
- Notifications to authors: 22 June 2021
- 35 accepted submissions

More details on the paper selection procedure and the conference can be found at https://www.clarin.eu/event/2021/clarin-annual-conference-2021-virtual-event.

Table of Contents

How to Perform Linguistic Analysis of Emotions in a Corpus of Vernacular Semiliterate Speech with the Help of CLARIN Tools

Rosalba Nodari
Università di Siena
rosalba.nodari@unisi.it

Luisa Corona
Università dell'Aquila
luisa.corona@univaq.it

Abstract

Research has shown that words are constitutive of emotions and that language contributes to shape feelings. However, less is known about how people with basic literacies can use language to maintain, create and recreate affective bonds, and how they express themselves through the language of emotions. In this respect, digital humanities tools can help shed some light on linguistic encoding of emotions. This proposal aims to show the potential of the CLARIN infrastructure tools for carrying out such analysis on a particular corpus of letters written in the 60s' by Michela Margiotta, a semiliterate Italian woman affected by tarantism, to the anthropologist Annabella Rossi. The research will show how corpora of semiliterate letters can pose several problems when conducting research using digital humanities tools. In this respect, different methodologies will be compared in order to verify how CLARIN tools can help in the detection of encoded emotion in written documents.

1 Introduction

Research in linguistics has shown how speakers can manipulate language in order to manifest feelings and evoke emotions in their listeners (Bednarek 2008). Affect in particular is not confined to the private domain of the subjectivity of the individual, but it is usually expressed and manifested through interaction. Emotion itself can be said to be one among several types of stances (Ochs 1996, Dubois 2009, Du Bois and Kärkkäinen 2012), and speakers can contribute to the co-construction of feelings through the use of specific linguistic cues that can index specific values. Lexemes, syntactic structure and encoding strategies can be perceived, processed and linked to peculiar moods. Therefore, linguistic studies and discourse analysis have focused on how emotions and scopes can be expressed though text structure and specific rhetorical moves that may be considered appropriate for discourse types as related to specific textual genre.

Studies devoted to the language of emotion have been conducted using different strategies developed under the umbrella of the digital humanities tools. In particular, lexicon based, rule based or learning based models have explored the role of textual resources such as neighbour words, word frequency, or terms in contexts, as cues for the expression of emotions (Lindquist, Gendron and Satpute 2016).

Research on corpora of written letters can be considered a fruitful field of inquiry. Historically, letters allowed people to maintain personal connections: relationships were thus constructed in the absence of the subjects themselves, with letters allowing the negotiation of affective bonds (Cancian 2010, 2012; Lyons 2013). Crucially, this communicative need was experienced even by people with basic literacy and little familiarity of written practices, and corpora of vernacular written speech can be associated with members of all classes (for Italy see, for example, Caffarena 2005). Letters written by speakers with basic literacies all share a series of linguistic features that have been analysed by linguists (De Mauro 2015 [19701]; Spitzer 1976). However, little is known about how people with basic literacies can use language to maintain, create and recreate affective bonds, and how they express themselves

through the language of emotions. In addition, analysis undertaken using corpus linguistics and digital humanities tools is still somewhat underrepresented (but see Vitali 2020). Corpora of semiliterate letters pose several problems for conducting research with digital tools. The transcription of semiliterate letters usually reflects the written habits of the writers, thus maintaining, for example, misspelling, orthographical errors, malapropism, vernacular forms. Furthermore, investigations like sentiment analysis usually require adaptation and first-pass cleaning of the data coming from written text, and this can sometimes alter the very nature of written vernacular forms that typically follow oral communication patterns.
The scope of this paper is, then, to propose to the CLARIN community an interesting case study that can be used to test and implement CLARIN resources. In particular, we want to underline how digital tools can help in tagging and analysing Italian written corpora of semiliterate writers, and how a quantitative approach can help improving knowledge regarding how emotion is encoded in written texts. However, given the nature of the specific corpus that we will use as a case study, we aim at highlighting some of the main problems encountered when conducting this specific type of analysis. CLARIN infrastructure offers, in fact, a wide variety of applications to discover, explore, exploit, annotate, analyse or combine language data. Nevertheless, many of these resources are not available for Italian and, particularly, are not even suitable for the analysis of vernacular forms. When dealing with corpora of semiliterate writers, these problems are especially amplified. We believe that this kind of research suggestion, despite all the difficulties that it poses, can be of interest to the CLARIN infrastructure. If CLARIN can help in offering tools that can be modified in order to be suitable for the analysis of semiliterate speech, researchers will benefit them for future applications.

2 The data under scrutiny: the description of Michela Margiotta corpus

In 1959, the anthropologist Annabella Rossi (1933 - 1984) is invited by Ernesto De Martino to join him and his équipe on some ethnographic fieldworks in Salento (Apulia, Italy), to document and research the phenomenon of tarantism. In this occasion Annabella Rossi meets Michela Margiotta (1898 – 1983) during the feast of San Paul in the sanctuary of Galatina (LE). Margiotta was a woman from the nearby village of Ruffano (LE) who was affected by tarantism and the "male di San Donato", the latter being typically used in the Salento area to refer to epileptic seizures or even pseudo-seizures of psychogenic nature. Rossi shows deep empathy for Margiotta's suffering and a curiosity about her story. From this moment onwards, Margiotta spontaneously decides to start a correspondence with Rossi, with the aim to help her in documenting tarantism in Salento, possessions, and other relevant ethnographic facts.
The correspondence between Michela Margiotta and Annabella Rossi spans around 6 years, from 1959 to 1965. In 1970 Rossi decided to collect and publish the letters received from Margiotta in a book entitled *Lettere da una tarantata*, where the woman is anonymised using the pseudonym of Anna.
Following its publication, this volume has been regarded as an important documentation of women's writing in what De Mauro (2015 [1970[1]]) defined as *'italiano popolare unitario*' "unitary popular Italian", a substandard variety of Italian language used by illiterate and semiliterate people. It is also considered a key text in anthropological literature, as it allowed to reflect on the relationship between the observer – the anthropologist – and the observed – the informant, and to highlight the asymmetry of power emerging in field research (see Apolito 2006, 2015).
The epistolary of Michela Margiotta comprises a total of 65 letters: 20 letters were dictated to a more literate person (according to a widespread and well-documented practice among uneducated people), whereas the other 45 were written directly by Margiotta and are listed in the corpus in chronological order. Crucially, despite previous research conducted on several front found no trace of Rossi's replies to any of these letters. Looking at the epistolary it can be assumed that the anthropologist has, at times, replied to Margiotta. Nevertheless, Rossi's role in this epistolary relationship can only be reconstructed from the references to letters, cards and gifts that are contained in Margiotta's letters. A personal investigation undertaken by the legal heirs of Michela Margiotta has confirmed that none of the letters received by Margiotta were ever found in her house after her death.
In addition to the epistolary corpus, we have been able to identify Margiotta in a long interview contained in *Archivio Sonoro della Puglia* (Apulia Sound Archive), recorded and conducted by Annabella Rossi in 1965. In this occasion, Margiotta talks about various topics related to tarantism and about her own personal experience of "sickness" and possession.

Overall, this archival material allows to have a significant amount of documentation about one single person, that can be of interests for many scholars. In particular, this is one peculiar instance of having access to both the written production and the speech of a semiliterate woman. The anthropological material can guide in the interpretation of her peculiar use of the language. The study of Michela Margiotta can therefore help in laying the foundations to develop a protocol for the analysis of emotional speech.
In her preface to *Lettere da una tarantata*, Annabella Rossi identified several recurring themes in Michela Margiotta's letters and grouped them into six main thematic nuclei. Overall, it can be said that, in her letters, Margiotta talked about different topics, but the whole corpus is characterised by a tension between Margiotta and the anthropologist. It can be argued that for Margiotta the letters were a possibility to construct a bond with a woman that, at least in appearance, was sincerely interested in her personal story, such as family tension, her refusal of the marital status, her phobic relationship with the males. Through these letters, Margiotta expresses herself and tries to change an apparent interviewer-interviewee relationship into something more personal, in order to be recognised as an 'individual' rather than just an informant. For this reason, the letters are characterised by an emotional tension. During these 6 years of writing, Margiotta appears to feel more and more that the relationship is a one-way one. Rossi's expectation and Margiotta's goal are deeply different, and Margiotta has to continuously negotiate the absence of Annabella, while at the same time, re-constructing her relationship with Rossi as friends, lovers, or mother and daughter. Margiotta uses the writing practice not only to describe or express her feelings, but also to reach in her letters and through her letters, a shared ground of emotion with Annabella Rossi.

2.1 CLARIN resources and semiliterate speech

The linguistic analysis of Michela Margiotta shall have, as a scope, the detection of linguistic patterns that can be related to the expression of emotion. To do so, researchers have to rely on corpus processing and annotation. However, a number of *caveats* are needed.
Margiotta is, in fact, semiliterate and her written Italian is characterised by misspelling, vernacular forms, and other phenomena commonly attributable to the *italiano popolare*. Since it is precisely in this struggle with a code that Margiotta does not master well that she can express her emotions with, researchers who are interested in her stylistic practice should not normalise her written speech. However, given the above, Margiotta's written practice doesn't appear to be suitable for a series of linguistic analysis.

As a first example, automatic basic sentiment analysis tools, such as the one that can be found in Voyant, are not suitable for this corpus because Margiotta vocabulary cannot be compared to the generic database that typically used when conducting sentiment analysis. In this respect, Margiotta semiliterate corpus exacerbates the problems that are faced when dealing with vernacular forms.

Another problem that must be faced regards the different written solutions that can be found in Margiotta letters. As for other semiliterate writers, Margiotta is not consistent with her lexical solutions. Sometimes she tends to write complex utterances as a monorhematic item (as, for example, *teneriandare* < *te ne eri andare* 'you had to go'), to merge forms (especially articles and prepositions (*laratiolino* < *la ratiolino* 'transistor radio', *couna voce* < *co una voce* 'with a voice') or to split forms (*sono a rivata* < *sono arrivata* 'I arrived'). In this regard, the analysis of frequencies and keywords in contexts is particularly problematic. The use of stop words is, indeed, not sufficient because it tends to eliminate parts of the lexemes. Additionally, the statistical analysis of frequencies cannot deal with different written forms, that can be counted for several times.

Finally, the style of Margiotta's letters is characterised by the use of constructions and strategies which are typical of the oral domain. Part-of-speech tagging and dependency parsing tools often encounter and show some difficulties and inconsistencies when performing NLP on speech transcriptions. Again, since the written style of Margiotta is entirely discursive, we expect to encounter similar issues.

3 Conclusion

Over the past few years, CLARIN infrastructure has helped many researchers in dealing with non-conventional objects that posed unexpected issues in linguistic analysis. For example, corpora of vernacular speech such as the Gra.Fo project (Calamai and Bertinetto 2014) have been a useful testing grounds for automatic speech recognition and segmentation, or for metadata annotation (Calamai and Frontini 2016). Therefore, with the Michela Margiotta case study, we intend to present to the CLARIN community with an example of something that includes issues of both written and vernacular speech corpora. Currently, the usage of digital humanities tools can present limits that often outweigh the pros. In this respect, we offer the Margiotta written digitised corpus as a case study that can help in developing such tools. Michela Margiotta corpus can be seen as an almost perfect opportunity to develop tailored tools for linguistic analysis. Given that Margiotta corpus comprises several documents from a single writer collected over time, it avoids additional complications such as regional variation or speakers' idiosyncrasies, which is then likely to simplify the tools developing process. This case study will therefore allow to develop some basic guidelines for tools tailoring which can then be used to analyse other illiterate corpora of vernacular speakers.

References

Apolito, P. 2006. *Con la voce di un altro. Storia di possessione, di parole e di violenza.* L'Ancora, Napoli.

Apolito, P. 2015. E sono rimasta come lisolo a mezzo a mare. In Rossi, A., *Lettere da una tarantata*. Squilibri, Roma: 13-62.

Bednarek, M. 2008. Analyzing language and emotion. In Bednarek, M. (ed.), *Emotion talk across corpora*. Palgrave Macmillan, Basingstoke: 1-26.

Caffarena, F. 2005. *Lettere dalla Grande Guerra. Scritture del quotidiano, monumenti della memoria, fonti per la storia. Il caso italiano.* Unicopli, Milano.

Calamai, S. and Bertinetto, P. M. 2014. *Le Soffitte Della Voce. Il Progetto Grammo-Foni*. Vecchiarelli Editore, Manziana.

Calamai, S. and Frontini, F. 2016. Not quite your usual kind of resource. Gra.fo and the documentation of Oral Achives in CLARIN. In *Proceedings of the CLARIN Annual Conference 2016*, Aix-en-Provence. https://hal.archives-ouvertes.fr/hal-01395027/document

Cancian, S. 2010. *Families, lovers, and their letters: Italian postwar migration to Canada*. University of Mannitoba Press, Winnipeg.

Cancian, S. 2012. The language of gender in lovers' correspondence, 1946-1949. *Gender and History*, 24: 755-765.

De Mauro, T. 2015 [1970[1]]. Per lo studio dell'italiano popolare unitario. In Rossi, A., *Lettere da una tarantata*. Squilibri, Roma: 63-82.

Du Bois, J. W. 2009. Interior dialogues: The co-voicing of ritual in solitude. In Senft G. and Basso E. B. (eds.), *Ritual communication*. Berg, Oxford: 317-340.

Du Bois, J. W., and Kärkkäinen, E. (2012). Taking a stance on emotion: Affect, sequence, and intersubjectivity in dialogic interaction. *Text & Talk*, 32(4): 433-451.

Lindquist, K. A., Gendron, M. and Satpute, A. B. 2016. Language and emotion: Putting words into feelings and feelings into words. In Barrett, L. F., Lewis, M. and Haviland-Jones, J. M. (eds.), *Handbook of emotions*. 4th edition. Guilford Press, New York: 579-594.

Lyons, M. 2013. *The writing culture of ordinary people in Europe, c. 1860-1920*. Cambridge University Press, New York.

Ochs, E. 1996. Linguistic resources for socializing humanity. In Gumperz, J. and Levinson, S. (eds.), *Rethinking linguistic relativity.* Cambridge University Press, New York: 407-437.

Rossi, A. 2015 [1970[1]]. *Lettere da una tarantata*. Squilibri, Roma.

Spitzer, L. 1976. *Lettere di prigionieri di guerra italiani, 1915-1918*. Bollati Boringhieri, Torino.

Vitali, G. P. 2020. What is a last letter? A linguistic/preventive analysis of prisoner letters from the two World Wars. In Marras, C., Passarotti, M., Franzini, G. and Litta, E. (a cura di), *Atti del IX Convegno Annuale AIUCD. La svolta inevitabile: sfide e prospettive per l'Informatica Umanistica. Quaderni di umanistica digitale:* 265-272.

Dependency Trees in Automatic Inflection of Multi Word Expressions in Polish

Ryszard Tuora
Institute of Computer Science
Polish Academy of Sciences
Warsaw, Poland
ryszardtuora@gmail.com

Łukasz Kobyliński
Institute of Computer Science
Polish Academy of Sciences
Warsaw, Poland
lkobylinski@ipipan.waw.pl

Abstract

Natural language generation for morphologically rich languages can benefit from automatic inflection systems. This paper presents such a system, which can tackle inflection, with particular emphasis on Multi Word Expressions (MWEs). This is done using rules induced automatically from a dependency treebank. The system is evaluated on a dictionary of Polish MWEs. Including such a tool into the CLARIN infrastructure, will be beneficial for processing morphologically rich languages.

1 Introduction

Language generation is mainly achieved by either template filling, or autoregressive neural models. Each approach has its use cases. Neural language generators excel at producing creative text. On the other hand templates are a much simpler and controllable solution, which works especially well for generating short messages. Sometimes these methods can be combined, e.g. in delexicalized neural generation.

As with other problems in the domain of NLP, research has so far been concerned mostly with English. Because English has very little morphological structure, this allows to entirely sidestep word inflection. And so template based approaches worked well, because words could simply be inserted into sentences, as long as they are put in a place that fits their grammatical role. Similarly neural generators did not have to choose between multiple different forms of a lexeme, reducing the output vocabulary severalfold[1], and as a consequence simplifying the problem.

In contrast, Polish has rich morphological structure and free word order. For this reason, it is often the case that there is no one proper place to insert a word into a sentence, even if it is known in advance which grammatical role it is supposed to fill. Instead, the grammatical role is partially reflected in the morphology of the word, and so the word must be inflected for the sentence to be well formed. For instance, the nominative case is often associated with the subject role, whereas accusative case corresponds to the object of the sentence. A template based approach would have to respect these constraints.

2 Single-word inflection vs Multi-word inflection

Automatic inflection systems have been the subject of extensive study in the previous years [2]. The basic case is inflecting a single word. Given a word w (not necessarily identical to its lemma), with a particular morphological profile p – a set of morphological features – and the target set of morphological features t, the goal is to find a form of the same lemma $w*$, which is associated with the morphologically closest profile $p*$ satisfying all features from t. It should be noted that in many cases t will not include values for all the morphological attributes, which means, that remaining features should be left untouched[3]. Such is

[1] As a rough measure of the scale of the problem, the biggest morphological dictionary for Polish – sgjp.pl – has 7.55 different forms for each nominal lexeme, and 26.47 forms for each verbal lexeme.

[2] Cf. the SIGMORPHON shared tasks (Cotterell et al., 2018), (Cotterell et al., 2017) which have dealt with automatic inflection across a broad spectrum of languages.

[3] And this is the sense of proximity, implied in "morphologically closest". It should be observed that sometimes this is not possible, because certain feature combinations are illegal. This is why we use a distance metric, as opposed to a hard constraint of satisfying remaining features.

the case e.g. when it is our goal to obtain a plural form of a word, without altering its grammatical case.

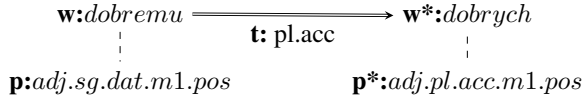


Figure 1. Inflecting "dobremu" into "dobrych", only the number ($sg \rightarrow pl$) and case ($dat \rightarrow gen$) are affected (note that *w* is not the lemma form).

A more extensive treatment of morphology should also allow working with multi-word expressions (MWE), as such phrases are powerful vehicles of meaning, and are often used to refer to named entities, which are of particular importance in some NLG applications. This subject is considerably less explored in literature: the most detailed work for Polish is Savary et al. (2009), where a set of handwritten rules is used to tackle a very particular application.

In the case of MWE, apart from the morphological structure of each word, there is also the internal syntactic structure of the phrase, which is often reflected in morphological features among the components. The internal syntax of the MWE can affect the process of inflection, because of the grammatical relations of governance and agreement. Therefore the phrase must be inflected in its entirety.

Governance is the more difficult relation, as it is often lexical in nature, for example the preposition "z", requires that the noun which is bound by it, be in the genitive case. Fortunately, unless the MWE would have to be constructed from scratch, we can assume that the lexical layer is fixed. Therefore governance is largely irrelevant, and given that the original phrase already satisfies governance constraints, it will also be the case for the inflected one. On the other hand, agreement between two words requires that inflecting one word with respect to one morphological attribute, brings about an identical alteration to the second word. Therefore, when inflecting an MWE, preserving the internal structure does not require an explicit representation of governance, but must take into account the agreement relations. The latter can be done by propagating the morphological features along their corresponding agreement relations.

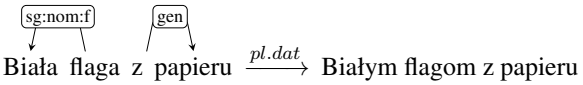


Figure 2. Morphosyntactic relations inside the phrase "Biała flaga z papieru" (*white flag made of paper*). Because "Biała" modifies "flaga" these words must both agree on the number (sg), case (nom), and gender(f), on the other hand "papieru" is introduced by a preposition "z" which lexically requires a noun in the genitive. When inflecting this into plural and dative case, the adjective modifier is modified, as it must still remain in agreement, while the rest of the phrase is left unaltered.

It should be noted, that this paper assumes that inflection does not alter the number of words in the phrase. This assumption does not hold for all cases (e.g. the introduction of auxiliary verbs in future tense), but it is satisfied in nominal and adjective phrases, which form majority of the use cases.

3 Implementation

3.1 Dictionary-based inflection

Inflection of a single word can be done by using an inflection dictionary as follows. First the set of all forms associated with the lemma of the word *w* is recovered, and filtered down to forms which satisfy all the features from target profile *t*. Secondly, the list of these forms is sorted by the size of the symmetric difference defined on morphological profiles. Lastly, the top rated word is returned as the form which satisfies the desired morphological profile, and is the closest to the original form. In this implementation Morfeusz2 (Woliński, 2014) was used for interfacing with the SGJP dictionary.

	m1	others
pos	**dobrych**:4	dobre:6
com	lepszych:6	lepsze:8
sup	najlepszych:6	najlepsze:8

Table 1. The set of all forms belonging to the lexeme for "dobremu": the adjective DOBRY (*good*), which has been narrowed down to forms satisfying the target morphology *t* (*pl.acc*). This leaves two attribute values to still be selected: degree (rows) and gender (columns). The morphologically closest form (in bold) is chosen based on the size of the symmetric difference between profiles (given after the colon), in this case the profile closest to the original *adj.sg.dat.m1.pos* is in *m1* gender and *pos* degree.

3.2 Neural inflection

Alternatively – and this is the common approach nowadays – a supervised learning solution can be used to tackle single-word inflection. For comparison, a BiLSTM seq2seq model was trained on the lexical data from the dictionary mentioned above. The added bonus over a dictionary-based method, is the possibility of generalizing into neologisms and domain vocabulary which is not covered by SGJP. The model architecture is closest to Faruqui et al. (2015), with the exception that a separate decoder for each feature was trained there, whereas here, one model is used for all possible transformations, with one-hot vector representation of target features as an additional input. The model is trained on the SGJP dictionary, in order to discover generalizable patterns.

3.3 Dependency relations as a proxy for agreement relations

In accordance with the arguments presented above, extending these solutions to MWEs, would require an explicit treatment of morphological agreement. In the absence of an existing formalism for dealing with this phenomenon, it is possible to utilize a different grammatical formalism, as an approximation. This is because agreement is not independent from other grammatical relations. The most popular example of accomodation: morphological agreement between an adjective, and a noun which it is modifying, is also one of the simplest structures studied by both dependency, and constituent approaches to sentence structure. Therefore it should be possible, to gain important insight into agreement, by just looking at the grammatical structure of a phrase.

The approach chosen here, was that of dependency grammar in the UD annotation scheme. The goal would be then, to have a system of rules for morphological agreement expressed in terms which match the morphological formalism (here, the National Corpus of Polish tagset (Przepiórkowski et al., 2012)), and the grammatical formalism (namely that of the largest UD treebank for Polish - PDB (Wróblewska, 2018)). Such a rule would have the form $dep \rightarrow attrs$ where *dep* corresponds to the dependency label, and *attr* represents the list of attributes, for which agreement between the head, and dependent occurs. For example adjective modifiers of nouns agree with respect to number, case and gender, and so a rule $amod \rightarrow number.case.gender$ is introduced.

Although morphological agreement is a well-studied phenomenon of the Polish language, we were not able to find any existing rulesets formulated in terms of the chosen formalism. For this reason we have decided to induce such a ruleset in an empirical fashion, i.e. by studying statistical regularities in a corpus containing both morphological, and dependency annotation - in this experiment, we have chosen the PDB treebank. Such a process assumes, that a high enough correlation of attribute values between dependents and heads, of a given dependency relation type, is good grounds for explaining it by reference to morphological agreement. This procedure is clearly far from perfect, as there can be different factors at play. For instance, it is often the case that nouns and their nominal modifiers agree with respect to number, but the reasons behind this are usually semantic. Therefore, the rule induction, should not be taken as a theoretical proposal in the domain of grammar, but rather as a heuristic procedure, which can be used to arrive at useful simplifications. As a result, we obtain a table of frequency of agreement given a dependency label of the relation, and the morphological attribute. From this table, we extract all the agreement rules, by selecting those combinations which exceed a predefined threshold of 95%.

3.4 MWE Inflection algorithm

Given the ruleset, the inflection algorithm of an MWE proceeds recursively, using the dependency tree of the phrase, and morphological tags of all the tokens. First, the root of the expression is selected and inflected, using the single-word algorithm. Subsequently, the values for each altered attribute, are propagated to all the children of the node in the dependency tree. A morphological feature can ”pass” through a dependency relation only if this is enabled by some rule in the ruleset, i.e. its morphological attribute is on the right hand side of the rule corresponding to the label of the dependency relation in question.

3.5 MWE lemmatization

A related use case, pertains to MWE lemmatization. This is an important subject outside of NLG, for example in entity resolution. A trivial solution for MWE lemmatization can be based on concatenating lemmas of individual tokens. Unfortunately this rarely produces satisfying lemmas, as all the information about internal structure of the MWE is lost. The algorithm for MWE inflection can be applied in this case as well. At the beginning, the root of the phrase is lemmatized, and all the morphological alterations are recorded[4]. Subsequently, these alterations are propagated along dependency arcs, just as in the case of MWE inflection. Lemmatization of Polish multi-word phrases has been previously discussed by Marcińczuk (2017) and was one of the tasks proposed for the 2019 edition of the PolEval competition (Ogrodniczuk and Łukasz Kobyliński, 2019).

4 Evaluation

Because the system must be provided with the expression to be processed, along with its full morphosyntactic tags, and labeled dependency tree of the phrase, it naturally fits in a general-purpose NLP pipeline, such as COMBO (Rybak and Wróblewska, 2018). In this experiment, both the dependency parsing, as well as morphological tagging was done by the *polish-herbert-large* model for COMBO.

As of now, there are no public datasets for training, or evaluating MWE inflection systems for Polish. Obtaining natural data for such an application is difficult, as it would require finding natural occurrences of phrases, which differ just by the morphological characteristics of some phrase (and any other alterations which are grammatically necessitated by this). For this reason, we opted for a different way of evaluation.

One resource which might prove useful, are inflectional dictionaries of MWEs, like SEJF (Czerepowicka, Monika and Savary, Agata, 2018). SEJF contains different inflectional variants of MWEs such as noun phrases, proper names and idiomatic adjective phrases. Because each entry is marked with the base form of the MWE, and the morphological profile of the inflected form, one can inflect the base form into the target profile, and compare it against the gold standard. Additionally, the same data can be used in a reverse configuration, for evaluation of MWE lemmatization. Table 2 contains the accuracy measures achieved on this dataset, for both dictionary-based, and neural versions of the MWE processing algorithm.

	dictionary based	neural based
inflection accuracy	90.54	85.21
lemmatization accuracy	79.97	78.96

Table 2. Form accuracy for two methods of inflecting MWEs, as evaluated on 50k examples from SEJF.

5 Conclusions

The results show that the proposed algorithm allows for high accuracy with respect to inflecting MWEs. The system presented here can be extended into other languages, and will be an interesting addition to

[4]This requires a comparison of morphological profiles between the original form, and the lemma, and so we need to know the morphological profile of the lemma. For this reason the lemma has to be processed by the tagger once more, independently of its original context.

the CLARIN infrastructure. This would require inducing rules for the given language, training a neural inflection component, and building a pipeline with a matching tagger and dependency parser. The resulting system could be utilized in most NLG tasks, but also for data generation and augmentation.

References

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2017. CoNLL-SIGMORPHON 2017 shared task: Universal morphological reinflection in 52 languages. In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 1–30, Vancouver, August. Association for Computational Linguistics.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Arya D. McCarthy, Katharina Kann, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. The CoNLL–SIGMORPHON 2018 shared task: Universal morphological reinflection. In *Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, pages 1–27, Brussels, October. Association for Computational Linguistics.

Czerepowicka, Monika and Savary, Agata. 2018. SEJF - A Grammatical Lexicon of Polish Multiword Expressions. In Vetulani, Zygmunt and Mariani, Joseph and Kubis, Marek, editor, *Human Language Technology. Challenges for Computer Science and Linguistics*, pages 59–73, Cham. Springer International Publishing.

Manaal Faruqui, Yulia Tsvetkov, Graham Neubig, and Chris Dyer. 2015. Morphological inflection generation using character sequence to sequence learning. *CoRR*, abs/1512.06110.

Michał Marcińczuk. 2017. Lemmatization of multi-word common noun phrases and named entities in Polish. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 483–491, Varna, Bulgaria, September. INCOMA Ltd.

Maciej Ogrodniczuk and Łukasz Kobyliński, editors. 2019. *Proceedings of the PolEval 2019 Workshop*, Warsaw, Poland. Institute of Computer Science, Polish Academy of Sciences.

Adam Przepiórkowski, Mirosław Bańko, Rafał L. Górski, and Barbara Lewandowska-Tomaszczyk, editors. 2012. *Narodowy Korpus Jezyka Polskiego*. Wydawnictwo Naukowe PWN, Warsaw.

Piotr Rybak and Alina Wróblewska. 2018. Semi-supervised neural system for tagging, parsing and lematization. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 45–54, Brussels, Belgium, October. Association for Computational Linguistics.

Agata Savary, Joanna Rabiega-Wiśniewska, and Marcin Woliński. 2009. Inflection of Polish Multi-Word Proper Names with Morfeusz and Multiflex. In Małgorzata Marciniak and Agnieszka Mykowiecka, editors, *Aspects of Natural Language Processing, Lecture Notes in Computer Science 5070*, pages 111–141. Springer Verlag.

Marcin Woliński. 2014. Morfeusz reloaded. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*, pages 1106–1111, Reykjavík, Iceland. European Language Resources Association (ELRA).

Alina Wróblewska. 2018. Extended and enhanced Polish dependency bank in Universal Dependencies format. In Marie-Catherine de Marneffe, Teresa Lynn, and Sebastian Schuster, editors, *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 173–182. Association for Computational Linguistics.

Corpora for Bilingual Terminology Extraction in Cybersecurity Domain

Andrius Utka
Vytautas Magnus University, CCL
Kaunas, Lithuania
andrius.utka@vdu.lt

Sigita Rackevičienė
Mykolas Romeris University
Vilnius, Lithuania
sigita.rackeviciene@mruni.eu

Liudmila Mockienė
Mykolas Romeris University
Vilnius, Lithuania
liudmila@mruni.eu

Aivaras Rokas
Vytautas Magnus University, CCL
Kaunas, Lithuania
aivaras.rokas@vdu.lt

Marius Laurinaitis
Mykolas Romeris University
Vilnius, Lithuania
laurinaitis@mruni.eu

Agnė Bielinskienė
Vytautas Magnus University
Kaunas, Lithuanian
agne.bielinskiene@vdu.lt

Abstract

The paper aims at presenting English-Lithuanian corpora for bilingual term extraction (BiTE) in the cybersecurity domain within the framework of the project DVITAS. It is argued that a system of parallel, comparable, and training corpora for BiTE is particularly useful for less resourced languages, as it allows to efficiently use strengths and avoid weaknesses of comparable and parallel resources. A special focus is given to the open nature of the data, which is achieved by publishing the data in CLARIN-LT repository.

1 Introduction

The model of combining several types of corpora has been chosen for the bilingual terminology extraction project DVITAS.[1] The aim of the project is to develop a methodology for automatic extraction of English and Lithuanian terms of a specialised domain from parallel and comparable corpora, as well as to create a publicly available bilingual termbase. Cybersecurity (CS) terminology has been chosen as a specialised domain for the project because of its particular relevance in today's digitalised world in which cyber hygiene skills are indispensable for every Internet user. The compiled termbase is believed to be valuable both for specialists of the domain and the general public, as well as drafters of legal and administrative documents, and translators.

The project aims at employing current deep learning terminology extraction methods. In 2020, the project team (Rokas et al., 2020) completed a pilot study on semi-supervised automatic extraction of Lithuanian CS terms from a Lithuanian monolingual corpus. A small-scale manually annotated dataset (66,706 word corpus with 1,258 annotated cybersecurity terms) was used as training data. The pilot study was performed in several stages: firstly, various baseline LSTM and GRU networks were tested using Adam optimizer and FastText embeddings; secondly, each of the best baseline LSTM and GRU networks were tested with various optimizers; and finally, the best model was compared with a model that has been trained using multilingual BERT embeddings (Rokas et al., 2020) . The latter approach proved to be the most efficient: Bidirectional Long Short-Term Memory model (Bi-LSTM) using multilingual Bidirectional Encoder Representations from Transformers (BERT) embeddings reached F1 score of 78.6%.

We believe that more comprehensive training data obtained from larger manually annotated gold standard corpora will allow to improve the obtained results. In studies by other scholars, who applied neural networks for term extraction as sequence labelling task and used larger annotated datasets, higher F1

[1] https://klc.vdu.lt/dvitas/en

score was achieved: e.g., Kucza et al. used a dataset with 78,567 annotated terms and with Bi-LSTM reached F1 score of 86.73% (Kucza et al., 2018).

The methodology used in the pilot study will be modified and tested on different configurations of neural networks taking into account the methods applied in related research. For instance, studies on sequence labeling tasks with multilingual BERT embeddings show that reduction of the number of languages to three in BERT models may help to achieve higher results compared with the ones achieved with multilingual BERT (Ulčar and Robnik-Šikonja, 2020).

2 Related Research

Bilingual/multilingual term extraction, which is widely used for terminographic purposes, is performed using both parallel and comparable corpora. Term extraction from parallel corpora has been already applied for several decades (Kupiec, 1993). Lately, the importance of comparable data is increasing, as more and more papers have appeared on term extraction from comparable corpora (Vintar, 2010; Delpech et al., 2012; Gornostay et al., 2012; Aker et al., 2013; Chu et al., 2016). Besides, since 2008 the *Workshop on Building and Using Comparable Corpora (BUCC)* has published a number of valuable research papers on the usage of comparable corpora for term extraction.

Researchers indicate several important advantages of using comparable data. Firstly, term extraction from comparable corpora provides valuable terminological data as these data reflect the usage of terminology in original languages which is much more natural than the usage of terminology in translations that are inevitably influenced by source languages. There is a strong possibility of having "inconsistencies in parallel corpora, which are then replicated by translators" (Postolea and Ghivirigă, 2016). Another important advantage of this approach is the possibility to include data sources of a much larger variety as data source search is not limited to translated resources. The third advantage is that comparable corpora are less expensive to build than parallel corpora. The last two are especially important for less-resourced languages which often lack parallel data.

Therefore, some scholars have introduced the idea of combining comparable and parallel corpora to benefit from the advantages provided by both (Bernardini, 2011; Morin and Prochasson, 2011; Biel, 2016; Giampieri, 2018) or yet some researchers concentrate solely on comparable corpora (Steyaert and Rigouts Terryn, 2019; Vintar et al., 2020).

3 Corpora System for Bilingual Terminology Extraction

Five CS corpora have been compiled for this project: a parallel corpus of English texts and their Lithuanian translations (approx. 1.4 million words), a comparable corpus composed of two subcorpora: original English texts and original Lithuanian texts (approx. 4 million words), and three training (gold standard) corpora (approx. 0.1 million words each), which will be manually annotated. The system of corpora and a flowchart of BiTE is presented in Figure 1.

The analysis of the cybersecurity sources revealed that this domain is highly heterogeneous and it encompasses diverse types of information accumulated in various discourses. Ideally, the cybersecurity corpora should be representative of the whole cybersecurity domain and its constituent genres. In order to fully represent the CS domain, we need to consider four discourses as sources of information: legislative & administrative, academic, expert, and popular (the discourses identified by (Wall, 2007)).

Most sources are suitable for compilation of the comparable corpus, which will consist of the original texts in English and Lithuanian. Meanwhile, the sources suitable for the parallel corpus (English original texts and their translations into Lithuanian) are much more sparse.

Legislative and executive sources contain textual information on cybersecurity, such as cybersecurity strategies, laws, government resolutions, minister orders, etc. Official national and EU legally binding and non-binding documents are commonly accessible without any restrictions. The documents of these categories can be acquired for both comparable and parallel corpora (see Table 1 and Table 2).

Another important source of relevant information is scientific research publications on the cybersecurity topic. However, access to academic and scientific publications is often restricted. Most relevant scientific sources are published by major publishing companies and protected by intellectual property

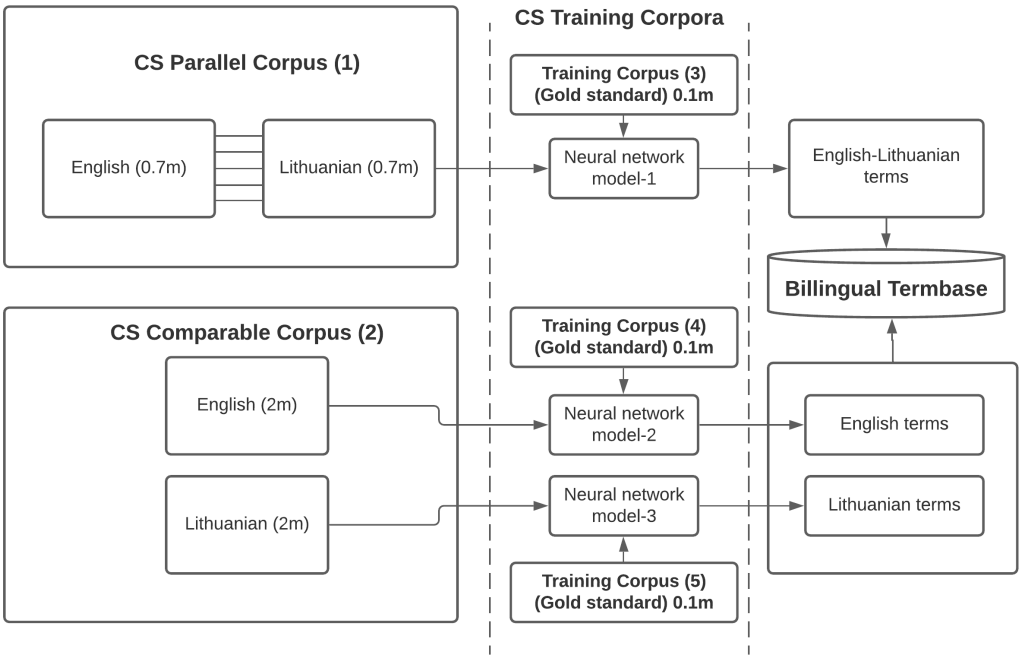


Figure 1: Corpora system for BiTE

rights. As we should ensure proper usage of these texts, the amount of texts suitable for the comparable corpus is rather limited. In fact, it is almost impossible to acquire original and translated academic texts for the parallel corpus.

We had to rely on the inclusion of rather large bulk of media texts into the corpus, due to less-restricted accessibility of these texts. In order to avoid an influx of general terms into the corpus, we tried to include more specialised media sources.

The CS comparable corpus compiled for the project includes texts from the time period of 2010-2020. The categories, subcategories and their approximate proportions within the corpus are presented in Table 1.

Main categories	**Subcategories**	**Proportion**
Legislative and executive documents	CS strategies, laws, government resolutions, minister orders	10%
Official non-binding documents and informational texts	Reports and recommendations of the National Cybersecurity Centres; booklets and posters	15%
Academic texts	Scientific articles, monographs, MA and PhD theses, textbooks	25%
Media texts	Mass media articles, specialised media articles	50%

Table 1: Structure of the comparable corpus (2010-2020).

The parallel corpus includes the EU legal acts and other documents from the time period of 2010-2020. The documents are extracted from the EUR-Lex database and other EU institutional repositories (see Table 2). As mentioned previously, for the parallel corpus it is almost impossible to acquire original

and translated academic texts and it is likewise difficult to find translated media articles. Therefore the corpus relies solely on EU documents.

Main categories	**Subcategories**	**Proportion**
Legally binding documents (secondary legislation)	Regulations of the European Parliament and of the Council; Directives of the European Parliament and of the Council; Decisions of the European Parliament and of the Council	60%
Official non-binding documents	Communications of the European Commission; Reports of the European Commission; Recommendations of the European Commission; Opinions of the Committees of the EU; Briefing papers of the Court of Auditors	40%

Table 2: Structure of the parallel corpus composed of the EU documents (2010-2020).

Training (gold standard) corpora have been composed of the same text categories as the main corpora. The comparable training corpus encompasses legally biding documents, official non-binding documents and informational texts, academic texts, and media articles. Parallel training corpus is composed of the most important EU legal acts and other documents on cybersecurity issues.

It is important to note that the task of depositing all 5 corpora into CLARIN-LT repository as open access resources is included in deliverables of the project.

4 Concluding Remarks

In the full paper we will present a more lengthy discussion on related research and state-of-the-art approaches to BiTE. Moreover, we will present a detailed discussion on different intellectual property restrictions and how we have dealt with them when compiling parallel and comparable corpora. Finally, the full paper will provide links to the compiled corpora in the CLARIN-LT repository.

Acknowledgements

The research is carried out under the project "Bilingual Automatic Terminology Extraction" funded by the Research Council of Lithuania (LMTLT, agreement No. P-MIP-20-282). The project is also included as a use case in COST action "European Network for Web-Centred Linguistic Data Science" (CA18209).

References

Ahmet Aker, Monica Lestari Paramita, and Robert J. Gaizauskas. 2013. Extracting bilingual terminologies from comparable corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers*, pages 402–411. The Association for Computer Linguistics.

Silvia Bernardini. 2011. Monolingual comparable corpora and parallel corpora in the search for features of translated language. *SYNAPS - A Journal of Professional Communication*, 26.

Łucja Biel. 2016. Mixed corpus design for researching the eurolect: A genre-based comparable-parallel corpus in the pl eurolect project. *Polskojęzyczne korpusy równoległe. Polish-language Parallel Corpora*.

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2016. Parallel sentence extraction from comparable corpora with neural network features. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*. European Language Resources Association (ELRA).

Estelle Delpech, Béatrice Daille, Emmanuel Morin, and Claire Lemaire. 2012. Extraction of domain-specific bilingual lexicon from comparable corpora: Compositional translation and ranking. In Martin Kay and Christian Boitet, editors, *COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, 8-15 December 2012, Mumbai, India*, pages 745–762. Indian Institute of Technology Bombay.

Patrizia Giampieri. 2018. Online parallel and comparable corpora for legal translations. *Altre Modernità*, 20:237–252.

Tatiana Gornostay, Anita Ramm, Ulrich Heid, Emmanuel Morin, Rima Harastani, and Emmanuel Planas. 2012. Terminology extraction from comparable corpora for latvian. In Arvi Tavast, Kadri Muischnek, and Mare Koit, editors, *Human Language Technologies - The Baltic Perspective - Proceedings of the Fifth International Conference Baltic HLT 2012, Tartu, Estonia, 4-5 October 2012*, volume 247 of *Frontiers in Artificial Intelligence and Applications*, pages 66–73. IOS Press.

Maren Kucza, Jan Niehues, Thomas Zenkel, Alex Waibel, and Sebastian Stüker. 2018. Term extraction via neural sequence labeling a comparative evaluation of strategies using recurrent neural networks. In B. Yegnanarayana, editor, *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018*, pages 2072–2076. ISCA.

Julian Kupiec. 1993. An algorithm for finding noun phrase correspondences in bilingual corpora.

Emmanuel Morin and Emmanuel Prochasson. 2011. Bilingual lexicon extraction from comparable corpora enhanced with parallel corpora. In *4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, pages 27–34.

Sorina Postolea and Teodora Ghivirigă. 2016. Using small parallel corpora to develop collocation-centred activities in specialized translation classes. *Linguaculture*, 2016(2):53–72.

Aivaras Rokas, Sigita Rackevičienė, and Andrius Utka. 2020. Automatic extraction of lithuanian cybersecurity terms using deep learning approaches. In Andrius Utka, Jurgita Vaičenonienė, Jolanta Kovalevskaitė, and Danguolė Kalinauskaitė, editors, *Human language technologies - the Baltic perspective: proceedings of the 9th international conference, Baltic HLT, Kaunas, Vytautas Magnus University, Lithuania, 22-23 September 2020*, pages 39–46. IOS Press.

Kim Steyaert and Ayla Rigouts Terryn. 2019. Multilingual term extraction from comparable corpora: Informativeness of monolingual term extraction features. In Serge Sharoff, Pierre Zweigenbaum, and Reinhard Rapp, editors, *Proceedings of the 12th Workshop on Building and Using Comparable Corpora*, pages 16–25, Varna, Bulgaria, September.

Matej Ulčar and Marko Robnik-Šikonja. 2020. Finest bert and crosloengual bert: less is more in multilingual models. *arXiv e-prints*, June.

Špela Vintar, Larisa Grčić Simeunović, Matej Martinc, Senja Pollak, and Uroš Stepišnik. 2020. Mining semantic relations from comparable corpora through intersections of word embeddings. In *Proceedings of the 13th Workshop on Building and Using Comparable Corpora*, pages 29–34, Marseille, France, May. European Language Resources Association.

Spela Vintar. 2010. Bilingual term recognition revisited. *Terminology*, 16:141–158.

David S. Wall. 2007. *Cybercrime: The Transformation of Crime in the Information Age*. Polity.

Voices from Ravensbrück. Towards the creation of an oral and multilingual resource family

Silvia Calamai
Università di Siena
Siena, Italy
silvia.calamai@unisi.it

Jeannine Beeken
University of Essex
Colchester, United Kingdom
jeannine.beeken@essex.ac.uk

Max Broekhuizen
Erasmus School of History
Rotterdam, The Netherlands
maksbroekhuizen@gmail.com

Christoph Draxler
Ludwig Maximilian University
Munich, Germany
draxler@phonetik.uni-muenchen.de

Arjan van Hessen
University of Twente
Enschede, The Netherlands
a.j.vanhessen@utwente.nl

Henk van den Heuvel
Radboud University
Nijmegen, The Netherlands
h.vandenheuvel@let.ru.nl

Stefania Scagliola
University of Luxembourg
Belval, Luxembourg
scagliolas@gmail.com

Abstract

This paper describes a pilot project aimed at introducing a new type of corpus in the CLARIN resource family tree, called 'narratives'. To this end, a multilingual corpus of existing interviews with survivors of concentration camp Ravensbrück will be curated following CLARIN compliant standards. During WWII this German camp imprisoned 130.000 women from 20 different nationalities. This diversity creates the opportunity to build a unique corpus of gender specific interviews, covering the same topic, narrated in a similar structure, but voiced in different languages. The corpus will also be enriched with various types of annotation (transcripts e.g.).

1 CLARIN Resource families and oral history

The CLARIN Resource Family is a user-friendly overview per data type of available language resources in the CLARIN infrastructure aimed at the needs of researchers from (digital) humanities and social sciences and human language technologies.[1] Within this overview, there is only one entry for 'spoken corpora', which contains 90 data sets mainly targeted at phonetic, linguistic and speech technology research. We argue for a new type of entry, namely the datatype 'narratives', covering oral history interviews and other types of spoken narrative discourse, in both audio and textual form. Interviews, aside from oral history, are a central object of research in a broad variety of fields such as anthropology, psychology, literary studies, sociology, health studies, education, linguistics and cognitive science. Yet there seem to be scarce opportunities for comparison and cross-fertilisation between these disciplines.

It is our contention that oral history interviews represent an under-utilized but promising datatype outside the realm of history, and that the CLARIN infrastructure is the ideal 'home' for this type of 'family' to illustrate its multidisciplinary potential. With this project we want to substantiate this position, by bringing together and comparing similar interview data – retrospective spoken narratives of women – from different countries and languages – about the same topic: going through war and trauma in concentration camp Ravensbrück. First the characteristics of oral history and opportunities for digital humanities will be discussed. This is followed by a sketch of the data that is currently available and data

[1] https://www.clarin.eu/resource-families

in other languages that will be explored. We conclude with a description of the envisioned workflow for collecting, enriching and publishing the corpus.

2 The underutilized potential of Oral History

The key characteristic of oral history is that it is co-created and mediated through the use of language, speech and memory. Its content is thus *multi-layered*. It can also be appreciated in *multimodal* ways (i.e., seeing the interview, hearing the recording, reading the transcript). An oral history interview offers information through a single person, but from the perspective of interactions in a social context (e.g., the family, the village, the military unit, the company). These descriptions reveal the creation of identity vis-à-vis social, economic, political and cultural contexts.

Aside from what is said, one can also reflect on the organization of the narrative: what is not said, what is repeated again and again, the language and terms that are used. Another dimension to study is the interaction that takes place during the interview: the mediating process between interviewer and interviewee who together are co-creating an historical source. Typical questions relate to differences in gender, background, status and the relation between standard and vernacular speech in the course of the interview.

With regard to modalities, the encounter can be studied: (1) by listening to the audio-signal, (2) by identifying speaker-specific physiological information with digital tools (segmental and suprasegmental acoustics, silence, emotions, speech rate and timing), and (3) by reading the textual representation of what is uttered, which is characterized by the transcription convention. Regardless of the approach, the interview has value as singular source, as well as in the form of a collection (similarities and differences).

With regard to ethical and legal issues and re-use of data, oral history data has a different background than spoken corpora that are specifically created to study language. This is evident with regard to recordings made before the digital era, when giving consent for use of the source meant that it could be consulted at the premises of an archive under supervision of someone who can be held accountable. But even for projects initiated in the digital era, that can be accessed through the web in secured ways, the problem is encountered that consulting the data is a different research practice then processing the data.

The first is a one-to-one encounter: you listen to a signal (the audio of an interview) or read one transcript at a time, while 'processing' the data, for example to detect signals or patterns, entails an automated process: you need access to the file itself to run it through the software. In a way this is taking Alessandro Portelli's adage to 'bring back orality to oral history' one step further. He pledged for shifting the focus of research from the transcript to the auditory features of the source: the tone and rhythm of a voice (Portelli 1981).

We strive to go back to the signal itself and discover patterns in speech and non-verbal features. Technically this does not pose problems, but with regard to privacy, copyright and access control, we will have to enter a 'trading zone' with the 'guardians' of oral history data and design special measures to guarantee that the interviews will be treated in a respectful way (Calamai *et al.* 2019). Once such measures will be taken up, they will secure access to a 'bonanza' of data that is awaiting to be understood in 'new ways'.

3 The historical context of Ravensbrück and its suitability for our objective

Such 'bonanza' of data in terms of richness of voices, speakers, narrative styles and topical coverage, requires a clear structure for orientation. The biggest challenge is cutting down the abundance of variables that become visible when joining data from different contexts of creation, to a set of parameters that make a comparison meaningful within a particular paradigm. That too many variables within one theme, is not productive for such a comparison, is what we learned during a CLARIN workshop in Munich in 2018 (*Oral History: Users and their scholarly practices in a Multidisciplinary world*, Munich 2018). The idea was to bring together interviews on the topic of migration in different languages as a basis for experimenting with annotation, analysis, and emotion recognition tools. There was however too much diversity: different discourses on migration and different tools to apply on this data. The focus of the workshop was therefore put on breaking down 'silos' of knowledge, identifying the obstacles for the uptake of these tools in different disciplines that work with interview data.

With the donation of the archive of the Italian scholar Anna Maria Bruzzone, who interviewed five survivors of Ravensbrück for her book, to the University of Siena in 2016, a new opportunity for cross-

disciplinary multilingual research presented itself. Interviews about experiences in one camp have a more specific set of variables. Moreover the spoken memories of these women have been collected extensively in many of the 20 countries to which the women returned at the end of the war. This makes comparisons across languages and diverse historical contexts more viable.

The camp, built in 1939 and located in northern Germany, 90 km north of Berlin, was initially intended for social outcasts or so-called 'inferior beings' that had to be re-educated (Romani people, political dissenters, criminals and prostitutes). As the German occupation expanded to new territories, new types of female prisoners with either a Jewish background or who had been involved in rescue or illegal operations, were deported to the camp. This evolution explains the widely divergent background of the prisoners. Of the 130,000 women from 20 different nationalities that passed through it, 48,500 came from Poland, 28,000 from the Soviet Union, 24,000 from Germany and Austria, 8,000 from France, and thousands from other countries. More than 20,000 women among this population were Jewish, and 80 percent were political prisoners. Many of these prisoners were employed as forced laborers by Siemens & Halske. From 1942 to 1945, medical experiments were undertaken as well.

What all narratives about being imprisoned in Ravensbrück have in common is the gender perspective. What makes the diversity of the narratives interesting is the socio-cultural context in which the retrospective account is expressed. The story of a former Polish prisoner who immigrated to the USA and tells her story for the Shoah Visual Archive, can be quite different from someone with exactly the same background at the time, who returned to Poland after the liberation of the camp and contributes to a Polish interview project.

4 Data that is currently available

Web research and consultation of a number of authors has yielded enough data and commitment to be able to search for profiles that match those of the five narrators in Bruzzone's archive. We intend to first complement the Italian interviews with English, Dutch and German ones. In the long run we intend to expand the project to Eastern Europe, especially to Poland and Russia, where many survivors came from.

Bruzzone's Ravensbrück interviews consist of 14 audio cassettes, with a total duration of about 18 hours and 20 minutes. The analogue audio cassettes were digitized according to IASA standards (.wav format, 96000 Hz, 24 bit). The archive contains 4 long interviews. We know that for her publication, Anna Maria Bruzzone transcribed the recordings step by step, writing everything down that she heard, but unfortunately, the handwritten transcriptions were lost. [2] In 2016, the book was translated to German.

For Dutch, we have access to interviews held with Dutch Ravensbrück internees between 2007 and 2010, for a PhD study on the memory of Ravensbrück by Susan Hogervorst (2010). In case this material may pose legal problems, due to the absence of consent from narrators who have deceased,[3] we can fall back on the many project-generated interview collections that are publicly available in archives. These interviews have been created in the wake of the 50th anniversary of the second World War from the 1990s onwards, and their proliferation has been strongly influenced by the availability of digital technology and a push towards presence on the web. In the Netherlands this has resulted in the online resource *Getuigenverhalen*[4] which contains 3 interviews on Ravensbrück. The Visual History Archive contains 5 interviews in Dutch about Ravensbrück. The United States Holocaust Memorial Museum contains 1 interview in Dutch about Ravensbrück.

With regard to interviews in German, we have access to the *Videoarchiv "Die Frauen von Ravensbrück"* (200 interviews), *Österreichische Lagergemeinschaft Ravensbrück und Freundinnen* (34 interviews). In English, we have the *Visual History Archive* (20 interviews) and the *Imperial War*

[2] Hopefully, with digital repositories this will never happen again.

[3] Legal issues may occur at different levels in case of oral archives, due to the transposition of the General Data Protection Regulation (GDPR) into different national laws. In Italian law, just as an example, GDPR applied also to dead people. At a general level, one may consider the right of the interviewers, that of the interviewees, and that of third parties mentioned during an interview. In some national laws, interviews are also protected by the laws on authors' rights.

[4] http://getuigenverhalen.nl/home

Museum GB (8 interviews). We are confident that in some of the countries involved in the CLARIN network, additional similar material might be found, in further languages.

5 Towards a CLARIN Resource Family for Oral History

The planned project consists of two phases: 1. basic curation of available data and exploration of data for expansion, 2. curation and enrichment of all available data. In the first phase, five Italian interviews from the Bruzzone collection will be described and transcribed via the Transcription Chain in the Oral History portal, a project supported by Clarin[5]. We will prepare the hosting of this corpus in a CLARIN Centre and generate the metadata according to an appropriate CMDI profile for oral history. This means that data created within a historical framework will be described in a format and structure that adheres to CLARIN methodology. To search in other oral history collections for interviews on Ravensbrück in English, German and Dutch, that match with the Bruzzone archive, we will use a profile of the five Italian narrators and take into consideration their background, the length of the interviews and the interview approach (chronological semi-structured interviews with probing for details). We will use the archival material identified in section 4 as a backlog.

In the expansion phase we will curate and enrich the collected data by adding transcriptions, time stamps at word level, and phonetic and suprasegmental information. We will also add annotations on e.g. the use of specific language, emotions expressed in non-verbal modes of communication (i.e., laugh, pauses and silences, breathing). Aside from exploring the content, we will also document the feasibility of the re-use of oral history data, within the framework of FAIR open data, taking into account the particularity of the interview as source of knowledge, and identifying technological and legal obstacles for the re-use and dissemination of such material. Overcoming these obstacles will hopefully increase the visibility of this type of data and foster the interest for cross-disciplinary and multilingual approaches to research with interviews.

References

Beccaria Rolfi, L., Bruzzone, A.M. 2020. *Le donne di Ravensbrück. Testimonianze di deportate politiche italiane*, Torino, Einaudi.

Beccaria Rolfi, L., Bruzzone, A.M. 2016. *Als Italienerin in Ravensbrück. Politische Gefangene berichten über ihre Deportation und ihre Haft im Frauen-Konzentrationslager*, Herausgegeben von Johanna Kootz, Metropol Verlag Berlin.

Calamai, S., Kolletzek C., Kelli, A. 2019. *Towards a protocol for the curation and dissemination of vulnerable people archives*. In: Skadina, I. & Eskevich, M.: Selected papers from the CLARIN Annual Conference 2018, Pisa, 8-10 October 2018. Linköping University Electronic Press, Linköpings universitet: 28-38 https://ep.liu.se/ecp/159/003/ecp18159003.pdf.

Draxler, C., Van den Heuvel, H., Van Hessen, A., Calamai, S., Corti, L., and Scagliola, S. 2020. A CLARIN *Transcription Portal for Interview Data. Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC2020)*. pp. 3346-3352 http://www.lrec-conf.org/proceedings/lrec2020/pdf/2020.lrec-1.411.pdf

Hogervorst, S. *Onwrikbare Herinnering. Herinneringsculturen van Ravensbrück in Europa*. 2010. Uitgeverij Verloren, Hilversum.

Portelli, C. 1981. On the peculiarities of oral history. *History Workshop Journal*, 12: 96-107. https://doi.org/10.1093/hwj/12.1.96

Scagliola, S., Corti, L., Calamai, S., Karrouche, N., Beeken, J., Van Hessen, A., Draxler, Chr., Van den Heuvel, H., Broekhuizen, M., and Truong, K.. 2020. *Cross disciplinary overtures with interview data: Integrating digital practices and tools in the scholarly workflow.* In: Simov, K., & Eskevich, M.: Selected Papers from the CLARIN Annual Conference 2019 Leipzig, 30 September - 2 October 2019. Linköping Electronic Conference Proceedings 172:15, 126-136. https://doi.org/10.3384/ecp2020172.

[5] https://www.phonetik.uni-muenchen.de/apps/oh-portal/

ParlaMint: Comparable Corpora of European Parliamentary Data

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

Maciej Ogrodniczuk
Institute of Computer Science PAS, Poland
maciej.ogrodniczuk@gmail.com

Petya Osenova
IICT-BAS, Bulgaria

Andrej Pančur
Institute of Contemporary History, Slovenia

Nikola Ljubešić
Jožef Stefan Institute, Slovenia

Tommaso Agnoloni
CNR-IGSG, Italy

Starkaður Barkarson
Árni Magnússon Institute for Icelandic Studies

María Calzada Pérez
Universitat Jaume I, Spain

Çağrı Çöltekin
University of Tübingen, Germany

Matthew Coole
Lancaster University, the UK

Roberts Darģis
IMCS UL, Latvia

Luciana D. de Macedo
Univ. Federal de Minas Gerais, Brazil

Jesse de Does
Dutch Language Institute,
the Netherlands

Katrien Depuydt
Dutch Language Institute,
the Netherlands

Sascha Diwersy
Univ. Paul Valéry Montpellier 3, France

Dorte Haltrup Hansen
University of Copenhagen, Denmark

Matyáš Kopp
Charles University, the Czech Republic

Tomas Krilavičius
Vytautas Magnus University, Lithuania

Giancarlo Luxardo
Univ. Paul Valéry Montpellier 3,
France

Maarten Marx
Universiteit van Amsterdam,
the Netherlands

Vaidas Morkevičius
Kaunas University of Technology, Lithuania

Costanza Navarretta
University of Copenhagen, Denmark

Paul Rayson
Lancaster University, the UK

Orsolya Ring
Centre for Social Sciences, Hungary

Michał Rudolf
Institute for Computer Science PAS, Poland

Kiril Simov
IICT-BAS, Bulgaria

Steinþór Steingrímsson
Árni Magnússon Institute for Icelandic Studies

István Üveges
University of Szeged, Hungary

Ruben van Heusden
Universiteit van Amsterdam, the Netherlands

Giulia Venturi
CNR-ILC, Italy

Abstract

This paper outlines the ParlaMint project from the perspective of its goals, tasks, participants, results and applications potential. The project produced language corpora from the sessions of the national parliaments of 17 countries, almost half a billion words in total. The corpora are split into COVID-related subcorpora (from November 2019) and reference corpora (to October 2019). The corpora are uniformly encoded according to the ParlaMint schema with the same Universal Dependencies linguistic annotations. Samples of the corpora and conversion scripts are available from the project's GitHub repository. The complete corpora are openly available via the CLARIN.SI repository[1] for download, and through the NoSketch Engine[2] and KonText[3] concordancers as well as through the Parlameter[4] interface for exploration and analysis.

1 Introduction

ParlaMint[5] (July 2020 – May 2021) was a project that built on the achievements of the ParlaCLARIN community and methodology and was financially supported by CLARIN-ERIC. The mission of ParlaMint was to turn existing contemporary diverse cross-national parliamentary data into resources that are comparable, interpretable and highly communicative with respect to society (NGOs, citizens, researchers, etc.). The ParlaMint project started with the creation of recent corpora of parliamentary sessions for 4 parliaments: Bulgarian, Croatian, Polish and Slovene. The project was then extended with 13 additional parliamentary corpora of the following countries: Belgium, the Czech Republic, Denmark, France, Hungary, Iceland, Italy, Latvia, Lithuania, the Netherlands, Turkey, and the UK. In addition, Spanish parliament data were added on a voluntary basis.

The project aimed to provide data and tools for focused observations on trends, opinions, decisions on lock-downs and restrictive measures as well as on the consequences with respect to health, medical care systems, employment, etc. in times of emergencies. For the ParlaMint project the emergency case is obvious – the COVID-19 pandemic. However, the methodology is scalable also to other events, such as economic crises, etc. Thus, the main aims of the project were: to compile a collection of parliamentary corpora from a number of countries and in a number of languages in a harmonized format, covering both current data and older, reference data; to process the corpora linguistically; to index the data with popular concordancers so that interested parties can search and extract the relevant comparable information; to make the data, workflow descriptions, related standards and lessons learnt publicly available; to show through appropriate use cases that the CLARIN resources and technology serve societal needs.

Considerable effort was already put into data from European Parliament, so we have at disposal valuable and well-synchronized resources like Europarl (Koehn, 2005),[6] JRC-Acquis (Steinberger et al., 2006)[7] or DCEP: Digital Corpus of the European Parliament (Hajlaoui et al., 2014).[8]

At the same time, there are many ongoing national initiatives ranging from parliament-focused corpora to task-oriented ones. Within large EU initiatives, such as CLARIN-ERIC, identification was performed of the available resources within European countries. It is worth mentioning that parliamentary data were one of the CLARIN Key Resource Families (Fišer et al., 2018).[9]

A number of related workshops have also been organized on the topics of gathering, standardizing, processing, maintaining, visualizing and using parliamentary data, in particular: CLARIN-PLUS Work-

[1] https://www.clarin.si/repository/xmlui/handle/11356/1432 and https://www.clarin.si/repository/xmlui/handle/11356/1431
[2] http://www.clarin.si/noske/
[3] https://www.clarin.si/kontext/corpora/corplist
[4] https://parlamint.parlameter.org/poslanske-skupine
[5] https://www.clarin.eu/content/parlamint
[6] http://www.statmt.org/europarl/
[7] https://ec.europa.eu/jrc/en/language-technologies/jrc-acquis
[8] https://ec.europa.eu/jrc/en/language-technologies/dcep
[9] https://www.clarin.eu/resource-families/parliamentary-corpora

shop "Working with Parliamentary Records"[10] (2017); two ParlaCLARIN workshops at LREC 2018 and 2020 (Fišer et al., 2018; Fišer et al., 2020)[11] or CLARIN Interoperability Committee ParlaFormat workshop[12] (2019) on standardization of parliamentary data.

Parliamentary data have also been subject of growing interest of the digital humanities reflected in search for synergies with the natural language processing community. This resulted in such events as *Computational Analysis of Political Texts* tutorial[13] offered at the top venues of computational social science and natural language processing (IC2S2 2019[14] and ACL 2019[15]) or *Big Data and the Study of Language and Culture: Parliamentary Discourse across Time and Space* workshop[16].

The paper is organized as follows: in the next section the structure and availability of the ParlaMint corpora is outlined. Section 3 briefly showcases the participating languages and parliaments. Section 4 concludes the paper.

2 Structure and availability of the corpora

ParlaMint contains 17 corpora with 16 languages (the Belgian corpus is bilingual Dutch/French), and comprises 22 thousand files, over 3.5 million speeches and almost 500 million words. It defines over 11 thousand persons and over 1.5 thousand "organisations", i.e. political parties, parliamentary groups etc.

In Figure 1 we give an overview of the ParlaMint corpora. The left side gives the time period covered by each corpus, with the dashed line (November 2019) splitting the period into "reference" and COVID-19 subcorpora. The middle part gives the country code, and the right part shows the number of words contained in the corpora. As can be seen, most corpora start in 2015, with the earliest speeches from 2009, and, while most corpora end mid-2020, the latest extends to April 2021. As for sizes, by far the largest corpus, both per year and in total, is that of the UK, with even the fact that it contains the speeches of both the House of Lords and of the House of Commons not fully explaining its size, but must be (as it is with the French) a result of longer or more sessions of their parliaments. In the opposite direction, the outlier is the Hungarian corpus, where its small size is due to the fact that it contains only interpellations and urgent questions from plenary sessions of the parliament.

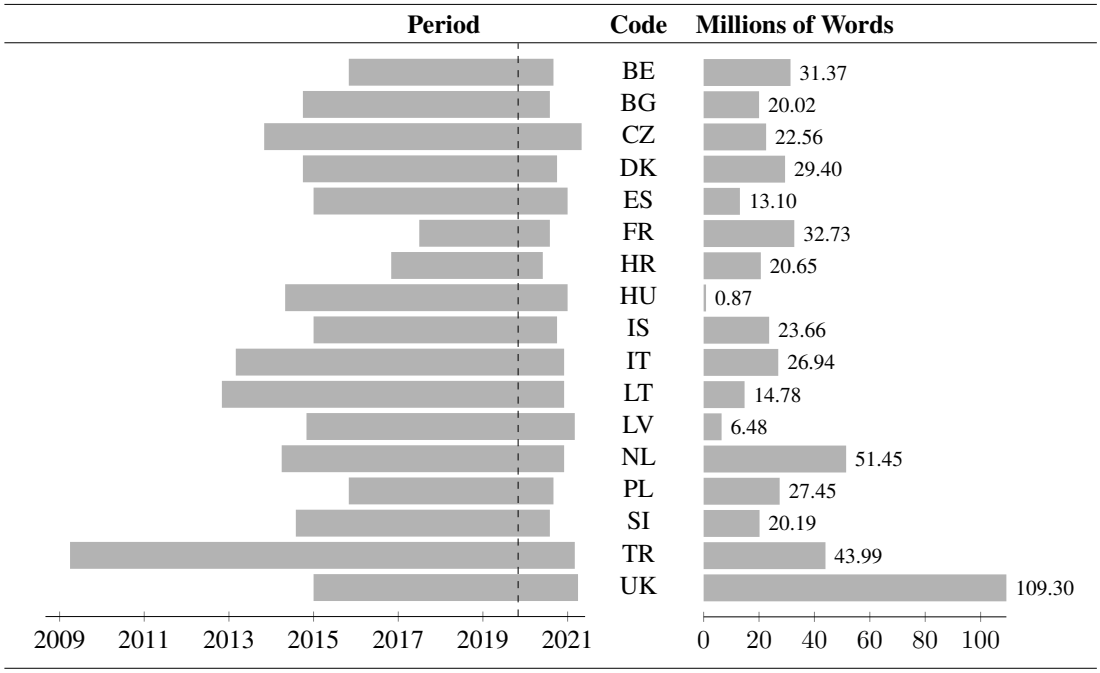
The corpora have extensive metadata about the speakers (speaker name, gender, party affiliation, MP status). They are structured into time-stamped terms, sessions and meetings, with each speech being marked by its speaker and their role (chair, regular speaker). The speeches contain also marked-up transcriber comments, such as gaps in the transcription, interruptions, applause, etc.

The corpora are encoded according to the Parla-CLARIN TEI recommendation[17] but have been validated to conform to the much stricter ParlaMint schemas, available from the ParlaMint GitHub repository.[18] This repository includes, apart from the XML schemas, also content validation scripts, scripts to convert the corpora into other formats, as well as samples from all the available corpora.

The corpora are available under CC BY via the CLARIN.SI repository in two variants, the "plain text" (Erjavec et al., 2021b) and the linguistically annotated one (Erjavec et al., 2021a). The former includes all the metadata and structured transcription in XML and in derived plain text format, while the latter adds linguistic annotations, which include named entities, lemmatisation, morphological features and syntactic parses according to the Universal Dependencies recommendations.[19] This version also includes the corpora in derived CoNLL-U and so called vertical formats. Samples of the "plain text" and linguistically annotated corpora, as well as the samples in several derived formats are also available from the ParlaMint GitHub repository.

[10] https://www.clarin.eu/event/2017/clarin-plus-workshop-working-parliamentary-records
[11] https://www.clarin.eu/ParlaCLARIN, https://www.clarin.eu/ParlaCLARIN-II
[12] https://www.clarin.eu/event/2019/parlaformat-workshop
[13] https://poltexttutorial.wordpress.com/
[14] 5th International Conference on Computational Social Science, https://2019.ic2s2.org/
[15] 57th Annual Meeting of the Association for Computational Linguistics, https://acl2019.org/
[16] Collocated with 40th Intl. Computer Archive of Modern and Medieval English conference, http://icame.uib.no/
[17] https://clarin-eric.github.io/parla-clarin
[18] https://github.com/clarin-eric/ParlaMint
[19] https://universaldependencies.org

Figure 1: The time period and number of words of the ParlaMint corpora.



3 Compilation of the ParlaMint corpora

The corpora of the following countries are included in ParlaMint: Belgium, Bulgaria, Croatia, the Czech Republic, Denmark, France, Hungary, Iceland, Italy, Latvia, Lithuania, Poland, Spain, the Netherlands, Slovenia, Turkey and the UK.

First of all, these countries have different political and thus, parliamentary systems. For example, there are unicameral (Bulgaria, Croatia, Denmark, Hungary, Iceland, Latvia, Lithuania, Turkey) and bicameral parliaments (Belgium, the Czech Republic, France, Italy, Poland, Slovenia, Spain, the Netherlands, the UK), each with its own specifics, which is reflected in the structure of the particular corpora, e.g. whether they distinguish sessions, sittings, and meetings. The steps of getting the data, converting them to the ParlaMint schema and annotating it linguistically also varied across the corpora.

Getting the data required either scraping it from the parliamentary websites (Belgium, Bulgaria, the Czech Republic, Hungary, Iceland, Latvia, Spain, Turkey); obtaining via Parlameter API (Croatia); retrieving from an already maintained parliamentary corpus (Poland and Slovenia); downloading from a server (Denmark, France, the Netherlands); obtaining through parliamentary API (UK) or through a service center at the parliament (Italy).

Data conversion employed various strategies such as: incremental and semi-automatic transformation from HTML to basic TEI XML and then to the ParlaMint format through XML constraints (Bulgarian) or through XSLT stylesheets and Python, Perl and Bash scripts (Belgian, Dutch, French, Spanish); automatic conversion through Perl scripts with heuristics only for difficult parts such as the transcriber comments (Croatian, Czech, Danish); automatic conversion through Python scripts with possible corrections of data during the process (Hungarian, Icelandic, Latvian, Polish, Turkish); transformation with XSLT, and some manual interventions upstream (Slovene) or adding necessary extensions to XSLT (English); automatic conversion with JAVA code (Italian). The main challenges of the conversion were related to re-structuring the data, and esp. adding mark-up to the previously unstructured data.

Linguistic processing included the UD-based morphosyntactic annotation and a named entity annotation with the traditional NEs: Person, Location, Organization and Misc.

This step was also approached differently by the groups depending on the availability of these tools for the language and their quality and performance. Thus, for some languages pre-trained pipelines were used that follow the same model. For example, the CLASSLA pipeline[20] was used for the annotation of Bulgarian, Croatian, and Slovene corpora. Italian, French and Spanish relied on the Stanza NLP pipeline, while for English the Stanford NLP pipeline was used. In the Spanish case, the Stanza NLP pipeline was aided by AnCora Treebanks and corpora.[21]

Other languages used language-specific models either different for each step, or in a combined piped mode, which was the case for Belgian, Czech, Danish, Dutch, Hungarian, Icelandic, Latvian, and Polish.

Some corpora contain additional linguistic information, e.g. Croatian and Slovene have also the MULTEXT-East (Erjavec, 2012) morphosyntactic annotations, while Czech also contains their own highly detailed and nested NE annotations.

4 Conclusions

The ParlaMint project establishes an innovative strategy for handling parliamentary data and processing them in times of any emergency period (COVID-19 is just a showcase). The novelties relate to unified handling of cross-lingual and cross-parliament comparable data, and to the quick access of all interested parties to these data.

The project output was already used in several studies. ParlaMint took part in the Helsinki Digital Humanities Hackathon DHH21 (19–28.05.2021).[22] The corpora were explored in three practical showcases: on Science and Expertise in Parliaments[23] on a comparative analysis of the available corpora[24] and on the Parlameter service of the ParlaMint project.[25]

The Parla-CLARIN TEI encoding is becoming a de-facto standard for national parliamentary data, and it will be further developed to cover more detailed and specific metadata across languages and parliaments. The created openly available corpora can serve as a baseline for further updates. Such uniform updates across the corpora would strongly support various methods of comparative research across parliaments and political systems.

We believe that the availability of comparable multilingual parliamentary data will boost further the research in the areas of digital humanities, linguistics, politology, sociology, psychology as well as in all the related branches of sciences.

Acknowledgements

We would like to thank CLARIN-ERIC for the financial support of ParlaMint.

The work on Bulgarian Parliamentary data was partially supported by the Bulgarian National Interdisciplinary Research e-Infrastructure for Resources and Technologies in favor of the Bulgarian Language and Cultural Heritage, part of the EU infrastructures CLARIN and DARIAH – CLaDA-BG, Grant number DO1-377/18.12.2020.

The work on the Czech Parliamentary data was partially supported by the Ministry of Education, Youth and Sports of the Czech Republic, Project No. LM2018101 LINDAT/CLARIAH-CZ.

The work on the Danish ParlaMint corpus was partially supported by the Department of Nordic Studies and Linguistics at the University of Copenhagen through CLARIN-DK.

The work on Hungarian Parliamentary data was partially supported by the Ministry of Innovation and Technology NRDI Office within the framework of the Artificial Intelligence National Laboratory Program, No. NKFIH-870-8/2020; and received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 951832 (OPTED).

[20] `https://pypi.org/project/classla/`
[21] `https://universaldependencies.org/treebanks/es_ancora/index.html`
[22] `https://dhhackathon.wordpress.com/2021/05/28/parliamentary-debates-in-the-covid-times/`
[23] See the video: `https://www.youtube.com/watch?v=K4y03qr4WoU`
[24] See the video: `https://www.youtube.com/watch?v=ddBHvbuzke4`
[25] See the video: `https://www.youtube.com/watch?v=h1292E_vtO8`

The work on Polish parliamentary data was partially supported by project CESAR (CEntral and South-east europeAn Resources, a European CIP ICT-PSP project, grant agreement 271022), CLARIN-PL (a Polish Ministry of Science and Education project, grant numbers DIR/WK/2016/02 and DIR/WK/2018/01) and MARCELL (Multilingual Resources for CEF.AT in the legal domain, a CEF-TC-2017-3 – eTranslation grant, grant agreement INEA/CEF/ICT/A2017/1565710, co-financed by the Polish Ministry of Science and Higher Education: research project 4082/CEF/2018/2, funds for 2018–2020).

The work on the Spanish Parliamentary corpus was supported by the Spanish Ministry of Science and Innovation, PID2019-108866RB-I0 / AEI /10.13039/501100011033, "Original, translated and interpreted representations of the refugee cris(e)s: methodological triangulation within corpus-based discourse studies".

The work on the Latvian Parliamentary data was partially supported by the CLARIN-LV, European Regional Development Fund project "University of Latvia and institutes in the European Research Area – Excellency, activity, mobility, capacity" (1.1.1.5/18/I/016) and Latvian State Research Programme's project "Digital Resources for Humanities: Integration and Developmen" (VPP-IZM-DH-2020/1-0001).

The work on the Slovenian Parliamentary data was partially supported by the Research infrastructures CLARIN.SI and DARIAH-SI, and the Slovenian Research Agency research programme P2-103 "Knowledge Technologies".

We thank Mindaugas Petkevičius, Monika Briedienė and Andrius Utka for their help in creating the Lithuanian corpora.

We thank Bart Jongejan who contributed to the creation of the Danish corpus.

References

Tomaž Erjavec et al. 2021a. *Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint.ana 2.1*. Slovenian language resource repository CLARIN.SI. `http://hdl.handle.net/11356/1431`.

Tomaž Erjavec et al. 2021b. *Multilingual comparable corpora of parliamentary debates ParlaMint 2.1*. Slovenian language resource repository CLARIN.SI. `http://hdl.handle.net/11356/1432`.

Tomaž Erjavec. 2012. MULTEXT-East: Morphosyntactic Resources for Central and Eastern European Languages. *Language Resources and Evaluation*, 46(1):131–142.

Darja Fišer, Jakob Lenardič, and Tomaž Erjavec. 2018. CLARIN's Key Resource Families. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Darja Fišer, Maria Eskevich, and Franciska de Jong, editors. 2020. *Proceedings of the Second ParlaCLARIN Workshop*, Marseille, France. European Language Resources Association (ELRA).

Darja Fišer, Maria Eskevich, and Franciska de Jong, editors. 2018. *Proceedings of LREC 2018 Workshop ParlaCLARIN: Creating and Using Parliamentary Corpora*, Paris, France. European Language Resources Association (ELRA).

Najeh Hajlaoui, David Kolovratnik, Jaakko Väyrynen, Ralf Steinberger, and Daniel Varga. 2014. DCEP – Digital Corpus of the European Parliament. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the Tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiș, and Dániel Varga. 2006. The JRC-Acquis: A Multilingual Aligned Parallel Corpus with 20+ Languages. *CoRR*, abs/cs/0609058.

The nature of Icelandic as a second language: An insight from the learner error corpus for Icelandic

Isidora Glišić
University of Iceland
Reykjavik, Iceland
isg14@hi.is

Anton Karl Ingason
University of Iceland
Reykjavik, Iceland
antoni@hi.is

Abstract

The Icelandic L2 Error Corpus is an expanding collection of texts written by users of Icelandic as a second language, published on CLARIN. It currently consists of 17508 manually-annotated errors in different categories pertaining to grammar, spelling, lexical and other issues. The corpus was used to perform a contrastive interlanguage analysis using a native speaker reference corpus comparing it to the Icelandic Error Corpus. This paper presents the corpus and the first results of the analysis.

1 Introduction

The popularity of Icelandic as a learner language is a quite novel phenomenon and teaching materials are still in development. With new language technology efforts in Iceland, it is finally possible to create ICALL (Intelligent Computer-Assisted Language Learning) solutions for Icelandic and a major step towards this is creating a learner error corpus. At the moment of writing, the Icelandic L2 Error Corpus is a collection of 70 texts, predominantly student essays, annotated for various types of errors. The corpus contains a total of 12081 revision spans and 17508 error instances, where a revision span is a word or a span of words that have been corrected in the annotation process and an error instance is a link between a revision span and a categorization of an error found in the span. This corpus is still growing and will be utilized in analysing learners' interlanguage for the purpose of perfecting teaching materials (both electronic, textbooks and syllabi) and automatic correction tools.

The paper is structured as follows. Overview of previous research on learner interlanguage for Icelandic and the introduction to the new Icelandic L2 corpus is in section 2. Section 3 describes the methods that we used. Section 4 presents the results of a comparative analysis with the Icelandic general error corpus.

2 Resources for studying L2 errors in Icelandic

Until relatively recently, not many foreigners were interested in learning Icelandic and no textbooks or teaching methodology existed. The first contrastive analysis of the learner language emerged in the 1980s (Sigmundsson, 1987), and it was not until the '90s that the first textbooks started being published and even more recently that attention was drawn to learner errors (Þorvaldsdóttir and Garðarsdóttir, 2013; Ólafsson, 2016). Finally the learner corpus has been published (Ingason et al., 2021) and is at this moment still in development. The corpus in its current form is provided in open access via the CLARIN repository and GitHub.

The Icelandic L2 Error Corpus[1] currently consists of 70 texts from 27 adult second language speakers of Icelandic (mostly aged between 20–40) with 13 different first languages, containing 17508 categorized error instances. Further analysis of the corpus data will follow in section 4. The texts are previously unpublished and obtained directly from their authors, who choose whether the text is to be published

[1]The corpus is available at: https://repository.clarin.is/repository/xmlui/handle/20.500.12537/106

under their name or anonymously. The texts are for the most part student essays submitted for evaluation in various courses at the university. The mean number of words per text is 1780 but this number varies when separated by skill level (with the mean for A1 texts being 324 words and 5177 words for C2) as the nature and type of texts vary – the highest skill level texts typically being long academic essays and parts of or entire MA theses. More numerical data based on skill level is depicted in Table 1.

Level	Files	Total words	Total errors	Errors/1000w
A1	12	3889	1095	281.56
A2	19	11901	2532	212.76
B1	8	10439	1919	183.83
B2	11	19504	3367	172.63
C1	9	21940	2715	123.75
C2	11	56953	2911	103.79

Table 1: Total number of files, words, errors and errors per 1000 words per skill level.

The advantage of using student essays is the accessibility of texts from subjects with different first and second language background. Furthermore, it is also relatively easy to estimate their proficiency level based on how far along they are in their studies. On the other hand, due to the nature of the texts (academic essays) some types of errors tend to be more prominent than in other types of writing. Apart from that, many generic errors may have been corrected already as the texts tend to be polished for better academic success. Additionally, as participation is strictly voluntary, it has been challenging to obtain sufficient amount of texts to be able to draw any conclusions based on their demographic features.

The standard for assessing the stage of learners' interlanguage is the Common European Framework of Reference for Languages (CEFR) (Piccardo et al., 2018).[2] The scale is particularly important in evaluating learner errors, as specific types of errors typically emerge on specific proficiency levels, with typical stagnation and regression points (Thewissen, 2013). The Icelandic as a second language program is separated into a one-year Practical diploma in Icelandic which covers the proficiency level A1–A2 and a 3-year bachelor degree where the students are estimated to be on the level B1–B2 by the end of the first year, and reach B2—C1 by the end of the program (Garðarsdóttir and Þorvaldsdóttir, 2020).

How the corpus was built and the process of extracting and analyzing relevant data will be explained in the next section.

3 Methods

The texts for the Icelandic L2 Error Corpus were collected through an online consent form and manually proofread and annotated for errors. Microsoft Word's track changes feature was used for this because it preserves both the original version and the corrected version. Both versions of the text were extracted and converted, using a Python script, into a single augmented TEI format XML document with labeled enumerated sentences, words and punctuation, and revision spans with unique id numbers containing errors. The errors were analysed and annotated manually and the annotators would label one or several error codes in each revision span. Figure 1 shows an example of a revision span containing several error codes. It demonstrates that a revision span can have both multiple codes for different errors, as well as codes which apply to the same error in which case they share the same index (*idx*). So in this example, the errors with the id "255-1" refer to the first two words in the revision span, whereas the last word needs to be covered by a different error type.

The annotation system used for error labeling was originally developed for the Icelandic Error Corpus (Ingason et al., 2020) previously released through CLARIN, which contains errors in native speaker texts,

[2]For more details about proficiency level assessment scale, see: `https://www.coe.int/en/web/common-european-framework-reference-languages/table-1-cefr-3.3-common-reference-levels-global-scale`

and later expanded with new labels that were specific to the L2 errors. The error tagset consists of 19 categories which are further divided into subclasses. Some subclasses are very narrow while others are more wide-ranging (notably the Grammar and Punctuation category) and in total there are 259 error codes. A list of all the codes along with an example and a description is available at `https://github.com/antonkarl/iceErrorCorpusSpecialized/blob/master/IEC_ErrorCodes.pdf`.

After the dataset of TEI documents has been finalized, statistical analyses are conducted that include quantifying the number of texts, revision spans and error occurrences in the corpus, as well as contrasting the L2 error corpus with the Icelandic Error Corpus by ranking the frequency of the error codes extracted as the number of errors per 1000 words. Moreover, each document contains metadata including the author's first language, other languages, length of residence in Iceland, length of study of Icelandic, and proficiency level. This data is stored to extract specific information on errors based on these parameters that can be used in further research.

```
5990    <w>gera</w>
5991    <w>ráð</w>
5992    <w>fyrir</w>
5993    <revision id="255">
5994      <original><w>tvö</w><w>myndbrigði</w><w>fyrir</w></original>
5995      <corrected><w>tveim</w><w>myndbrigðum</w><w>í</w></corrected>
5996        <errors>
5997          <error xtype="case-collocation" idx="255-1" eid="0" />
5998          <error xtype="nominal-inflection" idx="255-1" eid="0" />
5999          <error xtype="wrong-prep" idx="255-2" eid="0" />
6000        </errors>
6001    </revision>
6002    <w>þágufallsendingu</w>
```

Figure 1: An example of revision spans with multiple error codes.

4 Data analysis

The method we used is contrastive interlanguage analysis (CIA) which compares varieties within one language using two types of comparison: comparing learner language with native speaker reference corpora (L2 vs. L1) or comparing different varieties of learner language (L2 vs. L2) (Granger, 2008). The former can uncover the distinguishing features of L2 language use while the latter allows us to assess the generalizability of interlanguage features across different factors, learner and task based. As an error corpus for L1 Icelandic has recently been published (Ingason et al., 2020), this provides us with the possibility to make a CIA based on the first type mentioned.

Corpus	Files	Total words	Revisions	Categorized Errors	Errors/1000w
Icelandic Error Corpus	4046	1137941	44261	55346	44.56
Icelandic L2 Error Corpus	70	124626	12081	17508	140.73

Table 2: Numerical data for both L1 and L2 Icelandic error corpora.

As Table 2 demonstrates, the number of errors per 1000 words is significantly higher in L2 texts than in the general corpus. This is not surprising as learner errors are quite frequent, and particularly on lower proficiency levels the text can be so convoluted and inaccurate that making revisions proved to be a challenge as sometimes entire sentences needed to be rewritten for the text to be semantically coherent. However, it must be noted that the learner error corpus contains significantly fewer and less genre-diverse texts and this may affect how our findings generalize to L2 users as a population.

The analysis also sheds light on a significant disparity in the frequency of certain error categories and subcategories in L2 Icelandic compared to L1 errors. The most frequent error categories in the L2 corpus are: grammar (43.57%), punctuation (12.14%) and wording (11.63%). Each other error category comprises 5% or less of total errors. In comparison, the category grammar accounts for only 11.8% in the general Icelandic Error Corpus. There are 35 error codes that appear only in the L2 corpus, 27 of which are unsurprisingly within the grammar category.

These errors mostly involve case government (*case-verb*, *case-collocation*, *case-prep*, *case-adj*) as it is not intuitive in the language learning process which case is governed by a certain preposition or verb. For example, L2 users will commonly misuse a phrasal verb and instead of interpreting it as a verb+particle combination they would take the particle to be a preposition as part of a prepositional phrase with the following noun and apply the case that preposition governs (so [*leysa af*] + accusative becomes *leysa* +[*af* + dative]). Such is the case also in Figure 1 as the preposition *fyrir* can govern either accusative or dative, but in this case the collocation *gera ráð fyrir* (e. *allow for, anticipate*) exclusively takes dative.

Other typical errors involve the use of grammatical voice, and inflectional errors in closed word classes. Inflectional errors in nouns or verbs are among the most common errors as well but they are also prominent in the L1 corpus. Fixed word order in Icelandic is not intuitive for the learner either which created two additional error subclasses within syntax. Another very specific error type for L2 in the lexical category is *context* – an incorrect word chosen for the specific context often prompted by a literal dictionary translation.

The frequency of error codes was ranked to identify to which extent subclasses differ in frequency between the corpora. If the error code does not appear in a corpus, the rank is by default higher by one than the total number of ranks. The relative rank (Δ rank) between the corpora was calculated for each error code. A high number indicates a large difference in ranks between corpora for an error code, and a low number indicates similar rankings.

Error Codes	Category	Rank L1	Rank L2	Δ rank
wording	wording	1	1	0
nonword	nonword	3	3	0
date-period	punctuation	99	99	0
extra-conjunction	punctuation	32	33	1
comma4colon	punctuation	89	90	1

Table 3: Error codes with most similar rankings between the corpora

Table 3 shows the error codes that have the most similar frequency rankings in the L1 and L2 corpora. Two types of errors that are ranked among the highest in both corpora are *wording* and *nonword* error, the latter being possibly a simple typing error or an attempt to write a word form that does not exist, whereas the former is the most general error type which includes any type of formulating a phrase or a clause in a wrong way, and is often combined with other error types.

5 Conclusion

This paper introduces the Icelandic L2 Error Corpus, the first learner error corpus for Icelandic, which is a collection of texts written by users of Icelandic as a second language. The majority of the texts are essays submitted by students in the Icelandic as a second language program at the University of Iceland. The texts have been manually annotated for errors based on an error tagset previously built for the general Icelandic Error Corpus. Both corpora are openly accessible via CLARIN repository. First CIA results are also presented, comparing the L2 corpus with the general corpus.

At this point, the corpus consists of 12081 revision spans and 17508 categorized error instances. The preliminary results show a large disparity in the quantitative distribution of errors in the Icelandic L2 Error Corpus and the general Icelandic Error Corpus. This disparity relates to both the occurrence of different error categories, where grammar related errors are 4 times more prominent in the L2 corpus,

and the total error rate, which is 3 times as high for the L2 corpus compared to the native speaker referent. Moreover, it is still more than twice as high when the L2 speakers have reached the highest proficiency level.

Further analysis of the data is underway and we are in the process of compiling preliminary results of the L2 vs. L2 CIA. As the corpus is still small and the distribution of features such as first language is not as wide, the focus will be on the proficiency level and length of residence, which tend to intertwine. We hope to expand this corpus to provide further possibilities to analyse various features of learner language that are difficult to highlight due to the limited size of the sample. With the expansion of the corpus, it has potential to become an important asset for learning Icelandic.

References

María Garðarsdóttir and Sigríður Þorvaldsdóttir. 2020. A processability approach to the development of case in L2 Icelandic. *Language, Interaction and Acquisition A cross-theoretical and cross-linguistic perspective on the L2 acquisition of case systems*, 11(1):68–98.

Sylviane Granger, 2008. *Learner Corpora in Foreign Language Education*, pages 1427–1441. 01.

Anton Karl Ingason, Lilja Björk Stefánsdóttir, and Þórunn Arnardóttir. 2020. Icelandic error corpus (IceEC) version 0.9. CLARIN-IS.

Anton Karl Ingason, Lilja Björk Stefánsdóttir, Þórunn Arnardóttir, Xindan Xu, and Isidora Glišić. 2021. The Icelandic L2 error corpus (IceL2EC) version 1.1. CLARIN-IS.

Enrica Piccardo, Tom Goodier, and Brian North. 2018. *Council of Europe (2018). Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume with New Descriptors.* Strasbourg: Council of Europe Publishing., 01.

Svavar Sigmundsson. 1987. Íslenska í samanburði við önnur mál. *Íslenskt mál og almenn málfræði*, 9.

Jennifer Thewissen. 2013. Capturing L2 accuracy developmental patterns: Insights from an error-tagged EFL learner corpus. *The Modern Language Journal*, 97(S1):77–101.

Gísli Hvanndal Ólafsson. 2016. Grammar and linguistic structures at level A1 of Icelandic. Master's thesis, University of Iceland, Unpublished, 6.

Sigríður Þorvaldsdóttir and María Garðarsdóttir. 2013. Fallatileinkun í íslensku sem öðru máli. *Milli mála: Tímarit um erlend tungumál og menningu*, 5:45–73.

Insights on a Swedish Covid-19 corpus

Dimitrios Kokkinakis
SpråkbankenText, Department of Swedish
University of Gothenburg
Gothenburg, Sweden
dimitrios.kokkinakis@gu.se

Abstract

The COVID-19 pandemic has had a serious impact on people all over the world, from mental and physical health to economic downturn to education and social relationships, while political decisions in many countries have had a profound impact on the lives of all people regardless of age. Many of these effects can be studied with statistical and qualitative data such as collected questionnaires and sickness absence rates. But large-scale studies require expertise in multiple domains and from many points of view. *SpråkbankenText* continuously collects text from various sources. In order to fill the gap in the lack of an available Swedish COVID-19-related dataset, we started to build a Swedish COVID-19 corpus (sv-COVID-19). Various tools for e.g. lexical, semantic or pragmatic/discourse analyses can be then applied in order to answer relevant questions on e.g. how people, on a larger scale than what can be obtained through qualitative studies, experienced their everyday life through the different phases of COVID-19 crisis, or how political decisions and their consequences are described and discussed.

1 Introduction

Since the beginning of 2020 the COVID-19 pandemic has had an impact on almost all kind of human activities. Many of the effects of this global crisis can be studied with statistical and qualitative data such as collected questionnaires (SOM, 2020) and sickness absence rates. But large-scale studies require expertise in multiple domains and from many points of view (Tan, et al. 2020). For instance, figures on unemployment or depression can be supplemented and enriched with online discussions and studied through the collection and analysis of text from various media. Language usage and communication practices and style are no exceptions to this reality. Language evolves constantly in order to accommodate for this global situation, giving rise to new words and phrases which spread almost as fast as the virus. A collection of a range of texts from this period and their NLP analysis would allow for the detailed investigations of e.g. the lexical/language use (i.e. the use of compound words, idioms, metaphoric expressions and style), the semantic exploration (i.e. attitudes to the elderly or vaccination and other emerging and rapidly fading micro-events) or various pragmatic/discourse analyses (i.e. effects of the pandemic on the communication in various institutional settings) in order to answer relevant questions on e.g. how people, on a larger scale than what can be obtained through qualitative studies, experienced their everyday life through the different phases of COVID-19 crisis. The language use changes in tandem with changes on the pandemic scene and in order to fill this gap in the Swedish arena, we have collected and continue to collect a Swedish COVID-19-related corpus of various types of text that can be used to support language-related research. In the following section we present a description of the corpus and a few examples of the many ways the corpus can be used, here for lexical exploration.

2 A Current Description of the Corpus

The current version of the corpus we describe consists of ca 5 000 articles, starting from the time when the first media articles about the new virus started to appear in the Swedish media landscape, roughly

during the beginning of the second week of January 2020. The data comes from various Swedish sites and the corpus collection is not static. The aim is to enhance and enrich the corpus with regular updates as well as with Swedish data from Twitter and blogs. Moreover, the content is in plain text, UTF-8, format, cleansed from boilerplate and duplicate content. Due to licensing and privacy restrictions for some of its content, we cannot provide a direct download. However, we provide access by making the corpus searchable where the order of individual sentences is presented in a randomized order. The latest updated version of the corpus is available from the SpråkbankenText's corpus search interface, Korp (Borin, Forsberg & Roxendal, 2012): https://spraakbanken.gu.se/korp/#?corpus=sv-covid-19.

For each document/article in the sv-COVID-19 corpus certain basic types of metadata information are provided, information which can also be exploited in linguistic studies (Schäfer & Bildbauer, 2013). These include the: *title*, *URL source*, *publication date*, in the normalized format YYYY-MM-DD, and, if applicable and available, each article's *authorship* information, that is the author of each document/article. Moreover, we have added a domain/genre label to each article according to the following list of labels: *PBLMD* ('PuBLic MeDia', such as the 'Swedish Radio, SR'); *NEWS* (several Swedish online newspapers); *MDCL* ('MeDiCaL', scientific content from e.g. the 'J of the Swedish Medical Association'); *ATHRTS* ('AuTHoRiTieS', official authority sites content from the e.g.'Public Health Agency of Sweden'); *RSRCH* ('ReSeaRCH', such as research reports from hospitals); *PRDCLS* ('PeRioDiCaLS', periodical literature and magazines); *BLOG* and *CHAT* (content from official and unofficial blog and chat sites and services). The first corona-related articles in the corpus is from the 6th of Jan. 2021; on average there are 175-380 articles per month with an average of 250k words per month.

3 The coronavirus and its vocabulary

The pandemic vocabulary, words and phrases used by experts and non-professionals to describe the crisis, has become frequent and familiar, used and overused in various, sometimes partly overlapping discourse contexts (cf. de Smedt, 2021). On February 11, 2020, the World Health Organization (WHO[1]) announced an official name for the disease that is causing the 2019 novel coronavirus outbreak. The official new name of this disease became 'coronavirus disease 2019', abbreviated as COVID-19. Formerly, this disease was referred to in various ways, e.g. as 'new/novel coronavirus' (*nya coronaviruset*), '2019-nCoV', 'nCoV-19', 'CV19', 'C-19' or 'Cov*d-19'. Sometime occurrences were also available in stigmatizing and discriminatory terms, such as the 'China virus' *Kinavirus* or the Wuhan virus' *Wuhanvirus*; two terms frequently used at the beginning of the pandemic. Note that both 'coronavirus' and 'COVID-19', and their ambiguous variants such as 'corona', are used interchangeably in the various media; coronavirus is a common type of virus officially named "severe acute respiratory syndrome coronavirus 2; SARS-CoV-2; while COVID-19 is the disease caused by SARS-CoV-2.

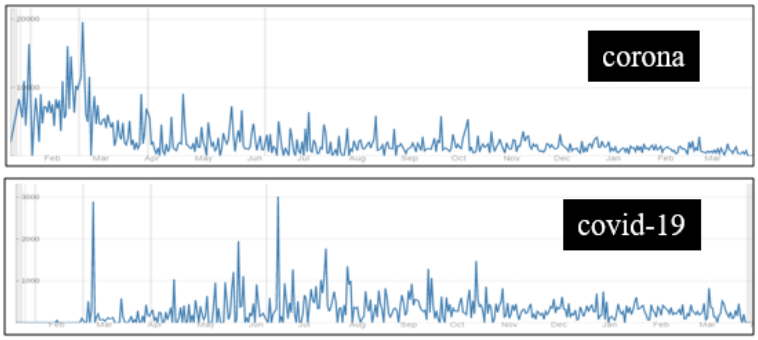


Figure 1. Trend line for the raw frequencies of 'coronavirus' and 'COVID-19' in the Sv-Covid-19.

Figure 1 illustrates two examples of the raw frequencies of 'COVID-19' and 'coronavirus' ('corona') in the Sv-COVID-19 corpus which are (currently) the most frequent content words with over 4500 occurrences for 'coronavirus' and over 7200 occurrences for 'COVID-19'. Note that for the purpose of the graph in figure 1 we do not make any semantic distinction of the two terms, in the sense that many

[1] Similarly on May 31, 2021 WHO announced non-stigmatising labels for SARS-CoV-2 variants using letters from the Greek alphabet, e.g. *delta* for the unfounded Indian variant, *gamma* for the Brazilian one etc.

people use both names interchangeably. Here, it can be clearly seen that 'corona' was overrepresented during the start of the pandemic, at least until the middle of Feb. 2020, when WHO announced the official name of the disease 'COVID-19'.

3.1 Insights into the sv-COVID-19

We summarize here two of the very many ways the corpus can be analysed, explored and presented and we provide some concrete examples to illustrate how this can be achieved by taking a more lexical-oriented perspective described. Also, since at the time of the writing, the COVID-19 situation is fast changing in a complex, and sometimes unpredictable, manner, so the data presented should be seen as a snapshot and results should be interpreted with caution; (cf Schweinberger et al., 2021).

Topic Modelling: a topic model is a type of statistical model for discovering the latent semantic structures or "topics" (clusters of similar words) that occur in a collection of documents. The sv-COVID-19 corpus contains documents of roughly similar lengths, with very few exceptions, that makes it suitable to explore using unsupervised topic modelling, without a need to split the documents. During preprocessing stopwords and also named entities of type 'numex' (numerical) and 'timex' (Kokkinakis, 2004) were removed since they tend to occur as noise in the estimated topics of the model. All words were then converted to lowercase and punctuation markers were removed. We used Latent Dirichlet Allocation (LDA) Blei, Ng & Jordan (2003) implemented in the R-package 'topicmodels' and used in Niekler & Wiedemann (2020), and we experimented with different parameters in order to find an optimal set of topics *k* which can be evaluated qualitatively. As in other similar research we applied Gibbs sampling with *k=20*, the number of iterations set to *i=500* and a low value for *alpha=0.2*, which controls the topic distribution within the documents. Figure 2 shows the aggregated mean topic distributions per month over the period between January–December 2020. The visualization presents a macro view of language change; it clearly shows that topics during the beginning of 2020 (far left in Figure 2) were dominated by words such as 'China', 'Wuhan', 'Chinese' *kinesisk*, 'outbreak' *utbrott*, 'infection' *smitta* and 'animal' *djur*. While topics during the end of 2020 are dominated by words such as 'vaccine' *vaccin*, 'vaccinate' *vaccinera* and 'side effect' *biverkning* and also 'mutated' *muterad*, *variant* and 'British' *brittisk* (far right in Figure 2).

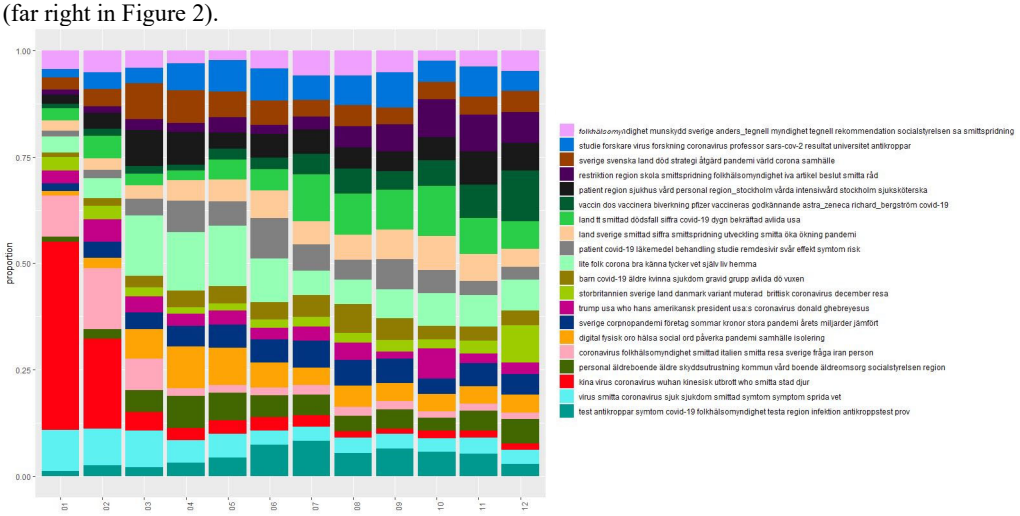


Figure 2. Aggregated mean topic distributions in the sv-COVID-19 corpus in 2020.

Sentiment analysis: is the process of detecting positive or negative sentiment in text (cf. Liu, 2012). The most basic models focus on polarity (*positive*, *negative*, *neutral*). Sentiment analysis is very important because it helps us quickly understand the overall opinions stated in the documents. For the task we explored two open source lexical resources for Swedish: a sentiment list with ca 1800 entries (https://spraakbanken.gu.se/en/resources/sentimentlex) and the SenSALDO lexicon with roughly 12000 entries (https://spraakbanken.gu.se/resurser/sensaldo). Both resources contain both single and multi-word, but are built in a rather different manner. For instance SenSaldo (Rouces et al., 2018) is based on

the Roget's thesaurus. For instance, the word *supernöjd* 'super happy/satisfied' is coded in the first as: *word: supernöjd*, *polarity: pos*, *strength: 3* and *confidence: 0.75* while the same entry in SenSaldo is coded as: *Saldo sense*: *supernöjd..1* and *polarity 1*. Nevertheless, the format was harmonized and merged for the sake of the experiment presented here. Figure 3 shows the negative sentiments in the near context of the word 'elderly' *äldre* in the sv-COVID-19 corpus.

Figure 3. Negative sentiments in the sv-COVID-19 corpus.

4 Conclusion/Future Plans

As the COVID-19 virus became a pandemic in March 2020, the amount of (time-stamped written) data, such as news/newspaper reports, scientific articles, social media posts, surveys and other information about the virus and its symptoms, prevention, management and transmission became massively available. This paper present a new Swedish corpus, the sv-COVID-19, a coronavirus corpus designed and collected over the internet. Such well-defined and timestamped corpus can be applied for both micro- and macro-linguistic exploration and we have illustrated two such scenarios, topic modelling and sentiment analysis. Although the sv-COVID-19 corpus cannot be compared with other mega collections in the same domain, such as English (https://www.english-corpora.org/corona/ and https://www.sketchengine.eu/covid19/) it still provides a good picture of the pandemic reality in the Swedish language landscape. In the near future we intend to expand the content with data from Swedish social media. We believe that such content can reveal a wealth of insights into the language and its speakers during that period (e.g., attitudes and concerns) their coping strategies and psychological outcomes, disinformation themes, but also provide guidance and information on the development and progression of the disease at the scientific and societal level.

References

Blei, D.M., Ng, A.Y. and Jordan, M.I. 2003. Latent Dirichlet allocation. *J of Machine Learn Res*. 3, 993–1022.

Borin, L., Forsberg K. and Roxendal J. 2012. *Korp – the corpus infrastructure of Språkbanken.* The 8th Language Resources and Evaluation Conf. (LREC). Pp 474–478. Istanbul: ELRA.

Kokkinakis D. 2004. *Reducing the Effect of Name Explosion.* Workshop: Beyond Named Entity Recognition, Semantic labelling for NLP tasks. The 4th Language Resources and Evaluation (LREC). Lissabon: ELRA.

Liu B. 2012. *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers.

Niekler, A. and Wiedemann, G. 2020. GPLv3. https://tm4ss.github.io/.

Rouces, J., Tahmasebi, N., Borin, L. and Rødven Eide, S. 2018. *SenSALDO: Creating a Sentiment Lexicon for Swedish.* The 11th Language Resources and Evaluation Conf. (LREC). Miyazaki, Japan.

Schweinberger, M, Haugh, M, and Hames S. 2021. Analysing discourse around COVID-19 in the Australian Twittersphere. *Big Data & Society*. doi.org/10.1177/20539517211021437

Schäfer, R. and Bildhauer, F. 2013. *Web Corpus Construction*. Synthesis Lectures on Human Language Technologies. Graeme Hirst G. (ed). doi.org/10.2200/S00508ED1V01Y201305HLT022.

de Smedt, K 2021. Contagious "Corona" Compounding by Journalists in a CLARIN Newspaper Monitor Corpus. Selected papers from the CLARIN Annual Conference 2020. Linköping Electronic Proc 180: 180 83–92.

SOM. 2020. Coronavirus SOM survey. gu.se/en/som-institute/publications/the-coronavirus-som-survey-2020.

Tan, K.H., Woods, P., et al. 2020. COVID-19 Insights and Linguistic Methods. *Southeast Asian J of English Language Studies*. Vol 26(2):1–23. doi.org/10.17576/3L-2020-2602-01

From Data Collection to Data Archiving: a Corpus of Italian Spontaneous Speech

Daniela Mereu
Free University of Bozen-Bolzano, Italy
daniela.mereu@unibz.it

Abstract

The interest in speech sciences for spontaneous speech has increased, and researchers have begun to study the characteristics of spontaneous and casual speech in different languages, on the basis of spontaneous speech corpora that allow to investigate large amount of data and to formulate more robust theoretical generalizations. For this kind of research on Italian, corpora of spontaneous speech suitable for phonetic analysis are very limited, because the available resources of spoken Italian are not always accompanied by audio files, or the recordings are not suited for acoustic analysis of speech.
The main aim of this proposal is to present a new corpus of Italian spontaneous speech, representing the variety of Italian spoken in Bolzano (South Tyrol, Italy). Special attention will be given to corpus construction procedures, from data collection to database creation. Finally, the way of archiving the corpus in a CLARIN repository will be discussed, in order to reflect on the best practices for making this corpus available to the scientific community and archiving it in a safe and long-term way.

1 Introduction

Recently, the interest in speech sciences for spontaneous conversations has increased, and researchers have begun to study the characteristics of spontaneous and casual speech in different languages, such as German (e.g., Kohler, 1990), Dutch (e.g., Ernestus, 2000; Oostdijk, 2000), English (e.g., Johnson, 2004), French (e.g., Torreira et al., 2010) and Czech (e.g., Ernestus et al., 2014). This kind of research has been driven by the creation of spontaneous speech corpora that allow the systematic investigation on large amounts of data and contribute to understand the empirical dimension of phonological competence.

For research of this kind on Italian, the available resources, i.e. corpora of spontaneous speech that are also suitable for phonetic research, are very limited. Although researchers interested in the study of spoken Italian can access several corpora, these resources are not always accompanied by audio files, or the recordings are not suited to be used for acoustic analysis of speech. To our knowledge, the only exception is represented by a subsection of the CLIPS-*Corpora e Lessici dell'Italiano Parlato e Scritto* corpus (Albano Leoni, 2007; Savy and Cutugno, 2009), stored in the BAS repository (https://clarin.phonetik.uni-muenchen.de/BASRepository/).

The increasing importance of corpora in the speech sciences is closely linked to the advantages they offer: on the one hand, speech corpora allow researchers to conduct systematic research on very large amounts of data; on the other, they contribute to a better understanding of the mechanisms regulating linguistic variation and, in particular, to the formulation of reliable generalisations about the effective

use of a given language variety in a linguistic community (cf. Ernestus and Baayen 2011). Within a research project whose general aim was to study speech variation in a bilingual community, these new theoretical challenges have inspired the construction of the DIA-Dialogic ItAlian corpus (Mereu and Vietti, 2021), a corpus of Italian dialogic spontaneous speech, representative of the Italian language variety spoken in the town of Bolzano (South Tyrol, Italy). This town is characterized by a high degree of linguistic variability, given by the presence of a bilingual Italian-German community, which means that, in Bolzano, Italian is spoken both by native speakers and Tyrolean-native speakers.

This proposal has two goals: a) to illustrate the procedures used to construct the corpus, and b) to reflect on the best practices for making this corpus available to the scientific community and archiving it in a safe and long-term way.

2 Corpus construction procedure

In constructing the DIA corpus we followed principles and methods derived from several disciplines, such as sociolinguistics, phonetics, corpus linguistics and language documentation. In particular, in order to create good quality and long-term data, we have followed recommendations, good practices, tools and methods of language documentation (cf. Austin, 2006).

Data collection sessions were composed of three different parts, all audio recorded: 1) a dialogic interaction in pairs, between people who know each other well, based on topics of interests for speakers; 2) a questionnaire on social networks, elicited by means of EgoNet software (McCarty, 2011); and 3) the reading of a list of sentences, designed to elicit phonological segments of interest for the analysis. Each session lasted approximately 2 hours for each pair of speakers, and it was recorded at 44,100 Hz and 16-bit depth with a Zoom H4 recorder, using headset microphones (Shure SM35). These strategies have made it possible to obtain spontaneous dialogues between pairs of speakers with high acoustic quality (WADA SNR: M = 58.12 dB, SD = 28.32 dB; WADA SNR Waveform Amplitude Distribution Analysis has been measured using Matlab, cf. Kim & Stern, 2008).

The speaker sample is made up of 40 speakers (age range 18-65; 14 M, 26 F), that represent different types of speakers: simultaneous bilinguals, sequential bilinguals, and late sequential bilinguals, or monolinguals. The speakers show also different social characteristics, in terms of levels of education (from middle school to university), and type of occupation.

Overall, the DIA corpus is made up of approximately 30 hours of speech (30 h 43’ 25’’): around 10 hours of dialogic spontaneous speech (9 h 49’ 32’’), 19 hours of speech from the interviews about social networks (19 h 03’ 26’’) and 1 hour and 50 minutes of reading speech (1 h 50’ 27’’). Currently, only the spontaneous speech data have been completely orthographically transcribed, and they consist of around 100,000 tokens.

Since both manual orthographic transcription and phonetic segmentation of spontaneous speech are very time-consuming procedures, we decided to make the transcription process as automatic as possible, from the orthographic to the phonemic transcription. To achieve this goal, we selected freely available tools and arranged them in a pipeline aimed at transforming raw audio recordings into annotated speech data. A first automatically generated orthographical transcription of the audio files has been created using the YouTube subtitle transcription system. This kind of transcription has been manually checked and, if necessary, corrected in ELAN by the author (Sloetjes and Wittenburg, 2008). A second check of all orthographic transcriptions was then made by a research assistant.

Audio files, with its corresponding orthographic transcription, have been processed in WebMAUS (Kisler et al., 2017), using the tools of forced alignment, automatic segmentation and labeling of speech signals. The output of this process is, for each speaker, an audio file with a time-aligned transcription file. The transcription file in TextGrid format consists of three tiers containing information at different levels: an orthographic transcription at the word level, a phonemic transcription of the entire words and a phonetic segmentation in the SAMPA alphabet. The latter tier of annotation is created by means of a system of forced alignment (Kisler et al., 2015), based on the phonemic transcription at the word level.

The result has been then manually corrected by the author. At present, 4 minutes of spontaneous speech for 18 speakers were phonologically segmented and corrected, i.e., more than 70 minutes.

The automation of the various data transcription processes drastically reduced the processing time of the data because the correction of the results obtained was much faster than a possible manual workflow (cf. Cangemi et al., 2019).

Once the data were phonologically segmented and annotated, we used EMU Speech Database Management System (EMU-SDMS), the collection of software tools offered by the Bavarian Archive for Speech Signals (Winkelmann et al., 2017). EMU allows researchers to create, manipulate, query and analyse speech databases through R. Audio files and transcriptions of data prepared have been stored in this database, and we analysed them directly from EMU.

In addition to the benefits of organizing annotations in a hierarchical structure, this database management system also allows users to formulate queries considering different metadata information, such as speaker, gender, age, language skills, social network information, etc., because metadata can be embedded as information about individual audio files.

To summarize, at the moment the corpus consists of data which are all in archive format: 20 dialogic interactions in the form of WAV sound files (44,100 sampling frequency/16 bit resolution), the transcription of the audio files (in eaf and TextGrid format), a rich set of metadata, and consent forms for registration and use of data for scientific purposes for all speakers.

3 Towards archiving the corpus

The next step is the long-term preservation and accessibility of the corpus. In accordance with research reproducibility standards (Berez-Kroeker et al., 2018), to facilitate the access to the corpus for a broad scientific community, we are currently working on sharing it in a CLARIN repository, which also guarantees long-term data preservation.

From the perspective of language documentation, the archiving of data represents an essential and integral part of the language documentation (cf. Austin, 2006: 89). Considering that there are very few corpora of Italian spontaneous speech suitable for phonetic analysis, it is particularly important that this step is achieved. In order to archive the corpus in a CLARIN repository, a number of processes still need to be addressed: a) records metadata has to be transformed into one of the accepted CMDI metadata profiles; b) informed consents with permission to share data from the participants must be obtained; c) audio files and text files must be anonymized.

4 Conclusions

The use of spontaneous speech poses important theoretical challenges and yields evident research benefits. In this respect, we have decided to invest our work in the creation of a new speech resource. Considering the methodological criteria applied in its realization, the DIA corpus is fit for linguistic studies from different perspectives, such as: a) the phonetic characterization of dialogic interaction in spontaneous speech, b) phonetic-phonological analysis, also in relation to c) the investigation of the role of external social factors and, d) the analysis of spoken Italian on different levels of language analysis (i.e. morphological, syntactic and lexical level).

References

Albano Leoni, F. 2006. Un frammento di storia recente della ricerca (linguistica) italiana. Il corpus CLIPS. *Bollettino d'Italianistica*, 4:122–130.

Austin, P. K. 2006. Data and language documentation. In: Jost Gippert, N. P. and Himmelmann, U. M. (eds.), *Essentials of language documentation.* Mouton de Gruyter, Berlin-New York:87–112.

Berez-Kroeker, A. L., Gawne, L., Kung, S. S., Kelly, B. F., Heston, T., Holton, G., Pulsifer, P., Beaver, D. I., Chelliah, S., Dubinsky, S., Meier, R. P., Thieberger, N., Rice, K. and Woodbury, A. C. 2018. Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics* 56(1):1–18.

Boersma, P. and Weenink, D. 2019. *Praat: Doing phonetics by computer*. http://www.praat.org/.

Cangemi, F., Fründt, J., Hanekamp, H. and Grice, M. 2019. A semi-automatic workflow for orthographic transcription and syllabic segmentation. In: Piccardi, D., Ardolino, F. and Calamai, S. (eds.), *Gli archivi sonori al crocevia tra scienze fonetiche, informatica umanistica e patrimonio digitale. Audio archives at the crossroads of speech sciences, digital humanities and digital heritage*. Officinaventuno, Milano:419–425.

Ernestus, M. (2000), *Voice assimilation and segment reduction in Dutch: A corpus-based study of the phonology-phonetics interface*. LOT, Utrecht, The Netherlands.

Ernestus, M. and Baayen, H.R. 2009. Corpora and exemplars in phonology. In: Goldsmith, J., Riggle, J. and Yu, A. (eds.), *The Handbook of Phonological Theory*. Wiley-Blackwell, Malden:374–400.

Ernestus, M., Kočková-Amortová, L. and Pollak, P. 2014. The Nijmegen corpus of casual Czech. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J. and Piperidis, S. (eds.), *Proceedings of LREC 2014: 9th International Conference on Language Resources and Evaluation*, 365–370.

Johnson, K. 2004. Massive reduction in conversational American English. In Yoneyama, K. and Maekawa, K. (eds.), *Spontaneous Speech: Data and Analysis. Proceedings of the 1st Session of the 10th International Symposium*. The National International Institute for Japanese Language, Tokyo:29–54.

Kim, C. and Stern, R.M. 2008. Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis. In *Proc. Interspeech. 9th Annual Conference of the International Speech Communication Association*, Brisbane, Australia: 2598–2601.

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

Kisler, T., Schiel, F., Reichel, U. and Draxler, C. 2015. Phonetic/linguistic Web Services at BAS. *Interspeech* 2015, 2609–2610.

Kohler, K. J. 1990. Segmental reduction in connected speech in German: Phonological facts and phonetic explanations. In Hardcastle, W. J. and Marchal, A. (eds.), *Speech Production and Speech Modelling*. Kluwer Academic Publishers, Dordrecht: 69–92.

McCarty, C. 2011. EgoNet. https://sourceforge.net/projects/egonet/.

Mereu, D. and Vietti, A. 2021. Dialogic ItAlian (DIA): the creation of a corpus of Italian spontaneous speech. *Speech Communication* 130:1-14.

Oostdijk, N. 2000. The spoken Dutch corpus. Overview and first evaluation. In Gravilidou, M., Carayannis, G., Markantonatou, S., Piperidis, S. and Stainhaouer, G. (eds.), *Proceedings of the Second International Conference on Language Resources and Evaluation*, volume 2:887–893, ELRA, Paris.

Savy, R. and Cutugno, F. 2009. CLIPS. Diatopic, diamesic and diaphasic variations in spoken Italian. In Mahlberg, M., Gonzalez-Diaz, V. and Smith, C. (eds.), *Proceedings of the 5th Corpus Linguistics Conference 2009 (CL2009), Liverpool*, 213:1–24.

Schiel, F. 1999. Automatic phonetic transcription of non-prompted speech. In: *Proceedings of the ICPhS 1999*:607–610.

Sloetjes, H. and Wittenburg, P. 2008. Annotation by category – ELAN and ISO DCR. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, 28-30 May 2008.

Torreira, F., Adda-Decker, M. and Ernestus, M. 2010. The Nijmegen corpus of casual French. *Speech Communication*, 52:201–212.

Winkelmann, R., Harrington, J. and Jänsch, K. 2017. EMU-SDMS: Advanced speech database management and analysis in R. *Computer Speech & Language*, 45:392–410.

IceTaboo: A database of contextually inappropriate words for Icelandic

Agnes Sólmundsdóttir
University of Iceland
ags46@hi.is

Lilja Björk Stefánsdóttir
University of Iceland
lbs@hi.is

Anton Karl Ingason
University of Iceland
antoni@hi.is

Abstract

We present IceTaboo, a database of 2725 words that are inappropriate or offensive to at least some speakers in some contexts. Every word is coded for part of speech, a classification of reasons that trigger a negative reaction among some speakers as well as information about the meaning expressed by the word. The database is released under an open CC BY 4.0 license on CLARIN and it is already being used in the development of an automatic proofreading tool, developed in collaboration with an industry partner in commercial software development. The proofreading tool, itself, is under development in an open repository on Github under an MIT license.

1 Introduction

The detection of offensive or contextually inappropriate language can be important for monitoring freely accessible discussion spaces on social media or for helping a user of a word processing system to avoid inappropriate expressions. A variety of resources of methods have been developed to address this challenge (Davidson et al., 2017; Risch et al., 2019; Pitsilis et al., 2018; Wu et al., 2019; Sigurbergsson and Derczynski, 2019; Pitenis et al., 2020; Mubarak et al., 2020; Çöltekin, 2020). One basic kind of a resource that can be integrated into such systems is a database of words that offend readers, in general or in some context. While such a database is not sufficient for detecting all instances of offensive language, it can serve as a useful first step. Furthermore, the development of such a database facilitates the emergence of a classification of the reasons that triggers reactions of this type in a given language.

In the project we present IceTaboo,[1] the Icelandic Taboo database, that was manually compiled at the Language and Technology Lab of the University of Iceland during 2020. IceTaboo contains 2725 words that are inappropriate or offensive to at least some speakers in some contexts. Every word is coded for part of speech, a classification of reasons that trigger a negative reaction among some speakers as well as information about the meaning expressed by the word. The database is released under an open CC BY 4.0 license on CLARIN and it is already being used in the development of an automatic proofreading tool, developed in collaboration with an industry partner in commercial software development. The proofreading tool, itself, is under development in an open repository on Github under an MIT license. Open access availability on CLARIN and an industry-friendly licensing policy ensures that the resource is ready to be used by any software developer that shows interest, thus supporting to the general policies in the current Language Technology Programme for Icelandic (Nikulásdóttir et al., 2020) that aim to make the delivered output as accessible to future development as possible.

2 Offensive language and automatic proofreading

Some previous work on offensive language focuses on training classifiers on annotated training data where the units being annotated are utterances labelled as offensive or not and the classification algorithm

[1] Accessible at: https://repository.clarin.is/repository/xmlui/handle/20.500.12537/64

assigns such utterances to an offensiveness category based on some features extracted from the item in question, for example using bag-of-word classification approaches (Kwok and Wang, 2013; Burnap and Williams, 2015).

Challenges arise when basing the classification on specific word forms, because while a racist word like *n****** triggers a reaction of an offensive experience in generally all contexts, a word like *gay* can be used either in a positive context or a negative one, depending on the intentions of the speaker who produces the utterance and the context in which the relevant discussion takes place. Furthermore, the reason for why a reader may feel offended are of various kinds. One sentence may be perceived as hate speech, invoking strong emotions, whereas another one may be perceived as promoting a negative stereotype, also a negative but less intense emotion. Any of these types of incidents are potential circumstances for intervention, such as filtering or tagging potentially offensive posts on social media or by suggesting a more appropriate choice of words to the user of some word processing software.

Detection of offensive language is an under-explored area of Natural Language Processing for Icelandic. Still, in the context of automatic proofreading, previous work has addressed the disambiguation of confusion sets of the *there/their*-type (Ingason et al., 2009; Friðriksdóttir and Ingason, 2020) and methods that predict the appropriate candidate from a confusion set are in principle applicable to the detection of an appropriate vs. inappropriate use of a word that is offensive in only some contexts. Furthermore, general advances in NLP for the language have paved the way for more advanced context analysis, such as the development of deep phrase structure parsing (Jökulsdóttir et al., 2019; Þorsteinsson et al., 2019; Arnardóttir and Ingason, 2020) – including systems released via CLARIN – which has allowed for complex pattern matching that involves syntactic context. Lemmatization system have also made it possible to normalize Icelandic words to their dictionary forms, a non-trivial task in a language with rich morphology (Ingason et al., 2008; Ingólfsdóttir et al., 2019), and a crucial step in methods that derive features from specific words, and thus important for detection of offensive language.

3 The IceTaboo resource

The IceTaboo database consists of a list of words in Icelandic that may in some way be considered inappropriate, taboo and/or loaded in use or meaning. These can be words such as; words that are biased against certain minorities (i.e. people of different races, abilities, genders or sexualities), words that are derogatory towards people, unnecessarily gendered, obsolete and so on. The list also includes words that are not very inappropriate but can be considered an unfortunate topic for children or politically loaded in any way. The words are grouped together in categories depending on either their meaning, form or use. Each word has then been marked with a short explanation (in Icelandic) on how they can be considered inappropriate and in what context. The words were collected through brainstorming sessions and systematic follow-ups to these (see below), but other similar lists from other sources were also used, i.e. a list of taboo words for children from the project Samrómur (Mollberg et al., 2020) and a list of taboo words for childrens Scrabble developed by the sofware company Miðeind.

This list does not contain actual information or data on the real opinion of the public towards these words. These words are merely thought to elicit a negative reaction for at least some speakers in at least some context. This list can therefore be a useful data set that serves as a starting point for further analysis. The list is already being used in the development of a commercial automatic proofreading system in collaboration with the software company Miðeind.[2] The database identifies the following classes of offensive words:

(1) **Classes of offensive words:**
Generally inappropriate, swear words, words associated with alcoholism or drug addiction, disability words, health-related words, words regarding stupidity, gendered words (generally, or ones that discriminate against either women or men), nasty adjectives, offensive profession names, collocations, LGBTQIA+ words used inappropriately, verbs of inappropriate actions, offensive words related to religion, offensive descriptions of people's appearance, words for gen-

[2]See: `https://github.com/mideind/GreynirCorrect`

tials, offensive prefixes, offensive words related to sex, offensive nationality words (often linked to oppression of some sort).

Additional classes:

(2) **Words with nuanced relationships with offensiveness:**
Inappropriate for children (while not so for adults), political terms (may trigger a negative reaction, depending on a person's political views), non-offensive (words that are not really offensive but have a nuanced meaning that may make sense to exclude in contexts that strive to remain neural), words with an alternative, non-offensive meaning (included to establish that the offensive counterpart reading is only attested in certain contexts).

An example entry is shown below. The word *fóstra* is often considered obsolete and degrading although some members of the profession still prefer to use this word and do not experience it as a negative expression.

- word: *fóstra* (roughly: 'a daycare babysitter')
- part-of-speech: noun
- code (see classes above): m (for profession)
- code2: (additional classification) NA
- meaning: preschool teacher
- reason for offensiveness: Now considered an obsolete and degrading term for the profession of preschool teachers, suggesting they are not a profession of educators.
- additional information (if needed): NA
- alternative non-offensive meaning: NA

The database was compiled manually, using the creativity of a number of research assistants, followed by a systematic search carried out mostly by the first author, supervised by the other two authors. Brainstorming sessions were held, in which RA's thought of all the most offensive words they could think of and the output of this work was used to establish the classification system. Then, each class of words was systematically studied, looking for synonyms or related words that might belong in the list, and electronic resources were used to look for compounds that contained inappropriate parts.

4 Conclusion

In the present paper, we have presented IceTaboo, a novel resource for processing offensive words in Icelandic. Although the work presented here involves manual annotation, and is already being used to highlight inappropriate words in a commercial automatic proofreading system via a simple lookup (after lemmatization), we believe it may also be of use in future development of systems that apply machine learning methods to automatic detection of offensive language. The words in the database can inform feature extraction steps of such systems and potentially make them more effective. We acknowledge that this release is only a small step to that end, yet we believe it is significant and has potential to facilitate further work on the topic in the context of Icelandic.

Various further avenues for future work remain to be explored. As suggested by reviewers, it would be interesting to test these materials on a wider audience to study the reactions of, for instance, younger vs. older speakers. It would also be worth trying to enhance the present work using methods from previous work on sentiment lexicons for various languages and to incorporate crowdsourcing methods in order to further expand the data. It is also a limitation of the present work that it mostly focuses on single-word items, meaning that future work might focus on adding more multi-word expressions to the resource.

Acknowledgements

We would like to thank other members of the Language and Technology lab at the University of Iceland for helpful discussions on this project. This work was supported by the Icelandic Student Innovation fund, grant nr. 2065110091.

References

Arnardóttir, Þ. and Ingason, A. K. 2020. A neural parsing pipeline for Icelandic using the Berkeley Neural Parser. In Navarretta, C. and Eskevich, M., editors, *Proceedings of CLARIN Annual Conference 2020*, pages 48–51.

Burnap, P. and Williams, M. L. 2015. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & internet*, 7(2):223–242.

Çöltekin, Ç. 2020. A corpus of Turkish offensive language on social media. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6174–6184.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*.

Friðriksdóttir, S. R. and Ingason, A. K. 2020. Disambiguating confusion sets in a language with rich morphology. In *Proceedings of ICAART 12*.

Ingason, A. K., Helgadóttir, S., Loftsson, H., and Rögnvaldsson, E. 2008. A mixed method lemmatization algorithm using a hierarchy of linguistic identities (HOLI). In Nordström, B. and Ranta, A., editors, *Advances in Natural Language Processing*, pages 205–216, Berlin, Heidelberg. Springer Berlin Heidelberg.

Ingason, A. K., Jóhannsson, S. B., Rögnvaldsson, E., Loftsson, H., and Helgadóttir, S. 2009. Context-sensitive spelling correction and rich morphology. In Jokinen, K. and Bick, E., editors, *Proceedings of NoDaLiDa 2009*, pages 231–234.

Ingólfsdóttir, S. L., Loftsson, H., Daðason, J. F., and Bjarnadóttir, K. 2019. Nefnir: A high accuracy lemmatizer for Icelandic. In *Proceedings of NoDaLiDa 2019*.

Jökulsdóttir, T. F., Ingason, A. K., and Sigurðsson, E. F. 2019. A parsing pipeline for Icelandic based on the IcePaHC corpus. In Simov, K. and Eskevich, M., editors, *Proceedings of CLARIN Annual Conference 2019*, pages 138–141.

Kwok, I. and Wang, Y. 2013. Locate the hate: Detecting tweets against blacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1621–1622.

Mollberg, D. E., Jónsson, Ó. H., Þorsteinsdóttir, S., Steingrímsson, S., Magnúsdóttir, E. H., and Gudnason, J. 2020. Samrómur: Crowd-sourcing data collection for Icelandic speech recognition. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3463–3467.

Mubarak, H., Rashed, A., Darwish, K., Samih, Y., and Abdelali, A. 2020. Arabic offensive language on twitter: Analysis and experiments. *arXiv preprint arXiv:2004.02192*.

Nikulásdóttir, A. B., Guðnason, J., Ingason, A. K., Loftsson, H., Rögnvaldsson, E., Sigurðsson, E. F., and Steingrímsson, S. 2020. Language technology programme for Icelandic 2019-2023. In *Proceedings of the 12th Conference on Language Resources and Evaluation*, pages 3414–3422.

Þorsteinsson, V., Óladóttir, H., and Loftsson, H. 2019. A wide-coverage context-free grammar for Icelandic and an accompanying parsing system. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, pages 1397–1404.

Pitenis, Z., Zampieri, M., and Ranasinghe, T. 2020. Offensive language identification in Greek. In *Proceedings of the 12th Conference on Language Resources and Evaluation*, pages 5113–5119.

Pitsilis, G. K., Ramampiaro, H., and Langseth, H. 2018. Detecting offensive language in tweets using deep learning. *Applied Intelligence*, 12:4730–4742.

Risch, J., Stoll, A., Ziegele, M., and Krestel, R. 2019. hpiDEDIS at GermEval 2019: Offensive language identification using a German BERT model. In *Proceedings of the 15th Conference on Natural Language Processing KONVENS*.

Sigurbergsson, G. I. and Derczynski, L. 2019. Offensive language and hate speech detection for Danish. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3498–3508.

Wu, Z., Zheng, H., Wang, J., Su, W., and Fong, J. 2019. BNU-HKBU UIC NLP team 2 at SemEval-2019 task 6: Detecting offensive language using BERT model. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 551–555.

The CIRCSE Collection of Linguistic Resources in CLARIN-IT

Marco Passarotti
CIRCSE Research Center
Università Cattolica del Sacro Cuore
Milan, Italy
marco.passarotti@unicatt.it

Rachele Sprugnoli
CIRCSE Research Center
Università Cattolica del Sacro Cuore
Milan, Italy
rachele.sprugnoli@unicatt.it

Abstract

In this paper, we present the collection of the linguistic resources for Latin made available by the CIRCSE Research Center in the CLARIN-IT repository. After an introduction about the history and the main research lines of the Center, the paper provides details both the lexical and the textual resources that were built across more than a decade at the CIRCSE and that are now accessible in CLARIN-IT.

1 The CIRCSE Research Center

The CIRCSE Research Center of the Università Cattolica del Sacro Cuore (Milan, Italy)[1] was founded in 2009 by Marco Passarotti and Savina Raynaud, to keep the legacy of a former Research Group (GIRCSE), which was started by the pioneer of linguistic computing father Roberto Busa at the end of the '70s, in strict connection with his course of Computational Linguistics at Università Cattolica (Bolognesi, 1999).

Following in the footsteps of father Busa, whose main contribution was the Index Thomisticus (IT) corpus collecting the opera omnia of Thomas Aquinas (Busa, 1974–1980), the research topics addressed at CIRCSE focus on building and disseminating linguistic resources and Natural Language Processing (NLP) tools for ancient languages, especially for Latin. Since its beginning, the core project of the Research Center was the Index Thomisticus Treebank, which aims to enhance the texts of the IT with syntactic annotation (Passarotti, 2019). From 2015 to 2017, the CIRCSE hosted a Marie Sklodowska-Curie IF grant[2] focused on Latin derivational morphology. Since 2018, the CIRCSE hosts the *LiLa: Linking Latin* ERC-Consolidator Grant.[3] The objective of LiLa is to build a Knowledge Base of interlinked linguistic resources for Latin according to the principles of the Linked Data paradigm (Passarotti et al., 2019), thus combining computational linguistics, semantic web and classical studies in an interdisciplinary perspective.

Reflecting its main research line, the CIRCSE contributes to CLARIN-IT by sharing the lexical and textual resources for Latin developed across more than a decade at the Center. Given the interdisciplinary nature of its resources, which provide (meta)data in an ancient language built and distributed according to state-of-the-art methods and formats in computational linguistics, the CIRCSE plays a strategic role in the CLARIN-IT context, considering that the Infrastructure aims to impact the entire research community that benefits from easily accessing linguistic data. Such community is large and diverse, including not only NLP scholars and computational linguists, but also digital and traditional humanists, who are interested in finding and accessing the CIRCSE resources in a wide, common repository such as CLARIN-IT.

[1] Centro Interdisciplinare di Ricerche per la Computerizzazione dei Segni dell'Espressione. https://centridiricerca.unicatt.it/circse_index.html.

[2] Agreement No 658332-WFL Word Formation Latin.

[3] https://lila-erc.eu/. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program – Grant Agreement No. 769994.

2 Linguistic Resources

This section presents details the linguistic resources of the CIRCSE Research Center that were made available through the ILC4CLARIN data center[4] in a dedicated collection[5] and the CLARIN Virtual Language Observatory. [6] Table 1 summarizes the type, size and format of the resources.

- LiLa Lemma Bank (Passarotti et al., 2019): a large collection of Latin lemmas, serving as the backbone to achieve interoperability between the resources in the LiLa Knowledge Base, by linking all those entries in lexical resources and tokens in corpora that point to the same lemma. Each lemma is described using a set of grammatical and morphological information, such as the Part-of-Speech and the inflection type; different written representations of the same lemma can be reported (e.g. *metrum*, *metron*, *metrom* for the lemma meaning "a measure").
- Index Thomisticus Treebank (IT-TB) (Passarotti, 2019): syntactically annotated portion of the IT corpus. The IT-TB includes the analytical (i.e. surface syntactic) dependency annotation of the entire "Summa contra Gentiles" (4 books), as well as of the concordances of lemma *forma* ("form") from "Scriptum super libros sententiarum magistri Petri Lombardi" (entire) and from "Summa Theologiae" (partial). The annotation guidelines are inspired by those of the analytical layer of the Prague Dependency Treebank.[7] The resource features also more than 2,000 dependency trees for as many sentences from "Summa contra Gentiles" annotated at the tectogrammatical (i.e. underlying syntactic) layer following the corresponding annotation guidelines of the Prague Dependency Treebank.[8]
- Latin Vallex v.1 (Passarotti et al., 2016): valency lexicon for Latin. The first version was built in close connection with the semantic/pragmatic annotation of the Index Thomisticus Treebank and the Latin Dependency Treebank. Data are stored in a single XML file, whose structure is the same of that for the valency lexicon for Czech PDT-VALLEX.[9]
- LatinAffectus (Sprugnoli et al., 2020a; Sprugnoli et al., 2020c): prior polarity lexicon of Latin lemmas developed by merging a Gold Standard and a Silver Standard. The Gold Standard was created by two experts of Latin language and culture following a multi-stage process and an extensive reconciliation phase. It features a five-way classification: 1 (fully positive), 0.5 (somewhat positive), 0 (neutral), -0.5 (somewhat negative), -1 (fully negative). The Silver Standard was built by deriving new entries starting from those in the Gold Standard through synonym, antonym and derivational relations.
- Index Graecorum Vocabulorum in Linguam Latinam (IGVLL) (Franzini et al., 2020): manually-corrected OCR of Günther Alexander Saalfeld's list of Latin loans from Ancient Greek (1874). It contains the Latin loanword (occasionally accompanied by variants), the Ancient Greek source lemma(s) (many lemmas include graphical, morphological and dialectal variants), and the link between the Ancient Greek lemma and its corresponding canonical forms in a machine-readable version of the Greek-English Liddell-Scott Jones lexicon (Blackwell, 2018).
- Word Formation Latin (WFL) (Litta et al., 2016): derivational morphology resource where lemmas are analyzed into their formative components, and relationships between them are established on the basis of Word Formation Rules (WFRs).
- EvaLatin 2020 Data (Sprugnoli et al., 2020b): training and gold test data released in EvaLatin 2020. The two shared tasks proposed in the campaign, i. e. Lemmatization and Part-of-Speech tagging,

[4] https://ilc4clarin.ilc.cnr.it/
[5] https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/handle/000-c0-111/525
[6] https://vlo.clarin.eu/search?0&fq=collection:CIRCSE
[7] http://static.perseus.tufts.edu/docs/guidelines.pdf
[8] https://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/t-layer/pdf/t-man-en.pdf
[9] http://ufal.mff.cuni.cz/pdt2.0/doc/data-formats/pdt-vallex/pdt-vallex-struct.html

NAME	TYPE	SIZE	FORMAT
LiLa Lemma Bank	lexicon	196,853 lemmas	RDF-Turtle
IT-TB	annotated corpus	450,000 nodes 25,000 sentences	XML
Latin Vallex v1	lexicon	1,426 lemmas 3,650 frames	XML
LatinAffectus	lexicon	2,437 lemmas	CSV
IGVLL	lexicon	1,763 lemmas	CSV
WFL	lexicon	34,951 relations 773 rules	SQL
EvaLatin 2020	annotated corpus	341,419 tokens 16 files	CoNLL-U
EDLIL	lexicon	1,874 lemmas	RDF-Turtle

Table 1: Information about the CIRCSE resources in CLARIN-IT.

were aimed at fostering research in the field of language technologies for Classical languages. The shared dataset consists of texts taken from the Perseus Digital Library, processed with UDPipe models (Straka and Straková, 2017) and then manually corrected by Latin experts. The training set includes only prose texts by Classical authors. The test set, alongside prose by the same authors represented in the training set, also includes poems and texts of the Medieval period.

- The Etymological Dictionary of Latin and the other Italic Languages (EDLIL) (Mambrini and Passarotti, 2020): collection of Proto-Italic and Proto-Indo-European reconstructed forms taken from the most recent "Etymological Dictionary of Latin and the other Italic Languages" (de Vaan, 2008), modeled following the Linked Data paradigm and released in RDF-Turtle format.

3 Examples of Use

By combining the resources described above, it is possible for the users of the CLARIN infrastructure to gather various types of linguistic information about Latin lemmas and their context of use in corpora. For example, the LiLa Lemma Bank reports that the lemma *dignus* ("worthy") is a first class adjective with a positive degree, having a deadjectival adverb (*digne* "worthily") and belonging to the same word formation family of other 25 lemmas such as the verb *digno* ("to deem worthy"). WFL describes how other lemmas derive from *dignus* through derivation or compounding: for example, *indignitas* ("unworthiness") derives from *dignus* by adding the prefix *in*(negation)- and the suffix *-tas/tat*. Semantic roles, called functors following the Functional Generative Description framework (Sgall et al., 1986) and expressing the types of relations between a word and its complements, are reported in Latin Vallex v.1: more specifically, *dignus* has one frame entry and one complement having the role of Patient that can be linguistically realized in four different ways. For example, in *[...] dono dignum esset* ("worthy [of] a gift"; Sallust, *Bellum Catilinae*, 54) the complement is realized with an ablative noun (*dono*). *Dignus* is also an entry in LatinAffectus with a fully positive polarity (+1) and it is a lemma appearing 177 times in the EvaLatin 2020 dataset and 36 times in the IT-TB. One of the occurrences of *dignus* in the IT-TB is annotated at both the analytical and the tectogrammatical layer. The occurrence in question is the following: *ipsa substantia angeli [...] est* ***dignior*** *rebus sensibilibus [...]* "the substance of an angel [...] is nobler than sensible things [...]" ("Summa contra Gentiles", book 1, chapter 3, number 5).[10] While the analytical dependency subtree of this portion of the sentence shows that the node for *dignior* is the nominal predicate of the copula verb *sum*, whose subject is *substantia*, the tectogrammatical tree includes the

[10]Translation taken from: `https://isidore.co/aquinas/ContraGentiles1.htm#3`.

semantic roles played by the content words of the sentence, reporting for instance that *substantia* is the Actor of the clause.[11]

By using the resources in the CIRCSE collection, it is also possible to perform etymological inquires. For example, the user can retrieve the Proto-Italic and Proto-Indo-European reconstructed forms that explain the history of the lemma *classis* ("class/army") from the EDLIL, that is **klāssi-* and **klh$_1$-d$^{(h)}$-ti-* respectively. In addition, the same lemma is described as a loanword of the Ancient Greek noun κλῆσις in the IGVLL.

References

Bolognesi, G. 1999. La «linguistica computazionale» nell'Università Cattolica del S. Cuore e l'origine del termine informatica. *Aevum*, 73:913–920.

Busa, R. 1974–1980. *Index Thomisticus*. Frommann-Holzboog, Stuttgart-Bad Cannstatt.

de Vaan, M. 2008. Etymological Dictionary of Latin: and the other Italic Languages. Brill, Amsterdam https://brill.com/view/title/12612.

Franzini, G., Zampedri, F., Passarotti, M., Mambrini, F., and Moretti, G. 2020. Græcissâre: Ancient Greek Loanwords in the LiLa Knowledge Base of Linguistic Resources for Latin. In Dell'Orletta, F., Monti, J., and Tamburini, F. (eds.), *Proceedings of the Seventh Italian Conference on Computational Linguistics*. Accademia university press, Collana dell'Associazione Italiana di Linguistica Computazionale.

Litta, E., Passarotti, M., and Culy, C. 2016. Formatio formosa est. Building a Word Formation Lexicon for Latin. In Corazza, A., Montemagni, S., and Semeraro, G. (eds.), *Proceedings of the Third Italian Conference on Computational Linguistics*. Accademia university press, Collana dell'Associazione Italiana di Linguistica Computazionale, vol. 2, Napoli, 185–189.

Mambrini, F. and Passarotti, M. 2020. Representing etymology in the LiLa knowledge base of linguistic resources for Latin. In *Proceedings of the 2020 Globalex Workshop on Linked Lexicography*. European Language Resources Association, 20–28.

Passarotti, M. 2019. The Project of the Index Thomisticus Treebank. In Berti M. (ed.), *Digital Classical Philology. Ancient Greek and Latin in the Digital Revolution*. De Gruyter GmbH, Berlin-Boston, 299–319.

Passarotti, M., Gonzalez Saavedra, B., and Onambele, C. 2016. Latin Vallex. A Treebank-based Semantic Valency Lexicon for Latin. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*. European Language Resources Association (ELRA), Portorož, Slovenia, 2599–2606.

Passarotti, M., Mambrini, F., Franzini, G., Cecchini, F.M., Litta, E., Moretti, G., Ruffolo, P., and Sprugnoli, R. 2019. Interlinking through Lemmas. The Lexical Collection of the LiLa Knowledge Base of Linguistic Resources for Latin. *Studi e Saggi Linguistici*, LVIII(1): 177–212.

Sgall, P., Hajicová, E., and Panevová, J. 1986. *The meaning of the sentence in its semantic and pragmatic aspects*. Reidel, Dordrecht.

Sprugnoli, R., Mambrini, F., Moretti, G., and Passarotti, M. 2020a. Towards the Modeling of Polarity in a Latin Knowledge Base. In *Proceedings of the Third Workshop on Humanities in the Semantic Web*. CEUR Workshop Proceedings, Heraklion, Greece, 59–70.

Sprugnoli, R., Passarotti, M., Cecchini, F.M., and Pellegrini, M. 2020b. Overview of the EvaLatin 2020 Evaluation Campaign. In Sprugnoli, R. and Passarotti, M. (eds.), *Proceedings of LT4HALA 2020 Workshop - 1st Workshop on Language Technologies for Historical and Ancient Languages*. European Language Resources Association (ELRA), Paris, 105–110.

Sprugnoli, R., Passarotti, M., Corbetta, D., and Peverelli, A. 2020c. Odi et Amo. Creating, Evaluating and Extending Sentiment Lexicons for Latin.. In *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association (ELRA), Paris, 3078–3086.

Straka, M. and Straková, J. 2017. Tokenizing, PoS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Association for Computational Linguistics, 88–99.

[11] The Actor role is semantically quite underspecified in the annotation guidelines of the Prague Dependency Treebank, which define it as "the human or non-human originator of the event, the bearer of the event or a quality/property, the experiencer or possessor" (`https://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/t-layer/html/ch07s02s01.html`).

'*Cretan Institutional Inscriptions*' Meets CLARIN-IT

Irene Vagionakis
University of Bologna, Italy
irene.vagionakis2@unibo.it

Paola Baroni
CNR-ILC – Pisa, Italy
paola.baroni@ilc.cnr.it

Riccardo Del Gratta
CNR-ILC – Pisa, Italy
delgratta@ilc.cnr.it

Angelo Mario Del Grosso
CNR-ILC – Pisa, Italy
angelo.delgrosso@ilc.cnr.it

Federico Boschetti
CNR-ILC & VeDPH – Pisa, Italy
federico.boschetti@ilc.cnr.it

Tiziana Mancinelli
VeDPH – Venezia, Italy
tiziana.mancinelli@unive.it

Monica Monachini
CNR-ILC – Pisa, Italy
monica.monachini@ilc.cnr.it

Abstract

This paper describes a project in the domain of Digital Epigraphy, named *Cretan Institutional Inscriptions* developed at the Ca' Foscari University of Venice. The project is supported by CLARIN-IT as part of the actions addressed to initiatives, projects and events in the field of Humanities and Social Sciences. The main goal is to make the project visible through CLARIN channels with the hope that it will be a forerunner for other digital epigraphy projects in CLARIN.

The article illustrates also the dockerization process applied to the *Cretan Institutional Inscriptions* project, currently hosted on the CLARIN-IT servers.

1 Project description

The EpiDoc collection named *Cretan Institutional Inscriptions* (Vagionakis, forthcoming) was created as a part of the PhD research project in Ancient Heritage Studies *Kretikai Politeiai: Cretan Institutions from VII to I century BC*, carried out by Irene Vagionakis at the Ca' Foscari University of Venice (UNIVE) from 2016 to 2019, under the supervision of Claudia Antonetti and Gabriel Bodard. The database, built by using the EpiDoc Front-End Services (EFES), collects the EpiDoc editions of 600 inscriptions shedding light on the institutions of the political entities of Crete from the VII to the I century BC.

The project, which contributes to the landscape of Digital Humanities – in particular to that of Digital Epigraphy, through the creation of a new Open Access online epigraphic resource – and could hopefully be a forerunner for other digital epigraphy projects in the Common Language Resources and Technology Infrastructure (CLARIN), has been a valuable opportunity for collaboration with the Venice Centre for Digital and Public Humanities (VeDPH) and the Italian node of CLARIN (CLARIN-IT) during its final testing and publication stages.

1.1 Aim of the research

The Archaic, Classical and Hellenistic history of Crete is a history characterised by a very high level of fragmentation. The frequent silences of the literary sources and the gaps in the epigraphic records have resulted in wide sectors of the island history still being overshadowed, and in a similar fate befalling many of its political entities. The institutional history of Crete, in particular, was greatly affected by such fragmentation, and also by the bulky presence of the albeit scarce literary sources. The alleged greater authoritativeness of authors such as Plato, Aristotle or Ephorus, in fact, often prompted to force what was witnessed by the uneven epigraphic records with the aim of relating to a single model the

contradictory information coming from different areas, flattening however in this way the variegations of the multiform landscape of ≪one hundred-citied Crete≫ (Hom. *Il.* II 649) in the name of the existence of only one unitary *Kretike politeia*. Within that framework, the PhD research project set itself the goal of collecting systematically the records pertaining to Cretan institutions in order to propose an up-to-date reconstruction of the administrative framework of the island political entities, highlighting the specificity of each context, from the rise of the *poleis* and their first epigraphic records in Greek alphabet(s) to the Roman conquest of Crete (VII-I century BC). By bringing together these so far scattered records in a searchable digital collection, the project aimed also at facilitating their finding, consultation and reuse.

1.2 The EpiDoc collection and the TEI catalogues

The core of the documentary basis of the research consisted of 600 Greek inscriptions, either directly mentioning institutional elements (as the decree from Knossos *I.Cret.* I 8 12 of the late II cent. BC: l. 1, ἔδοξεν Κνωσίων τοῖς κόσμοις καὶ τᾶι πόλι, 'the *kosmoi* and the *polis* of the Knossians decreed') or hinting at them through a revealing terminology (as the treaty between Hierapytna and an unknown *polis* *I.Cret.* III 3 6, of the late III or early II cent. BC: ll. 1–2, μηνὸς [- - -] τάδε ἔδ[οξ]εν τ[- - -], 'in the month of [- - -], [- - -] decreed these things').

For each inscription an XML edition compliant with the TEI EpiDoc epigraphic substandard (Elliott et al., 2020) was created, including a descriptive and a bibliographic lemma, the text of the inscription, a selective apparatus criticus and a commentary focused on the institutional data offered by the document, plus links to other related online resources. The EpiDoc markup was especially functional to the research questions in the encoding of the Greek texts inside `<div type="edition">`, where its semantic nature proved to be very helpful for the extraction of the institutional elements and for the analysis of variations in their type and function or sphere of competence, avoiding preconceived generalizations and valuing the specificity of each occurrence. The markup of the institutional elements, in fact, was based on the use of `<rs type="institution">` along with a customized combination of focused attributes: `@subtype` for specifying their typology (such as assembly or board), `@role` for specifying their function or field of action (such as voter or dedicant), `@ref` for specifying their political entity (usually their *polis*), `@key` for facilitating their indexing. A complete example of the markup of an institutional element is the following (*I.Cret.* I 22 4 A, l. 31): `<rs type="institution" subtype="official" role="eponym" key="damiorgos" ref="#olous">ἐπὶ <w lemma="δαμιοργός">δαμιοργοῦ</w> <persName type="attested" key="Leukos"><name nymRef="Λεῦκος">Λεύκου </name></persName></rs>`. In addition to `<rs>`, some other core TEI EpiDoc elements that were used are `<w lemma="">` for the lemmatization of institutional and other relevant terms, `<placeName type="" ref="">` for toponyms and ethnic adjectives, and `<persName type="" key="" ref="">` and `<name nymRef="" type="">` for prosopographic and onomastic elements (for officials, honoured individuals, foreign rulers and theonyms).

In addition to the epigraphic collection, the research outputs include also the creation of two TEI catalogues: one relating to the political entities of Crete (*poleis*, *koina*, dependent communities, extra-urban sanctuaries, hegemonic alliances); another one relating to the attested Cretan institutions (comprising assemblies, boards, officials, associations, civic subdivisions, social statuses, age classes, months, festivities and other celebrations, institutional practices, institutional instruments and public spaces).

1.3 Benefits of using EFES and EpiDoc

The EpiDoc Front-End Services (EFES) are an open source customisable tool for the online publication of ancient documents in EpiDoc XML, inscriptions in primis (Bodard and Yordanova, 2020)[1]. It comes as the EpiDoc specialisation of Kiln, an analogous framework for publishing collections of TEI XML documents, from which it was forked in 2017[2]. The main strengths of EFES, as well as of its ancestor

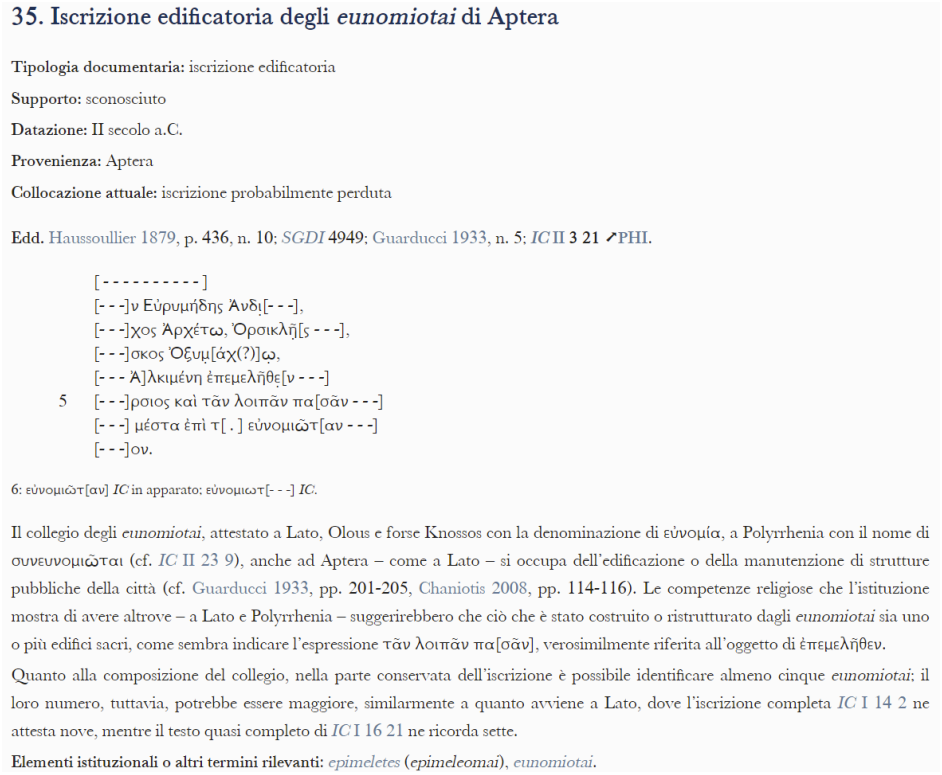
[1] EFES: code `https://github.com/EpiDoc/EFES`, documentation `https://github.com/EpiDoc/EFES/wiki`

[2] Kiln: code `https://github.com/kcl-ddh/kiln/`, documentation `https://kiln.readthedocs.io/en/latest/`

Kiln, are its comprehensiveness, ease of use and high customizability, which make possible the quick creation of a website provided with indices, textual search and browse facilities, even from persons without advanced IT skills. The aim of EFES, in fact, is to allow the production of such outputs also to smaller projects whose teams neither include IT experts devoted to the development of websites, nor have funds to be dedicated to that purpose.

The specific case of *Cretan Institutional Inscriptions* is particularly emblematic of the benefits deriving from the use of EFES, being it the first project carried out by a single person to have used it. Despite the awareness of the importance and usefulness of a collaborative approach to research, the case was that of an individual doctoral project to be completed in three years with no external support, thus perfectly matching the target of users expected by EFES. The timing of the first release of EFES in September 2017, at the end of the first year of the PhD research, was providential and allowed the creation of the website in the remaining two years.

With some customization of the provided XSLT stylesheets, in particular, it has been possible to generate from the EpiDoc markup of inscriptions several custom thematic indexes, recording the occurrences of institutional elements, relevant lemmas, prosopography, onomastics, toponyms, ethnics, theonyms. These indexes, especially that of institutions, are displayed in a tabular format, where all the pieces of information included in the markup are collected in separate columns and can be easily combined and compared with each other. Besides the indexes, the EpiDoc encoding allowed the creation of very specific search filters, thanks to which the inscriptions can be browsed not only according to their traditional metadata (type of document, type of support, date, provenance, current location, bibliographic reference) but also on the basis of the name (e.g. *agela*), type (e.g. tribe) and role (e.g. decreer) of the institutional elements and of the name of the places and divinities mentioned. Another benefit deriving from the EpiDoc encoding is the high level of accuracy of the textual searches performed on the collection, which ignore all the extremely frequent diacritics due to the epigraphic editorial conventions and which can be further refined by including the lemmatized base forms of the terms.



35. Iscrizione edificatoria degli *eunomiotai* di Aptera

Tipologia documentaria: iscrizione edificatoria

Supporto: sconosciuto

Datazione: II secolo a.C.

Provenienza: Aptera

Collocazione attuale: iscrizione probabilmente perduta

Edd. Haussoullier 1879, p. 436, n. 10; *SGDI* 4949; Guarducci 1933, n. 5; *IC* II 3 21 ↗PHI.

[- - - - - - - - - -]
[- - -]ν Εὐρυμήδης Ἀνδι̣[- - -],
[- - -]χος Ἀρχέτω, Ὀρσικλῆ[ς - - -],
[- - -]σκος Ὀξυμ[άχ(?)]ω̣,
[- - - Ἀ]λκιμένη ἐπεμελῆθε̣[ν - - -]
5 [- - -]ρσιος καὶ τᾶν λοιπᾶν πα[σᾶν - - -]
[- - -] μέστα ἐπὶ τ[.] εὐνομιῶτ[αν - - -]
[- - -]ον.

6: εὐνομιῶτ[αν] *IC* in apparato; εὐνομιωτ[- - -] *IC*.

Il collegio degli *eunomiotai*, attestato a Lato, Olous e forse Knossos con la denominazione di εὐνομία, a Polyrrhenia con il nome di συνευνομιῶται (cf. *IC* II 23 9), anche ad Aptera – come a Lato – si occupa dell'edificazione o della manutenzione di strutture pubbliche della città (cf. Guarducci 1933, pp. 201-205, Chaniotis 2008, pp. 114-116). Le competenze religiose che l'istituzione mostra di avere altrove – a Lato e Polyrrhenia – suggerirebbero che ciò che è stato costruito o ristrutturato dagli *eunomiotai* sia uno o più edifici sacri, come sembra indicare l'espressione τᾶν λοιπᾶν πα[σᾶν], verosimilmente riferita all'oggetto di ἐπεμελῆθεν.

Quanto alla composizione del collegio, nella parte conservata dell'iscrizione è possibile identificare almeno cinque *eunomiotai*; il loro numero, tuttavia, potrebbe essere maggiore, similarmente a quanto avviene a Lato, dove l'iscrizione completa *IC* I 14 2 ne attesta nove, mentre il testo quasi completo di *IC* I 16 21 ne ricorda sette.

Elementi istituzionali o altri termini rilevanti: *epimeletes* (*epimeleomai*), *eunomiotai*.

Figure 1. An inscription of the collection

Istituzione	Termine attestato	Individuo	Tipologia	Ambito / Ruolo	Località	Periodo	Occorrenze
Damiorgos	δαμιοργέω	Eteon f. Archetos	Magistrato o funzionario	Dedicante	Aptera	E	seg_60_984.2
Damiorgos	δαμιοργός	A-	Collegio	Eponimo	Olous	E	seg_23_548.2
Damiorgos	δαμιοργός	Arsias	Magistrato o funzionario	Eponimo	Olous	E	[ic1_22_4.B.61]
Damiorgos	δαμιοργός	Autosthenes	Magistrato o funzionario	Eponimo	Olous	E	[ic1_22_4.B.1] ic1_22_4.B.19
Damiorgos	δαμιοργός	Botrynos	Collegio	Eponimo	Olous	E	[seg_23_549.1]
Damiorgos	δαμιοργός		Collegio	Eponimo	Kydonia	E	[seg_41_731.3]
Damiorgos	δαμιοργός		Collegio	Eponimo	Polyrrhenia	E	ic2_23_7.B.1
Damiorgos	δαμιοργός	Leukos	Magistrato o funzionario	Eponimo	Olous	E	ic1_22_4.A.31 ic1_22_4.A.35
Damiorgos	δαμιοργός	Onasandros f. Parmenon	Collegio	Eponimo	Polyrrhenia	E	ic2_23_7.A.1
Damiorgos	δαμιοργός	Sosos f. Tasskos	Collegio	Eponimo	Polyrrhenia	E	[ic2_23_8.1]

Figure 2. An excerpt from the index of institutional elements

2 Cretan Inscriptions at VeDPH

The Venice Centre for Digital and Public Humanities (VeDPH) was inaugurated in 2019 and belongs to the Department of Humanities of the Ca' Foscari University of Venice (UNIVE-DSU). The mission of the centre is the promotion of interdisciplinary methodologies for "the collaborative development of durable, reusable, shared resources for research and learning" (https://www.unive.it/pag/39289/).

VeDPH not only promotes and funds new projects, but also is in charge of legacy projects developed at UNIVE-DSU over the past decades.

Since its foundation, a CNR-ILC detached research unit has been working in collaboration with VeDPH within the *Archipelago DPH* project, in order to ensure that the creation of new digital resources and the maintenance of the legacy ones are effectively durable, reusable and shared.

According to this vision, CLARIN-IT provides the necessary know-how through webinars and seminars at VeDPH, a state-of-the-art technological infrastructure to develop and test the new prototypes created by VeDPH affiliates, a suitable web infrastructure for (permanently or temporarily) hosting the legacy projects developed at UNIVE-DSU and all the CLARIN tools and strategies to make new data and legacy data as FAIR[3] as possible.

In this context, *Cretan Institutional Inscriptions* gave us the opportunity to test this model of collaboration between VeDPH and CLARIN-IT.

3 Cretan Institutional Inscriptions at CLARIN-IT

The Italian Consortium CLARIN-IT[4] has a strong interest in the field of Digital Classics and aims at including a large part of resources for historical languages in its repositories: see (Nicolas et al., 2018) for an overview of the consortium at a whole. The repository hosted at the ILC4CLARIN B Centre[5] already contains important resources, such as the ALIM archive (Ferrarini, 2017), recently presented also at the CLARIN Conference 2020 (Boschetti et al., 2020), as well as many resources[6] from the

[3]FAIR is an acronym for Findable, Accessible, Interoperable, and Reusable `https://www.go-fair.org`

[4]`https://www.clarin-it.it/en/`

[5]`https://ilc4clarin.ilc.cnr.it/en/` and `https://dspace-clarin-it.ilc.cnr.it/repository/xmlui`, respectively

[6]The described resources are available from the VLO: `https://vlo.clarin.eu/search?4&q=CIRCSE`

ERC project "LiLa: Linking Latin"[7]. The deposit and description of the *Cretan Institutional Inscriptions* follow this path and increase the number of resources for Ancient Greek available at the CLARIN Virtual Language Observatory (VLO). Indeed, if VLO is queried for some of the main keywords used to describe the *Cretan Institutional Inscriptions* in the repository[8] (such as, for instance, *epigraphy* or *epigraphic*), only few resources are returned. The authors hope that *Cretan Institutional Inscriptions* pave the way for other similar initiatives to be described in Italy and other countries belonging to CLARIN-ERIC.

3.1 Organizational aspects

In this section, we describe how the *Cretan Institutional Inscriptions* resources are organized in the CLARIN Italian National Consortium. Within the activities of supporting projects and events of the sector of Social Sciences and Humanities, CLARIN-IT hosts the portal of the initiative (`https://www.clarin-it.it/cretaninscriptions`)[9], where the same initiative is exhaustively described. This page contains the other links of the initiative: the search engine (`https://ilc4clarin.ilc.cnr.it/cretaninscriptions`)[10] and the handles of two items in the repository. Indeed, the authors agreed on describing both the dataset[11] and the search engine[12].

The strategy behind this organization is the following: the *Cretan Institutional Inscriptions* has its GitHub repository (`https://github.com/IreneVagionakis/CretanInscriptions`), which contains the dataset, the software for the search engine, some customization and the licenses of use. The GitHub repository contains the releases of the dataset as well. The authors decided to periodically deposit the various releases of the dataset in the CLARIN-IT repository, so that scholars can access the complete data without using the search engine. On the one hand, this approach guarantees the versioning of the dataset and the long term preservation of the data; on the other hand it shares the *Cretan Institutional Inscriptions* with the CLARIN community.

4 Dockerization

CLARIN-IT fosters the adoption of the DevOps methodology for the development, building and deployment of applications. This methodology increases *velocity* in each phase of the process (software development, testing and deployment on production servers), reduces the *variation* of unexpected issues (due for instance to different operative systems or different versions of software libraries) and facilitates the *visualization* of the running components[13]. Thus, CLARIN-IT team is adopting dockers in order to implement a sound development workflow as well as a long term preservation policies for their applications by increasing their collaborative implementation and portability.

With the aim of maximizing the benefits of this methodology, CLARIN-IT runs applications as separate dockers, managed through the Rancher open source environment[14]. In order to be easily managed on the CLARIN-IT servers, also *Cretan Institutional Inscriptions* has been dockerized. In this way, the different technologies and devices (such as the Operative System, the Java Virtual Machine, and the Web Server) are totally separated from technologies and devices of other applications running on the same server.

5 Conclusion

CLARIN-IT is opening up to areas of Digital Humanities that, until a few years ago, were not central to the CLARIN world, such as Digital Epigraphy. The authors hope that a project such as *Cretan Institutional Inscriptions* can contribute to widen the bridge between purely linguistic interests and other areas

[7] `https://lila-erc.eu/`

[8] `http://hdl.handle.net/20.500.11752/OPEN-550`, `http://hdl.handle.net/20.500.11752/OPEN-548`

[9] The PID for the URL is `http://hdl.handle.net/20.500.11752/1002`

[10] The PID for the URL is `http://hdl.handle.net/20.500.11752/1003`

[11] `http://hdl.handle.net/20.500.11752/OPEN-548`

[12] `http://hdl.handle.net/20.500.11752/OPEN-550`

[13] `https://goto.docker.com/rs/929-FJL-178/images/20150731-wp_docker-3-ways-devops.pdf`

[14] `https://www.docker.com`,`https://rancher.com`

of the Humanities. To this end we are implementing a DevOps methodology and a docker infrastructure aimed at hosting such a kind of initiatives.

6 Acknowledgment

We are grateful to Alessandro Enea, CNR-ILC, for the technical support to the porting of the *Cretan Institutional Inscriptions* on the CLARIN-IT servers. We thanks also Franz Fischer, VeDPH, for the organizational support.

References

Bodard, G. and Yordanova, P. 2020. Publication, Testing and Visualization with EFES: A tool for all stages of the EpiDoc XML editing process. *Studia Universitatis Babeș-Bolyai Digitalia*, 65(1):17–35, Dec.

Boschetti, F., Del Gratta, R., Monachini, M., Buzzoni, M., Monella, P., and Rosselli Del Turco, R. 2020. "Tea for two": the Archive of the Italian Latinity of the Middle Ages meets the CLARIN infrastructure. In *CLARIN Annual Conference 2020*, pages 121–125. CLARIN-Virtual Edition.

Elliott, T., Bodard, G., Mylonas, E., Stoyanova, S., Tupman, C., and Vanderbilt, S. e. a. 2020. EpiDoc Guidelines: Ancient documents in TEI XML (Version 9.2).

Ferrarini, E. 2017. Alim ieri e oggi. *Umanistica Digitale*, 1(1).

Nicolas, L., König, A., Monachini, M., Del Gratta, R., Calamai, S., Abel, A., Enea, A., Biliotti, F., Quochi, V., and Stella, F. 2018. Clarin-it: State of affairs, challenges and opportunities. In *Selected papers from the CLARIN Annual Conference 2017*.

Vagionakis, I. forthcoming. Cretan Institutional Inscriptions. A New EpiDoc Database. *Journal of the Text Encoding Initiative*.

Swedish Word Metrics: A Swe-Clarin resource for psycholinguistic research in the Swedish language

Erik Witte
Swedish Institute for Disability Research
Linköping University
Linköping, Sweden
erik.witte@liu.se

Jens Edlund
Speech, Music and Hearing
Royal Institute of Technology
Stockholm, Sweden
edlund@speech.kth.se

Arne Jönsson
Computer and Information Science
Linköping University
Linköping, Sweden
arne.jonsson@liu.se

Henrik Danielsson
Swedish Institute for Disability Research
Linköping University
Linköping, Sweden
henrik.danielsson@liu.se

Abstract

We present *Swedish Word Metrics* (SWM), a new CLARIN resource for calculations of lexical and sub-lexical metrics of Swedish words. The calculations at SWM are based on the AFC-list, which is a freely available lexical database with 816404 entries containing spellings, phonetic transcriptions, word-class assignments, and word frequency data. Besides allowing for easy access to the AFC-list data, the SWM site calculates metrics of orthographic and phonological neighbourhood density, phonotactic probability, orthographic transparency, as well as phonetic and orthographic isolation points. The source code for all calculations has been made publicly available and can be extended with more types of word metrics, whereby it forms a framework for continued word-metric developments in the Swedish language.

1 Introduction

Over the years, researchers within the field of psycholinguistics have noted that certain lexical and sublexical properties systematically impact the process of human word perception. Most prominently, the *word-frequency effect* cause words with high word frequency (WF), i.e. that often occur in spoken or written language, to be more quickly and accurately perceived than less frequently occurring words (Brysbaert et al., 2018).

Alongside WF, also other metrics are important for word recognition. Of these, the effect of neighbourhood density (ND) (Luce and Pisoni, 1998) have been extensively studied. The neighbourhood of a word is typically considered to consist of other similar words and described on a scale ranging from sparse neighbourhoods, with only a few neighbours, to dense neighbourhoods containing many similar words. In the auditory domain, phonologically similar words appear to compete with each other, making the process of word recognition more difficult in dense neighbourhoods compared to sparse. However, when orthographic neighbourhoods are considered, studies have indicated a reversed effect, making spoken high-density words easier to perceive than low-density words. Ziegler et al. (2003) speculated that this could be related to irregularities in how words are represented orthographically. Several studies have since shown that the level of orthographic transparency (OT) influences, not only the reading process, but also word recognition in the auditory domain (Dich, 2014). Also, the likelihood of encountering specific phonemes in different parts of words, a property often referred to as phonotactic probability (PP), seems to have a facilitatory effect upon word recognition (Vitevitch and Luce, 1999).

2 The AFC-list

Forming a base resource for calculations of psycholinguistic metrics in the Swedish language, the lexical database referred to as the AFC-list (Witte and Köbler, 2019) contains spellings, phonetic transcriptions, word-class assignments and WF data drawn from several freely available sources. The AFC-list contains 816404 entries, each constituting a unique combination of phonetic transcription and spelling. The distribution of the number of syllables per word in the AFC-list is presented in the leftmost pane of Figure 1.

Based on the AFC-list, Witte and Köbler (2019) defined Swedish versions of word metrics previously published for other languages — including neighbourhood density, orthographic transparency and phonotactic probability — and also developed several new word-metric algorithms specifically targeted towards Swedish phonology. Finally, Witte and Köbler (2019) calculated their metrics for all entries in the AFC-list and made the results publicly available in supplementary data files.

3 The Swedish Word Metrics website

The purpose of the current study was to make calculations of Swedish word metrics, as well as the content of the AFC-list, directly available via an internet website.

Towards this aim, we have created the Swedish Word Metrics[1] (SWM) website, which implements all word-metric calculations described by Witte and Köbler (2019) as well as functionality for searching directly in the AFC-list. In addition to the word metrics described in Witte and Köbler (2019), we have also implemented a few additional word metrics such as phonetic and orthographic *isolation points* (Marslen-Wilson and Welsh, 1978), as well as the *Orthographic Levenshtein Distance 20* (OLD20) developed by Yarkoni et al. (2008) into the SWM website. The SWM website is hosted by the Swedish national research infrastructure *Språkbanken Tal*, which is a member of Swe-Clarin, the Swedish node in the European CLARIN research infrastructure for language resources and technology.

The word-metric calculation functionality is found on the SWM website's *Calculator* page. Words to input into the calculator should be specified by their spellings, their phonetic transcriptions, and an optional word frequency value. All phonetic transcriptions need to adhere to the transcription convention described in Witte and Köbler (2019). To this aim, the calculator contains an optional transcription checking facility that performs phonological evaluations of the entered phonetic transcriptions. For words already present in the AFC-list, it is sufficient only to supply the spelling; all other data will be drawn automatically from the AFC-list. The calculator interface also allows for customisation of which word-metric calculations to run. This feature is useful since some metrics — for instance ND — can be quite time consuming to calculate.

In the SWM website's *AFC-list* page, the AFC-list can be searched either by specifying any combination of spelling, phonetic transcription and word frequency, or by using a SQL query, enabling full search term flexibility.

Both the word-metric calculation results and the AFC-list search results are returned in tables in which rows represent entered words or matching AFC-list entries and columns present the different word metrics and other descriptive properties of each entry. The tables contain up to a total of 70 columns, all described on the SWM website's *Info* page.

Below we will briefly describe a selection of the lexical and sub-lexical word metrics that can be calculated at the SWM website.

3.1 Frequency-weighted phonological neighbourhood density

Besides raw WF data, the AFC-list also contains the *Zipf-scale* value of each included word. The Zipf-scale is a WF metric developed for the specific purpose of capturing the WF effect upon word recognition (van Heuven et al., 2014). The Zipf-scale value takes into account both the

[1] https://www.sprakbanken.speech.kth.se/data/swm/

total size of the corpus from which the frequency data was derived as well as the number of word types in that corpus. In addition, the Zipf-scale value is constructed to assume that there exists a number of words in the language which are not included in the corpus used. The Zipf-scale values were used by Witte and Köbler (2019) to create a neighbourhood metric that takes both WF and ND into account. The metric was referred to as the Zipf-scale weighted phonetic neighbourhood density probability (PNDP) and expresses the Zipf-scale weighted probability of encountering a specific word given the Zipf-scale values of its phonological neighbours (Witte and Köbler, 2019). The SWM calculator identifies phonological neighbours as other existing words which differ from the target word by an edit distance (i.e., one insertion, deletion or substitution) of one, and share the same number of syllables. The comparison words are primarily taken from the AFC-list, but an option in the SWM calculator also allows manually entered words not present in the AFC-list to be included in the comparisons.

To illustrate, the words *bladet* [blɑːdɐ̣t] (the sheet) and *floder* [flʉːdɐ̣r] (rivers) are very similar in terms of syllabic structure, WF, OT and PP, but differ in their PNDP values (0.16 and 0.78, respectively) since *bladet* has several, more common, phonological neighbours (e.g. *badet* [bɑːdɐ̣t] (the bath), *blodet* [blʉːdɐ̣t] (the blood), *bladen* [blɑːdɐ̣n] (the leaves)) while *floder* has only one, less common, neighbour (*floders* [flʉːdɐ̣s] (rivers')).

3.2 Orthographic transparency

In order to calculate metrics of OT, specific segments of the phonetic transcriptions, here referred to as pronunciations, need to be matched to their corresponding segments in the spelling, i.e. graphemes. In the AFC-list, such grapheme/pronunciation correspondences are called *sonographs*. The SWM calculator determines the sonographs of each entered word using a rule-based parsing algorithm implementing a custom-made finite-state transducer of the same type as described in Witte and Köbler (2019). To exemplify, it parses the sonographs in the word *sakens* [sɑːkɐ̣ns] (the thing's) and the word *djungeln* [j̩øŋːɐ̣ln] (the jungle) into (s-s|a-ɑː|k-k|e-ɐ̣|n-n|s-s) and (dj-j̩|u-ø|ng-ŋː|e-ɐ̣|l-l|n-n), respectively.

Based on the AFC-list sonographs, Witte and Köbler (2019) determined word-specific probabilities for each grapheme to correspond to different pronunciations, here called *grapheme-to-pronunciation* (G2P)-OT, according to the method developed by Berndt et al. (1987). Since basing OT calculations directly upon grapheme-pronunciation correspondences makes the unmerited assumption that readers know the length of each encountered grapheme before parsing it. Therefore, Witte and Köbler (2019) also developed a modified OT metric referred to as the *grapheme-initial letter-to-pronunciation* OT (GIL2P-OT) in which OT accounts for both the process of identifying a grapheme given its first letter and the process of finding the appropriate pronunciation for that grapheme. The SWM calculator uses the probability data calculated in Witte and Köbler (2019) to calculate all three types of OT metrics. While for the example words *sakens* and *djungeln* given above, the WF, ND and PP values returned from the SWM calculator are very similar, the word-average GIL2P-OT values are relatively different (0.99 and 0.92, respectively). This difference is primarily related to the unusual situation in which a grapheme-initial letter *d* initiates the pronunciation [j̩].

3.3 Phonotactic probability

Besides the commonly used, and linguistically neutral, phonotactic metrics by Vitevitch and Luce (2004), Witte and Köbler (2019) introduced a new metric which determines PP separately for different types of syllables as well as different intra-syllabic positions. This metric is referred to as the *normalised stress and syllable structure-based* PP (SSPP). To illustrate, the SWM calculator output for the words *skrattet* [skratːɐ̣t] (the laughter) and *sniglar* [sniːglar] (snails) are very similar in WF, ND and OT, but differ largely in their SSPP values (0.97 and 0.89, respectively). The reason for this difference is the relatively rare occurrences of the bi-phones [sn] in syllable onsets, [iːg] in the rhyme, and [gl] across a syllable boundary.

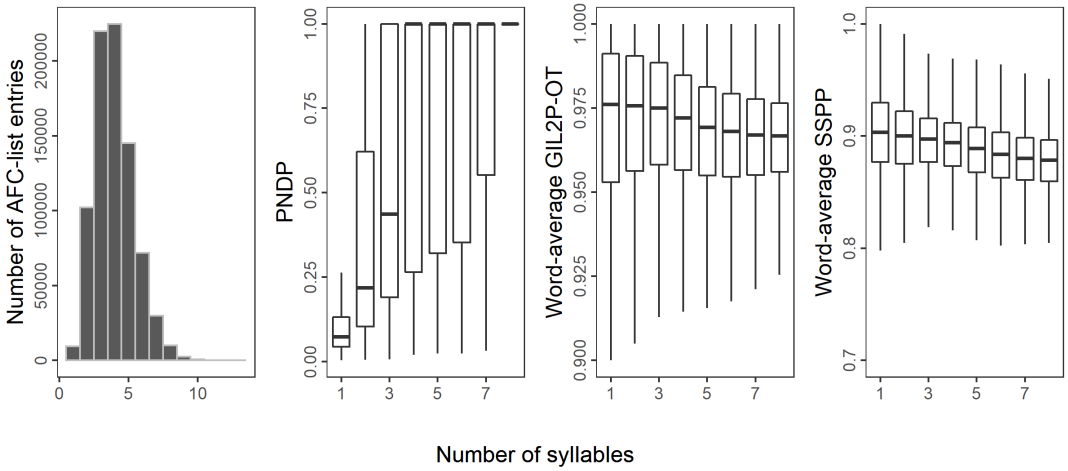


Figure 1. From left to right, a histogram presenting the word lengths (in number of syllables) of all entries in the AFC-list, and boxplots (outliers removed) presenting the distributions of Zipf-scale weighted phonetic neighbourhood density probability (PNDP) values, grapheme-initial letter-to-pronunciation orthographic transparency (GIL2P-OT) values, and word-average normalised stress and syllable structure-based phonotactic probability (SSPP) values for different word lengths in the AFC-list, respectively.

Figure 1 presents the distribution of PNDP, word-average GIL2P-OT and SSPP for different word lengths in the AFC-list. As is clearly seen, PNDP is lowest for short words and increases rapidly with increasing word lengths. PIP2G-OT and SSPP values are relatively stable across word lengths, both showing a slight decrease towards longer words.

4 Potential applications

Since most metrics in the SWM resource were developed with the aim to reflect and quantify different psycholinguistic phenomena, they can potentially be used to model different aspects of human word perception computationally. The large size of the AFC-list makes it an appropriate resource in the creation of experiments by which hypotheses stemming from such models can be tested, thus advancing current theories about human speech perception. Potentially, word metrics such as ND, OT and PP can also be used to improve the quality of synthetic speech. For example, the metrics could be utilised to adjust the local playback speed of speech-synthesis algorithms to assimilate natural variations in reading speed caused by the level of orthographic and phonotactic complexity of specific words, or by the presence of competing phonological neighbours. Furthermore, as OT has bearings upon the ease of phonological decoding of written words, it could likely be used to improve the predictive accuracy of readability metrics such as the commonly used Swedish LIX or other similar metrics (see Heimann Mühlenbock, 2013).

5 Source-code availability

The word-metric calculations used by the SWM website were written in an independent and cross-platform software library using .NET. This backend library is called *Swedish Word Metrics Calculations* (SWMC) and its source code has been made publicly available under an Apache-2.0 license via a link at the SWM website.

References

Berndt, R. S., Reggia, J. A., and Mitchum, C. C. 1987. Empirically Derived Probabilities for Grapheme-to-Phoneme Correspondences in English. *Behavior Research Methods, Instruments, & Computers*, 19(1):1–9.

Brysbaert, M., Mandera, P., and Keuleers, E. 2018. The Word Frequency Effect in Word Processing: An Updated Review. *Current Directions in Psychological Science*, 27(1):45–50.

Dich, N. 2014. Orthographic Consistency Affects Spoken Word Recognition at Different Grain-Sizes. *Journal of Psycholinguistic Research*, 43(2):141–148.

Heimann Mühlenbock, K. 2013. *I See What You Mean. Assessing Readability for Specific Target Groups.* Doctoral dissertation, University of Gothenburg, Gothenburg, Sweden.

Luce, P. A. and Pisoni, D. B. 1998. Recognizing Spoken Words: The Neighborhood Activation Model. *Ear and Hearing*, 19(1):1–36.

Marslen-Wilson, W. D. and Welsh, A. 1978. Processing Interactions and Lexical Access During Word Recognition in Continuous Speech. *Cognitive Psychology*, 10(1):29–63.

van Heuven, W. J. B., Mandera, P., Keuleers, E., and Brysbaert, M. 2014. SUBTLEX-UK: a New and Improved Word Frequency Database for British English. *Quarterly Journal of Experimental Psychology*, 67(6):1176–1190.

Vitevitch, M. S. and Luce, P. A. 1999. Probabilistic Phonotactics and Neighborhood Activation in Spoken Word Recognition. *Journal of Memory and Language*, 40(3):374–408.

Vitevitch, M. S. and Luce, P. A. 2004. A Web-Based Interface to Calculate Phonotactic Probability for Words and Nonwords in English. *Behavior Research Methods, Instruments, & Computers*, 36(3):481–487.

Witte, E. and Köbler, S. 2019. Linguistic Materials and Metrics for the Creation of Well-Controlled Swedish Speech Perception Tests. *Journal of Speech, Language, and Hearing Research*, 62(7):2280–2294.

Yarkoni, T., Balota, D., and Yap, M. 2008. Moving Beyond Coltheart's N: a New Measure of Orthographic Similarity. *Psychonomic Bulletin Review*, 15(5):971–979.

Ziegler, J. C., Muneaux, M., and Grainger, J. 2003. Neighborhood Effects in Auditory Word Recognition: Phonological Competition and Orthographic Facilitation. *Journal of Memory and Language*, 48(4):779–793.

Creating an Error Corpus: Annotation and Applicability

Þórunn Arnardóttir
University of Iceland
Reykjavík, Iceland
thar@hi.is

Xindan Xu
University of Iceland
Reykjavík, Iceland
xindanxu@hi.is

Dagbjört Guðmundsdóttir
University of Iceland
Reykjavík, Iceland
dagu@hi.is

Lilja Björk Stefánsdóttir
University of Iceland
Reykjavík, Iceland
lbs@hi.is

Anton Karl Ingason
University of Iceland
Reykjavík, Iceland
antoni@hi.is

Abstract

In this paper, we describe the Icelandic Error Corpus, a manually annotated error corpus for Icelandic. The Icelandic Error Corpus consists of texts from three sources: student essays, online news and Wikipedia articles, with a total of 56,794 annotated error instances. The corpus is used to analyze errors made by Icelandic native speakers, which are in turn used to guide the development of an Icelandic open-source spellchecker. The corpus is delivered in an augmented TEI format and published under an open-source license.

1 Introduction

The Icelandic Error Corpus is a collection of texts in modern Icelandic which are manually annotated for errors related to spelling, grammar, and other issues. The corpus consists of three genres: student essays, online news and Wikipedia articles. In total, the corpus consists of 4,044 texts with 44,268 revision spans and 56,794 categorized error instances. It is published under a CC BY 4 license and is available from the Icelandic CLARIN repository (Ingason et al., 2021).

A manually annotated error corpus is a useful resource for various tasks within language technology. It can be used to analyze real-world spelling and grammar errors, which in turn can be used to guide the development of a spellchecker. The Icelandic Error Corpus was created for this purpose and it is a novel kind of resource in the context of Icelandic. It reflects the mistakes that Icelandic informants make in written text and is used to measure and improve the performance of an automatic spelling and grammar corrector for Icelandic.

The paper is structured as follows. Section 2 discusses error corpora in general and currently available Icelandic spellcheckers. Section 3 describes the text sources used for the corpus while Section 4 describes the annotation process and Section 5 the annotation scheme. Section 6 gives an overview of the Icelandic Error Corpus and reports on statistical information on it, and we then conclude with Section 7.

2 Error Corpora and Icelandic Spellcheckers

Spelling and grammatical error correction are established tasks within natural language processing. Different methods are available for doing so, some of which are based on an error corpus, a collection of texts which have been annotated for errors. Error corpora can be generated automatically by comparing the edit history of texts (Grundkiewicz and Junczys-Dowmunt, 2014) or by identifying typo edits using a trained classifier (Hagiwara and Mita, 2020). They can also be created by manually annotating text,

which is the case for the Birbeck spelling error corpus (Mitton, 1980). The data comprising an error corpus also differ, some of them consisting of texts written by native speakers (Deksne and Skadina, 2014; Rosner et al., 2012) and others consisting of texts written by non-native speakers (Boyd et al., 2014; Tenfjord et al., 2006; Volodina et al., 2016). Errors within a corpus depend on the data it consists of, which can either be texts written by informants or word lists. An important task for a spellchecker is context-sensitive correction, especially when strings are pronounced the same but are semantically distinct, as in the English pair *there/their*. An Icelandic corpus of such strings already exists (Friðriksdóttir and Ingason, 2020b) and can be used along with a general error corpus, made up of real-world texts, when developing an Icelandic spellchecker. This corpus has been used in recent experiments involving different types of binary classifiers (Friðriksdóttir and Ingason, 2020a; Friðriksdóttir and Ingason, 2020b), expanding on earlier research that depended on more limited data sets (Ingason et al., 2009).

A few Icelandic spellcheckers exist, but they differ with respect to their accessibility as well as the set of features they implement. The Skrambi system is available through an online user interface[1] and is capable of context-sensitive spellchecking (Daðason, 2012). Another spellchecker, Púki, is only available through a fee. It includes a thesaurus and can therefore suggest synonyms for words in a text and learn new words and terms from the text itself. The only open-source Icelandic spellchecker is GreynirCorrect,[2] which is published under the MIT license. The tool returns both errors and suggestions on spelling and grammar. The Icelandic Error Corpus is used to improve the performance of this spellchecker by analyzing real-world examples of spelling and grammatical errors.

3 Data

Three text genres were used in the corpus: student essays, online news texts and Wikipedia articles. These sources were chosen for two main reasons: first, they reflect different styles of writing; and second, they are readily available because they have already been compiled and published, without annotation, as part of the Icelandic Gigaword Corpus (Steingrímsson et al., 2018). The student essays were written by high school students between the age of 16 and 20. These texts were published anonymously in the Icelandic Gigaword Corpus under a license that imposes certain restrictions on derived resources. Therefore, sentences within these texts had to be shuffled before they could be released under an open-source license. Texts in the online news and Wikipedia articles were published under an open-source license in the Icelandic Gigaword Corpus and therefore, they did not need to be shuffled before they could be published as part of the error corpus.

As mentioned in Section 2, the Icelandic Error Corpus is used for developing GreynirCorrect. It was therefore split into a training corpus and a test corpus, which allows the developers to measure the spellchecker's performance on data which the developers have not seen. Random sampling was used to split the corpus into a development corpus, 90% of the total, and a test corpus, the other 10%. Section 6 reports on the number of files and errors in the respective parts of the error corpus.

4 Annotation Process

The annotation process uses a layered approach which culminates in a collection of augmented TEI-format XML documents with the eventual error annotations. The process consists of five steps: text cleanup, proofreading, conversion to TEI-format XML, error code labeling and format checks.

First, a text is converted to the appropriate format, i.e. the XML-format files of the Icelandic Gigaword Corpus are converted to text format, and any extra information is removed. The second step in the process involves manual proofreading and correction using any tool that allows for correcting errors and preserving the original version of the text. Microsoft Word and its Track Changes feature were used for this purpose. The annotators working within this step, 8 in total, were solely proofreaders and could therefore specialize in this task, allowing for more precise and consistent corrections.

After the texts have been proofread, the incorrect and correct versions of each document are aligned and merged. This is done using a Python script which results in an XML structure that explicitly marks

[1] `http://skrambi.arnastofnun.is`
[2] `https://github.com/mideind/GreynirCorrect`

every correction made to the text, by using a revision span. The next step in the process consists of manual annotation of the errors, whereby annotators work with the XML structure, labeling each error with one or more error codes. The annotators working on this step were separate to the proofreaders and specialized in error code labeling. The final step, before publication, is checking each file's format, i.e. ensuring that the XML format is readable and that all labeled error codes are part of the annotation scheme.

5 Annotation Scheme

The annotation scheme developed for the corpus consists of three hierarchical levels: main categories, subcategories, and error codes used during annotation. The annotation scheme is similar to that used in the MERLIN corpus (Boyd et al., 2014). The main categories are six in total, the subcategories are 31 and the error codes are 253.[3] The annotation scheme evolved as more texts were annotated, being descriptive in that it reflects the errors which appear in the corpus and none beyond that. The error codes, the lowest level of the annotation scheme, are precise and there is a clear correspondence between an error and an error code, while the subcategories, the middle level, better reflect the error types in general, e.g. agreement errors, typographical errors, etc.

A group of four annotators, separate to the ones who proofread the texts, worked on error code labeling and created the annotation scheme simultaneously. We believe that this separation between proofreaders and annotators ensures more precise corrections, and it is in contrast to the approach taken in Deksne and Skadina (2014) and Rosner et al. (2012), where proofreaders also annotated the errors. Furthermore, the error annotation in Deksne and Skadina focuses solely on spelling errors and foreign words while the annotation scheme in Rosner et al. is similar to the one used in the Icelandic Error Corpus, only simpler. The first texts in the corpus were annotated by all annotators and then reviewed to ensure that the annotation was agreed upon. Additionally, all annotators had to agree on adding a new error code to the scheme.

Three steps were taken to revise the annotation scheme. First, specialists in language use consultation and spellchecking were consulted. As a result, error codes were refined and redundant error codes were merged or removed from the annotation scheme. Second, 10% of all instances of each error code was sampled and reviewed by the annotators. If a particular error code was incorrectly used for more than 33% of the cases in the sample, all instances of the error code were manually reviewed and corrections made. Third, all instances of each error code are reviewed while developing the spellchecker and corrections made when necessary. All steps lead to both a more refined annotation scheme and more accurate error code labeling.

6 Overview of the Error Corpus

The Icelandic Error Corpus consists of 4,044 texts, which were processed and annotated for errors. A total of 44,268 revisions were made and 56,794 errors annotated. These two numbers are different because a revision span can include more than one error. Table 1 shows the number of files, revisions and categorized error instances in each subcorpus and their respective text genres in the Icelandic Error Corpus. The corpus is delivered in augmented TEI-format XML documents, and is therefore machine-readable. As a result, some corpus management platforms particular to TEI-format files can be used to obtain information from the corpus.

The overall average number of errors per 1000 words in the Icelandic Error Corpus is 45.76. However, there are clear differences in the error rates between genres within the corpus. As is shown in Table 1, the number of errors per 1000 words is lowest in the online news, and highest in the Wikipedia articles. In the development corpus, the number of errors per 1000 words is similar between the online news (35.74) and student essays (37.83), whereas the number of errors per 1000 words is substantially higher in the Wikipedia articles (62.03; Table 1). This trend is also seen in the test corpus, although the number of errors per 1000 words is slightly higher for student essays, and slightly lower for Wikipedia articles.

[3]The complete annotation scheme is available at `https://github.com/antonkarl/iceErrorCorpus/blob/master/errorCodes.tsv`

Subcorpus	Files	Revisions	Categorized Errors	Errors/1000w
Development corpus				
Student essays	158	4,719	5,947	37.83
Online news	2,638	15,969	19,579	35.74
Wikipedia articles	881	20,216	26,786	62.03
Test corpus				
Student essays	18	645	828	43.30
Online news	267	1,334	1,663	32.74
Wikipedia articles	82	1,385	1,991	58.03
Total	4,044	44,268	56,794	45.76

Table 1. Overview of the number of files, revision spans and categorized error instances in both parts of the Icelandic Error Corpus.

Table 2 shows the 10 most common subcategories in the Icelandic Error Corpus, as indicated by the first column. It also lists the most common error codes within each subcategory, ordered by frequency, the subcategory's frequency and its proportion of all errors in the corpus. The most common error type is incorrect use of punctuation, such as when wrong quotes are used. This amounts to 25% of all the errors in the corpus. The second most prominent error type is "wording", which comprises 15% of all errors in the corpus. The remaining subcategories shown in Table 2 have a substantially lower frequency, with a proportion ranging from 7% to 3%.

Subcategory	Main category	Most common error codes	Freq	Prop (%)
punctuation	orthography	wrong-quots, extra-comma, missing-comma	13,357	25.46
wording	style	wording	7,734	14.74
spacing	orthography	missing-hyphen, split-compound, merged-words	3,663	6.98
nonword	orthography	nonword, compound-collocation	3,203	6.11
typo	orthography	missing-letter, letter-rep, extra-letter	2,981	5.68
style	style	nonit4it, it4nonit, fw4ice	2,920	5.57
insertion	vocabulary	extra-word, extra-words	2,885	5.50
syntax	grammar	missing-fin-verb, missing-sub, missing-obj	2,244	4.28
omission	vocabulary	missing-word, missing-words	1,763	3.36
capitalization	orthography	upper4lower-common, lower4upper-proper, upper4lower-proper	1,695	3.23

Table 2. Most common error types in the Icelandic Error Corpus.

7 Conclusion

In this paper, we have described the Icelandic Error Corpus, an open-source collection of texts which have been annotated for errors, and its purpose in developing an Icelandic spellchecker. The corpus consists of three text genres: student essays, online news and Wikipedia articles, which have been annotated for errors regarding spelling, grammar and other issues. The error corpus is published in an augmented TEI format, with revision spans marking the corrections made to a text and error codes for categorizing each error. In total, the corpus consists of 44,268 revision spans and 56,794 categorized error instances.

This manually annotated error corpus is important, not only for developing an open-source spellchecker, but also to depict real-world spelling and grammar errors which Icelandic informants make. The corpus facilitates the development of a spellchecker that takes into account the needs of native Icelandic speakers, so that it can detect and correct errors which are often produced by them.

Acknowledgements

This project was funded by the Language Technology Programme for Icelandic 2019-2023. The programme, which is managed and coordinated by Almannarómur (https://almannaromur.is/), is funded by the Icelandic Ministry of Education, Science and Culture.

References

Boyd, A., Hana, J., Nicolas, L., Meurers, D., Wisniewski, K., Abel, A., Schöne, K., Štindlová, B., and Vettori, C. 2014. The MERLIN corpus: Learner language and the CEFR. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1281–1288, Reykjavík, Iceland, May. European Language Resources Association (ELRA).

Daðason, J. 2012. Post-correction of Icelandic OCR text.

Deksne, D. and Skadina, I. 2014. Error-annotated corpus of Latvian. In *The Sixth International Conference "Human Language Technologies – The Baltic Perspective" (Baltic HLT 2014)*, pages 163–166, 09.

Friðriksdóttir, S. R. and Ingason, A. K. 2020a. Disambiguating confusion sets as an aid for dyslexic spelling. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, pages 1–5, Marseille, France, May. European Language Resources Association.

Friðriksdóttir, S. R. and Ingason, A. K. 2020b. Disambiguating confusion sets in a language with rich morphology. In *Proceedings of ICAART 12 (International Conference on Agents and Artificial Intelligence)*.

Grundkiewicz, R. and Junczys-Dowmunt, M. 2014. The WikEd error corpus: A corpus of corrective Wikipedia edits and its application to grammatical error correction. In Przepiórkowski, A. and Ogrodniczuk, M., editors, *Advances in Natural Language Processing*, pages 478–490, Cham. Springer International Publishing.

Hagiwara, M. and Mita, M. 2020. GitHub typo corpus: A large-scale multilingual dataset of misspellings and grammatical errors. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6761–6768, Marseille, France, May. European Language Resources Association.

Ingason, A. K., Jóhannsson, S. B., Rögnvaldsson, E., Loftsson, H., and Helgadóttir, S. 2009. Context-sensitive spelling correction and rich morphology. In *Proceedings of the 17th Nordic Conference of Computational Linguistics (NODALIDA 2009)*, pages 231–234, Odense, Denmark, May. Northern European Association for Language Technology (NEALT).

Ingason, A. K., Stefánsdóttir, L. B., Arnardóttir, Þ., and Xu, X. 2021. Icelandic error corpus (IceEC) version 1.1. CLARIN-IS.

Mitton, R. 1980. Birkbeck spelling error corpus. Oxford Text Archive.

Rosner, M., Gatt, A., Attard, A., and Joachimsen, J. 2012. Incorporating an error corpus into a spellchecker for Maltese. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 743–750, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Steingrímsson, S., Helgadóttir, S., Rögnvaldsson, E., Barkarson, S., and Guðnason, J. 2018. Risamálheild: A very large Icelandic text corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Tenfjord, K., Meurer, P., and Hofland, K. 2006. The ASK corpus - a language learner corpus of Norwegian as a second language. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May. European Language Resources Association (ELRA).

Volodina, E., Pilán, I., Enström, I., Llozhi, L., Lundkvist, P., Sundberg, G., and Sandell, M. 2016. Swell on the rise: Swedish learner language corpus for European reference level studies. *CoRR*, abs/1604.06583.

ALEXIA: A Lexicon Acquisition Tool

Steinunn Rut Friðriksdóttir, Atli Jasonarson,
Steinþór Steingrímsson, and Einar Freyr Sigurðsson
The Árni Magnússon Institute for Icelandic Studies
Reykjavík, Iceland
{srf2, atj9}@hi.is,
{steinthor.steingrimsson, einar.freyr.sigurdsson}@arnastofnun.is

Abstract

We present a new corpus tool, ALEXIA, which is designed to facilitate research using the Icelandic Gigaword Corpus but can be adapted to any text corpus. The tool aids the compilation and expansion of lexical databases and dictionaries by comparing the vocabulary of the database to that of the corpus in order to find gaps in the data. In particular, two well-known Icelandic language resources are incorporated into the design in order to explore the tool's usage. We describe the design and functionality of the tool, how it can be adapted to various data sources and the process of filtering out noise in order to get a clean list of word candidates. Additionally, we present an extensive list of manually collected stop words that can be used to minimize distortion in research results.

1 Introduction

Using automated methods can expedite the process of compiling dictionaries and lexical resources by identifying and collecting candidate words not included in the lexicon from a relevant text corpus. The vocabulary of the corpus is examined in order to find information that is representative of word frequencies in a given language or domain. While manual analysis has certainly made an important contribution throughout the centuries, the availability of computers radically changed lexicography in the middle of the 20th century (Kennedy, 1998, 5), when corpus linguistics shifted the focus towards a quantitative and descriptive approach to language analysis (Bonelli, 2010).

In this paper, we present a new corpus tool, ALEXIA (Friðriksdóttir and Jasonarson, 2021), whose purpose is to facilitate research in lexicography using the Icelandic Gigaword Corpus (IGC). In Section 2, we discuss the motivation behind the tool and its relation to the CLARIN infrastructure. Section 3 briefly discusses previous work and describes the aforementioned corpus. Section 4 describes the design and architecture of the tool, the filtering applied in order to get the best possible results and our manual compilation of corpus-specific stop words. We conclude in Section 5.

2 Motivation and Relation to CLARIN Infrastructure

In recent years, language technology (LT) has gained some momentum in Iceland, most recently in connection to a national language technology program aimed at building basic LT resources (Nikulásdóttir et al., 2020). This has resulted in a considerable increase of publicly available language resources and LT tools. All data created within the program are distributed with open licenses and made universally and permanently available in the CLARIN-IS repository. ALEXIA will serve as a useful tool to compile larger and better lexical resources within the program and after it comes to an end, and thus serve to improve the data available within the CLARIN infrastructure.

As the number of publicly available Icelandic language resources increases, the creation of lexical databases becomes easier. With the growing volume of data, however, the task of manually overseeing potential data gaps becomes ever more cumbersome. ALEXIA is intended to facilitate the construction or expansion of these lexical databases by sourcing large data sets in order to detect potential gaps in their

vocabulary. ALEXIA offers a structured way for the editors of these databases to compare the vocabulary to that of large corpora and at the same time collect statistics on the words' frequencies, individual word form frequencies and frequencies within specific types of text. The tool can also be used to compose new dictionaries, specialized dictionaries, terminologies or other lexicons.

ALEXIA also offers academic research potential, not least in a sociolinguistic context. When new words gain foothold in the language or when a word gains popularity in an unusual context, it can be an indicator of societal changes (Michel et al., 2011). To name examples from different areas of society, advances in technology require an entirely new vocabulary and gender equality movements and progression in LBTQ+ rights have been making their influence in heavily gendered languages and demand an updated vocabulary. A more recent example is the skyrocketing use of pandemic-related vocabulary that goes hand in hand with changes in the development of COVID-19, see e.g. Þorbergsdóttir and Steingrímsson (2021). Examining changes in word frequencies in various data or from different years is made easy by the use of ALEXIA.

3 Previous Work and Relation to the Icelandic Gigaword Corpus

Corpus-driven approaches to lexicography gained prominence in the 1980s, not least due to the success of the COBUILD project (Sinclair, 1987). Corpus-linguistic tools are often dependent on specific corpora, such as the Danish KorpusDK,[1] but can also be generic and corpus-independent, an example of which is XAIRA, an XML-based tool created for but not limited to the British National Corpus (Xiao, 2006). The functionalities of such tools can vary from simple frequency counts, as can be seen in Michael Barlow's MonoConcEsy,[2] to advanced features such as can be seen in Sketch Engine (Kilgarriff et al., 2014). ALEXIA is designed for the Icelandic LT-resource environment, centering around the IGC which is distributed in TEI P5 format, with options to take advantage of common lexical resources imported as plain text files. The tool can easily be adapted to other languages, having similar resource environments.

ALEXIA is designed to be used with the IGC (Steingrímsson et al., 2018) but is not limited to it. The IGC is a corpus of mostly contemporary texts in continuous collection, the 2019 version consisting of around 1.6 billion running words of text along with their morphosyntactic tags and lemmas. It is by far the largest available text corpus in Icelandic and therefore serves well as a representation of actual usage of the Icelandic language. It has various subcorpora from a variety of genres including parliamentary speeches, literature, media texts, etc. One of the primary objectives of the compilation of the IGC is that it is openly available and constantly expanding. Thus, it is ideal for experimenting with a tool like ALEXIA.

4 Architecture

In this section we discuss ALEXIA's design and intended use. As stated in Section 3, the tool is designed for but not limited to the IGC. ALEXIA is run through the command line and has two language options, Icelandic and English. The user can select the default settings of the tool, which involves either the Database of Icelandic Morphology (Bjarnadóttir et al., 2019) or the Dictionary of Contemporary Icelandic (Jónsdóttir and Úlfarsdóttir, 2019), or they can proceed to choose their own input data. In either case, the user is guided towards setting up the appropriate databases for the lexicon to be examined. Additionally, if the default settings are chosen, a filter database containing approximately 60 thousand stop words is created (discussed in Section 4.2). If the user chooses to use their own data, a filter database can also be created from a list of words provided by the user. The stop words can easily be expanded or modified depending on the individual user's need.

4.1 The Databases

The Database of Icelandic Morphology (DIM) contains inflectional paradigms of 303,067 words as of September 2021. Its development has been continuous since 2004 and its applicability for various assignments has increased over the years. It is widely used not only as a linguistic resource in academic

[1] https://ordnet.dk/korpusdk
[2] https://www.monoconc.com/

research and the development of LT resources but also as a reference for the general public. Dictionary of Contemporary Icelandic (DCI) was first published in 2016 and has been in constant expansion since then. It is based on the multi-lingual Scandinavian dictionary ISLEX (Úlfarsdóttir, 2014) and includes various information such as pronunciation audio, illustrative images, collocations and fixed expressions. As these two databases are in constant expansion, there is a need for quick and concise ways to determine gaps in the data. These databases are therefore ideal for exploring the functionalities of ALEXIA.

As previously stated, the databases are compared to the IGC. The corpus can be used in its entirety, representing the most common registers of written Icelandic, or specific subcorpora can be chosen. The latter option can be beneficial if the user intends to research a specific vocabulary (e.g. sports).

4.2 Filters and Stop Words

ALEXIA is designed around the IGC and therefore its default settings have a number of predefined rules for filtering a word candidate list. Using a tagged corpus can be very beneficial as it provides the option to exclude words tagged with certain parts of speech from the results that may not be well suited for dictionary compilations. Lemmatization also provides the option to exclude all oblique conjugations in order to lessen potential noise in the results. Words with certain POS-tags are excluded from the results, e.g. emails and websites, abbreviations and exclamations. Additional filters are applied, such as excluding words that start or end in a hyphen, and the user can optionally exclude proper names as they tend to overflood candidate lists.

However, not all errors are caught by these filters. As available corpora have increased in volume, the use of automatic POS-tagging and lemmatization has also increased. While their use has certainly sped up and facilitated work in computational linguistics, it is not without its limitations. When a corpus is not corrected by human annotators there will be a certain amount of noise included, especially when words are incorrectly tagged or lemmatized. Perhaps the largest contribution of our work is therefore the list of stop words compiled from the IGC. We have manually collected over 60 thousand typos, misspellings, outdated spellings (e.g. the use of ‘z’, which was replaced by ‘s’ in the official spelling standard in 1973), OCR errors, foreign words, improperly lemmatized words and other non-word tokens. The list greatly minimizes distortion when generating candidate lists.

4.3 Candidate Lists

After comparing the input data to the corpus and filtering the results, ALEXIA creates a candidate list of the user’s choice. The following lists are available:

1. Frequency lists where either all lemmas or all word forms that are not found in the input lexicon are displayed along with their frequencies in the comparison corpus. This can be utilized to estimate the usage of certain words and thus decide if they are suitable candidates. Additionally, information on nouns where the plural form is much more frequent than the singular form can be included, suggesting that a word might only exist in the plural.

2. Frequency lists including the top five collocation examples taken from the corpus. The collocations include the two previous words and the two words following the candidate word. This can be used to determine the context of a word, especially if it is often used in fixed expressions.

3. Frequency lists, where all word forms that appear with a given lemma in the comparison corpus are displayed. This can be useful for determining if a certain word form is the most common and thus if the word is only used in a certain context, e.g. a fixed expression. If word forms are chosen, all lemmas that appear with the word form are displayed. This can be useful for determining if a word can belong to multiple word classes and whether the automatized lemmatization delivers the expected results.

4. Frequency lists where individual word frequencies within certain types of text are displayed. This can be useful for building specialized vocabularies (e.g. related to news, math or sports domains).

5 Conclusion and Future Work

The lexicon acquisition tool, ALEXIA, aims to facilitate the creation and expansion of lexicons and dictionaries by comparing their vocabularies to text corpora, such as the Icelandic Gigaword Corpus. Additionally, it can be used as a resource for academic research, not least concerning sociolinguistics. We include several filtering options in order to deliver appropriate candidate lists, including a stop-word list, a list of POS-tags to exclude and an option to filter out proper nouns if desired. The resulting lists are based on frequencies within the comparison corpus but can include additional information intended to make the lexicographer's work as fast and easy as possible.

As the availability and size of corpora expands, so does the need for a corpus tool like the one we have presented here. Future development of ALEXIA could include ways to visualize in a graph the timeline of a word's use and thus give a better overview of when the word gained foothold in the language or if its usage is declining. We encourage anyone interested to use the tool and modify it as they wish.

Acknowledgements

This project was funded by the Language Technology Programme for Icelandic 2019–2023. The program, which is managed and coordinated by Almannarómur (https://almannaromur.is/), is funded by the Icelandic Ministry of Education, Science and Culture.

References

Bjarnadóttir, K., Hlynsdóttir, K. I., and Steingrímsson, S. 2019. DIM: The Database of Icelandic Morphology. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 146–154, Turku, Finland.

Bonelli, E. T. 2010. Theoretical overview of the evolution of corpus linguistics. In O'Keeffe, A. and McCarthy, M., editors, *The Routledge Handbook of Corpus Linguistics*, pages 14–28. Routledge, London and New York.

Friðriksdóttir, S. R. and Jasonarson, A. 2021. *ALEXIA: Lexicon Acquisition Tool for Icelandic 3.0*. CLARIN-IS, http://hdl.handle.net/20.500.12537/123.

Jónsdóttir, H. and Úlfarsdóttir, Þ. 2019. Íslensk nútímamálsorðabók. *Orð og tunga*, 21:1–26.

Kennedy, G. 1998. *An Introduction to Corpus Linguistics*. Longman, London and New York.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlỳ, P., and Suchomel, V. 2014. The Sketch Engine: ten years on. *Lexicography*, 1(1):7–36.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., The Google Books Team, Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A., and Aiden, E. L. 2011. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182.

Nikulásdóttir, A. B., Guðnason, J., Ingason, A. K., Loftsson, H., Rögnvaldsson, E., Sigurðsson, E. F., and Steingrímsson, S. 2020. Language Technology Programme for Icelandic 2019–2023. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*, pages 3414–3422, Marseille, France.

Sinclair, J., editor. 1987. *Looking up: an account of the COBUILD Project in lexical computing and the development of the Collins COBUILD English Language Dictionary*. Collins, London.

Steingrímsson, S., Helgadóttir, S., Rögnvaldsson, E., Barkarson, S., and Guðnason, J. 2018. Risamálheild: A Very Large Icelandic Text Corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 4361–4366, Miyazaki, Japan.

Úlfarsdóttir, Þ. 2014. ISLEX – a Multilingual Web Dictionary. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, pages 2820–2825, Reykjavik, Iceland.

Xiao, R. 2006. Xaira – an XML aware indexing and retrieval architecture. *Corpora*, 1(1):99–103.

Þorbergsdóttir, Á. and Steingrímsson, S. 2021. Orð ársins 2020: Sóttkví [Word of the Year 2020]. *Hugrás,* https://hugras.is/2021/01/ord-arsins-2020-sottkvi/.

CLARIN Knowledge Centre for Belarusian text and speech processing (K-BLP)

Yuras Hetsevich
UIIP of NASB,
Minsk, Belarus
yuras.hetsevich@gmail.com

Jauheniya Zianouka
UIIP of NASB,
Minsk, Belarus
evgeniakacan@gmail.com

David Latyshevich
UIIP of NASB,
Minsk, Belarus
david.latyshevich@gmail.com

Mikita Suprunchuk
Minsk State Linguistic University, Belarus
ms@philology.by

Valer Varanovich
Belarusian State University,
Minsk, Belarus
gamrat.vvv@gmail.com

Katerina Lomat
UIIP of NASB,
Minsk, Belarus
katerina.lomat@gmail.com

Abstract

This paper represents CLARIN Knowledge Center for Belarusian text and speech processing (K-BLP) which is based at the Speech Synthesis and Recognition Laboratory, the United Institute of Informatics Problems of the National Academy of Sciences of Belarus, Minsk. The CLARIN Knowledge Centre for Belarusian text and speech processing is part of the CLARIN ERIC, which holds the European ESFRI (European Strategy Forum on Research Infrastructures) certification as a landmark research infrastructure.

1 Introduction

The Speech Synthesis and Recognition Laboratory of the United Institute of Informatics Problems of the National Academy of Sciences of Belarus (https://ssrlab.by) established K-BLP center (Figure 1). It provides users with knowledge for text, speech and other data processing for Belarusian, Russian, and English. The K-BLP center proposes tools for text, speech and other data processing for languages, especially for the Belarusian language. The center also offers wide-ranging user support, guidelines and instructions for each service and material.

We are committed to widen the access to Belarusian developments in the computational linguistics environment and popularize our tools within the Republic of Belarus and abroad (Figure 2). It is very important to support available tools and promote them out to improve and facilitate the access for researchers in humanities and social sciences that contributes to wide-ranging user support, guidelines and instructions for each service. The main target audience of K-BLP are researchers in humanities and digital humanities with an interest in different aspects of computational linguistics and natural language processing.

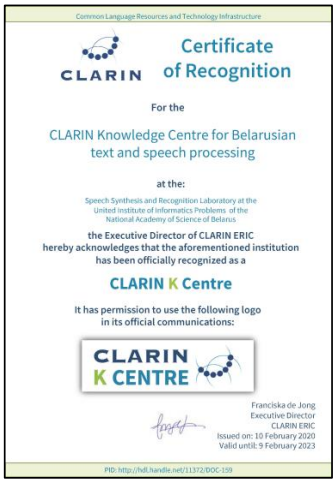


Figure 1. CLARIN Certificate of the Belarusian centre

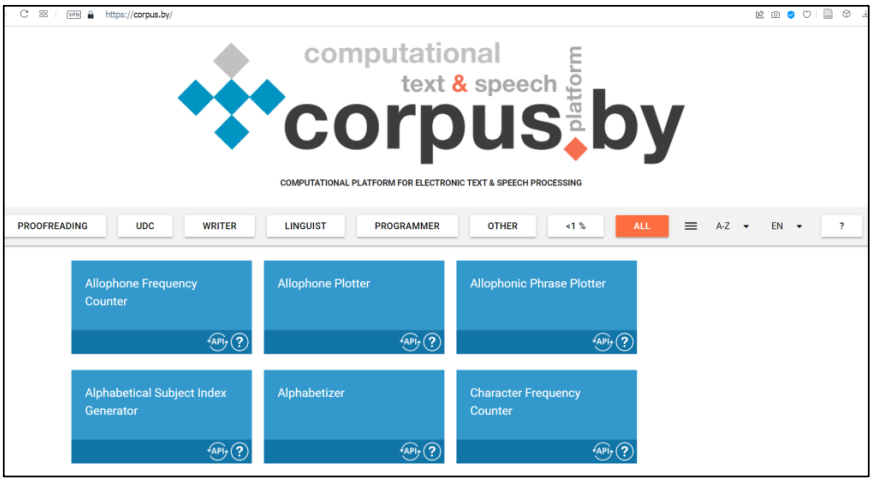


Figure 2. Overview of Belarusian text, speech and other data processors

2 K-BLP center initial activities

K-BLP was formed in September 2020 by the Speech Synthesis and Recognition Laboratory of UIIP NASB. Step by step, it started the process of CMDI metadata creation for all online resources, which means that part of the services is now available via the VLO. Currently, our centre offers Data processing services and tools (the corpus.by platform which includes over 65 services (Dzienisiuk, 2020), a speech intonation analyzer and trainer IntonTrainer (Lobanov, 2019), Belarusian NooJ module for convenient processing of Belarusian language via NooJ linguistic development environment), tutorials and exercises. All provided services can also be accessed through the links directly via http://www.corpus.by/. More information is available on the Speech Synthesis and Recognition Laboratory of UIIP NAS Belarus website.

The Laboratory works on such main scientific research directions as digitization of cultural heritage, high-quality text-to-speech synthesis, robust recognition of discrete and continuous word sequences, computer systems for the rehabilitation of people with hearing and vision disabilities. Except this, we work with systems, programs and platforms for processing big data, universal algorithms for stationery, online and mobile platforms for asynchronous input and output storing and issuing information from different platforms, semi-automatic systematization and processing of data by administrators of target programs (Figures 3–5). Our staff also uses the approaches to processing audio and text forms of speech, which is often found in the development of modern systems that work with the input and output of large-size speech (BigData) on different platforms.

We intend to create and maintain user infrastructure to support the sharing, use and sustainability of big data and tools for research in computational linguistics, the humanities and social sciences. Almost all our digital resources are open, free and available to scholars, researchers and scientists from all spheres through single sign-on access.

All products are made to solve the problems of developing algorithms, resources and methods of Internet input and Internet output of speech, saving and systematizing large volumes of speech. The results can be adapted for wide use in applied and practice-oriented research that requires processing large amounts of data at different levels.

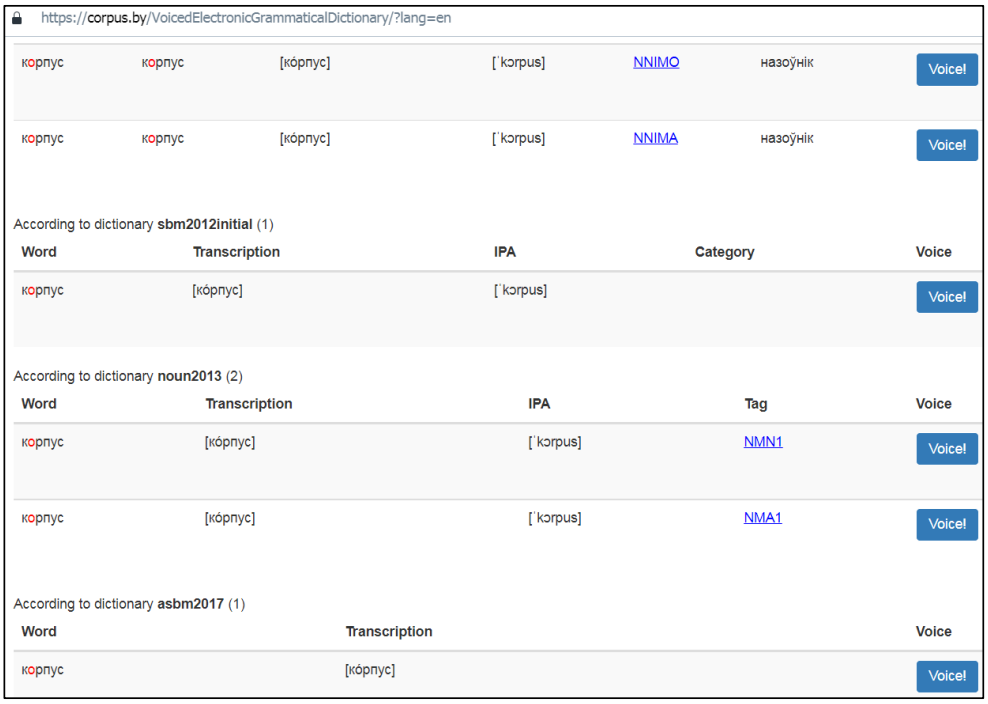


Figure 3. Voiced Electronic Grammatical Dictionary

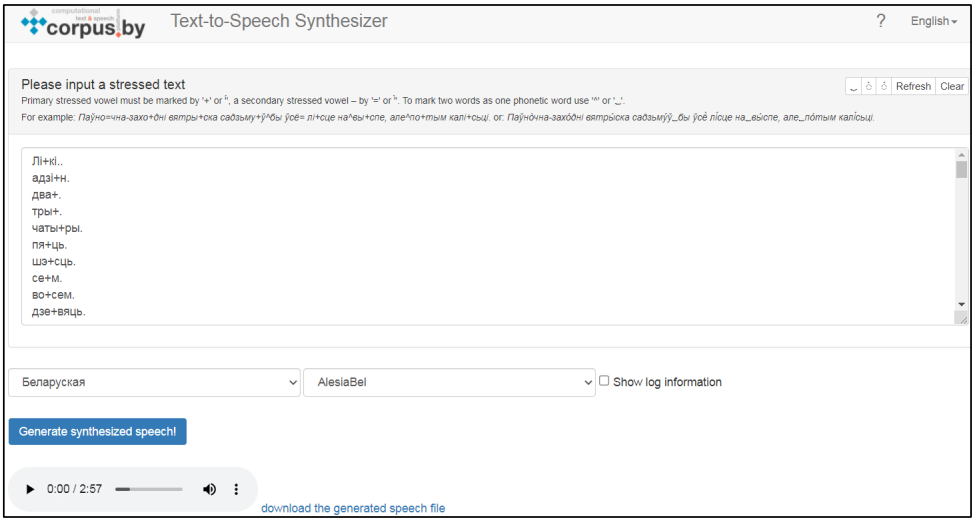


Figure 4. Text-to-Speech Synthesizer

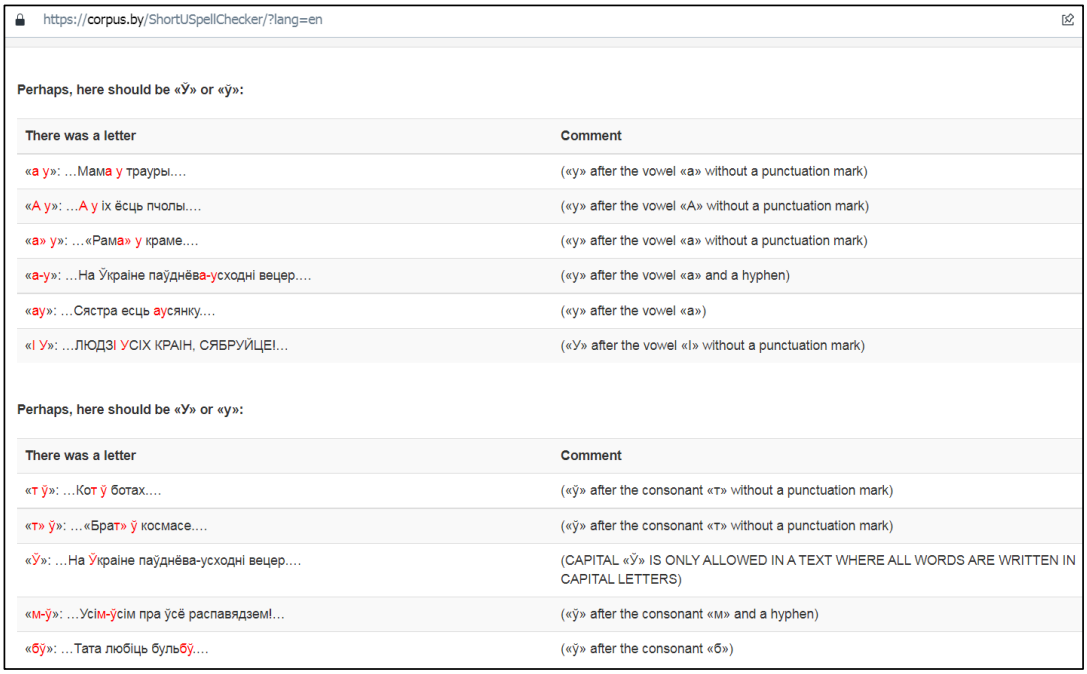


Figure 5. Non-syllable U Spell Checker: [u] or [w]

One more task is to provide a user-friendly overview of the available tools for researchers as well as to organize the overviews of developed methods and algorithms according to the types of data in the resources and listings sorted by language. Our team has great experience in accumulating big data in different formats and platforms. There are specialists in programming, front- and back-end development, project managers, computational linguists and philologists. We are open to create and develop new resources, tools, algorithms and methods according to users' demands.

3 K-BLP's main aims within CLARIN ERIC Research Infrastructure

The main task of K-BLP Center is to extend our resources and tools of natural language processing and organize them according to the types of data within the CLARIN Resource Families in the examples of other Resource families (cf. Franciska, 2020). Increasing the interest in Belarusian developments in computational linguistics and popularizing available tools and resources are dominant directions of K-BLP. To follow these aims, we should widen the number of scientific organizations of K-BLP (except the UIIP of NASB), add new resources and structuralize our Belarusian services within CLARIN classification. It is very important to promote available resources to facilitate access for researchers. That is why we propose wide-ranging user support, guidelines and instructions for each service. We also plan to create and maintain new tools for electronic text and speech processing in the Belarusian language.

Nowadays K-BLP has main strategic priorities such as:

1. To attract other scientific organizations and institutes with research centers for computer processing of the Belarusian language to widen K-BLP (such organizations as Belarusian State University, the Center for the Belarusian Culture, Language and Literature researches of the National Academy of Sciences and other).
2. To expand K-BLP with such resources as new Belarusian corpora (at least 3), dictionaries (nearly 5-7 items) and other tools for computer processing of Belarusian text and speech information (5-7 tools).
3. To annotate and systematize new resources and tools as consistent with description of all resources disposed in other CLARIN ERIC centers.
4. To optimize existing resources and tools in K-BLP according to CLARIN ERIC classification of resources.

5. To organize the overviews of developed Belarusian tools according to the types of data in the resources and listings sorted by language.
6. To provide a user-friendly overview of the available Belarusian language tools in the CLARIN infrastructure for researchers from digital humanities, social sciences and human language technologies.
7. To create and maintain an infrastructure to support the sharing, use and sustainability of Belarusian language data and tools for research in the humanities and social sciences.

We hope to implement our plans listed above in the near future with the help of CLARIN ERIC.

4 Conclusion

Building and running a distributed knowledge center K-BLP for computational linguistics and natural language processing of Belarusian requires samples, text descriptions, demos, courses and possible contacts with specialists of natural language approaches of Belarusian.

K-BLP provides knowledge about tokenization, morphological analysis, voiced electronic grammatical dictionaries, part-of-speech tagging, frequency counting, spell checking, text classification and other approaches used in speech and text processing. It offers special courses in language processing, data analysis and collecting research data for the fast entrance of humanities and others into the digital world of Belarusian data processing.

We are aimed at collecting Belarusian-language linguistic and computer resources for manual and automatic processing in one unit for popularizing the Belarusian language as much as possible. There is a variety of developments in Belarusian, but they are not in the public domain. For this, we want to conduct research in computational linguistics and modern standard Belarusian language and represent them within the K-BLP Center. The future idea is to participate with other CLARIN centres in joint European projects. The plan is to prepare main services and tools from "Computational platform for electronic text & speech processing www.corpus.by" for CLARIN Virtual Language Observatory.

The Speech Synthesis and Recognition Laboratory organises several courses in universities to educate students and researchers in computer linguistics. Several education online materials in English were prepared, such as "Lab 0 – How to be acquainted with text and speech processing services in 10 days?". Introduction into CLARIN project could be presented here, too. All this will allow the introduction of different tools for computational processing of Belarusian for all who are interested in it including foreign scientists and partners.

References

Lobanov, B. and Zhitko, V 2019. Software Subsystem Analysis of Prosodic Signs of Emotional Intonation. *Speech and Computer*: 21st International Conference, SPECOM 2019, Istanbul, Turkey, August 20–25, 2019, Proceedings / Eds. Albert Ali Salah, Alexey Karpov, Rodmonga Potapova. Springer: 280–288.

Dzienisiuk, D. A., Zianouka, Ja. S., Drahun A. Je. [et al.]. 2020. Platforma dlia apracouki tekstavaj i hukavoj infarmacyi dlia roznych tematycnych damienau bielaruskaj movy. *Yazykovaya lichnost' i èffektivnaya kommunikatsiya v sovremennom polikul'turnom mire* : materialy VI Mezhdunar. nauch.-prakt. konf., posvyashch. 100-letiyu Belorus. gos. un-ta, Minsk, 29–30 okt. 2020 g. / Belorus. gos. un-t ; redkol.: S. V. Vorobyeva (gl. red.) [i dr.]. – Minsk : BGU: 69–74.

Franciska, J., Maegaard, B., Fišer, D. [et al.] 2020. Interoperability in an Infrastructure Enabling Multidisciplinary Research: The case of CLARIN. *Proceedings LREC 2020*, 12th International Conference on Language Resources and Evaluation, ELRA. Mode of access: https://www.aclweb.org/anthology/2020.lrec-1.417. Date of access: 02.06.2021.

Enhancing CLARIN-DK Resources While Building the Danish ParlaMint Corpus

Bart Jongejan
Department of Nordic Studies and Linguistics
University of Copenhagen, Denmark
bartj@hum.ku.dk

Dorte Haltrup Hansen
Department of Nordic Studies and Linguistics
University of Copenhagen, Denmark
dorteh@hum.ku.dk

Costanza Navarretta
Department of Nordic Studies and Linguistics
University of Copenhagen, Denmark
costanza@hum.ku.dk

Abstract

In this paper we describe the Danish CLARIN resources, corpora, tools and workflow, which we used and enhanced in order to build the Danish ParlaMint corpus, as part of the CLARIN founded ParlaMint project. More specifically, the article accounts for the manual and automatic processes involved in the preparation of the Danish Parliamentary speeches with focus on the CLARIN-DK tools and Text Tonsorium workflow management. The tools annotated the speeches with metadata and linguistic information in compliance with the common ParlaMint TEI P5 format. As a spin-off of the project, the CLARIN-DK sentence tokenizer and the CST Named Entity Recognizer were improved. These tools, together with the CST-lemmatiser, Danish UD-Pipe software and several data transformation utilities, produced all the linguistic annotations in the correct format. We conclude the paper with a report of a pilot evaluation of the quality of some of the linguistic annotations in the Danish ParlaMint corpus.

1 Introduction

The ParlaMint project (Erjavec et al. 2021) founded by CLARIN ERIC[1] has *inter alia* the aim to create linguistically annotated corpora of parliamentary speeches from different countries which follow the same TEI P5 format. All the corpora cover at least the period 2015-2020, and their metadata contain information about each corpus, comprising the speakers, their parties, and the speech dates. The ParlaMint format and various tools for controlling the format can be found in the ParlaMint repository[2].

The Danish CLARIN group contributed to the second phase of the project and constructed the Danish ParlaMint corpus. The main challenge was not only to adapt the existing resources, tools and workflows available in CLARIN-DK to the task, but also to obtain the most correct output in the short lifetime of the project. As a side effect of this effort, a number of improvements to some of our tools and the Text Tonsorium workflow management system (Jongejan 2020) were done. These improvements and a first evaluation of part of the linguistic annotations in the Danish corpus are the subject of this paper.

The structure of the paper is the following. In Section 2, we shortly present related work, and in Section 3 we describe the collection and metadata annotation of the Danish corpus. In Section 4, the tools, workflow and their improvements are accounted for, followed by a short overview of the linguistic annotations and a first evaluation of some of these annotations (Section5). In Section 6, we conclude and present fuure work.

[1]https://www.clarin.eu/content/parlamint-towards-comparable-parliamentary-corpora

[2] https://github.com/clarin-eric/ParlaMint.

2 Related work

Currently, the transcriptions of parliamentary debates from many countries are publically available, among many the Parliament corpora in the CLARIN resource family (Fišer et al. 2019) in parallel with multilingual corpora, such as the EuroParl corpus (Kohen 2005) and the first ParlaMint corpora.

In Denmark, the Danish Parliament Corpus 2009–2017, v1, consisting of the Hansards of the sittings in the Chamber of the Danish Parliament, has been available as a collection through CLARIN-DK[3] since 2018. This corpus has been used for studying gender differences in the Danish parliamentary speeches (Hansen et al. 2018) and exploring the automatic classification of the subjects and speakers in them (Hansen et al. 2019, Navarretta and Hansen 2020). Moreover, the Danish CLARIN group has participated in the ParlaCLARIN initiative, which also resulted in international workshops, collecting researchers from different disciplines who address parliamentary data, e.g. Fišer et al. (2018), Fišer et al. (2020).

3 The Danish ParlaMint Speeches

The Danish Parliament speeches are made available online in XML format by the administration of the Parliament. The data from the present study partly overlap with data from The Danish Parliament Corpus 2009-2017, v1. The reports from the latest parliamentary year have not been published as the final edition, and are marked as "preliminary version" in the subtitles. The reports of all other meetings are the final editions.

Information from the Hansards was transformed to the TEI-format defined by ParlaMint. Besides structuring the information differently from the downloaded Hansards to secure that all ParlaMint corpora could be alike, new information like birth and gender was added from external sources while other types of information available in the data, such as the agenda title and MP title, were deleted.

The Danish ParlaMint corpus contains transcripts from 688 meetings with 287,144 speeches containing 29,401,796 words. In total 18 political parties are represented in the data with 454 MPs, whereof 193 females and 261 males. In the period covered there were three governments with three different prime ministers.

4 The Linguistic Annotations

The linguistic annotations in the ParlaMint corpus comprise PoS, lemmas, and dependency structures according to the Universal Dependency Grammars, as well as NER. They have been produced with the Text Tonsorium workflow management system (TT henceforth), which can be run through CLARIN-DK (Jongejan, 2020). All the tools needed to produce the final results, with one exception[4], were already integrated in the TT, but the capabilities of some of them had to be expanded. Furthermore, we found inadequacies and bugs in some of the tools that had to be fixed.

Tools integrated in the TT are run as web services. A web service consists of a wrapper, written in PHP, and the tool, most often a program that has a command line interface. Many wrappers apply a format transformation before sending input to the tool, and perform yet another transformation to produce the expected output format. For example, a web service accepting and producing TEI P5 encoded data may internally use the CoNLL-U format.

The TT can handle data in various formats (TEI, HTML, PDF, plain text, etc.), also in combination. In producing the ParlaMint corpus, all involved integrated tools outwardly use a TEI-format. The first workflow step takes unannotated TEI P5 files, and the last step produces linguistically annotated ParlaMint files. All intermediate data are stand-off annotation layers encoded as TEI P5 <text> elements that contain <spanGrp> elements, wherein each element carries the annotation of a token or a sentence. Most of the metadata in the input of the workflow are repeated in its output, but some metadata, such as the usage counts of TEI tags, are recomputed.

The TT already contained tools for annotating TEI P5 input with word and segment delimiters and with PoS tags and lemmas, while the Named Entity and syntax annotation layers were missing. Tools

[3] http://hdl.handle.net/20.500.12115/8.

[4] Except for one simple utility that merely distributes annotation types that are combined in a single file over several files, with one annotation type per file.

for NER and for syntax analysis were also part of the toolbox, but we had to add to them the capability to read and write stand-off annotations in the aforementioned TEI format. Finally, in order to facilitate the evaluation of the annotations, especially of the syntax parser, we added two new tools to the TT: a tool that converts the annotated TEI P5 output to CoNLL-U, and a tool that converts CoNLL-U to a graphical representation. In the following, we shortly account for the tools used to annotate the Danish corpus with linguistic information and the improvements added to them.

The Segmenter/Tokenizer: We improved the treatment of abbreviations, e.g. cases when the character after the final period is a colon, a semicolon or a comma. For example, we distinguished between abbreviations that cannot be the last token in a sentence and those that can. In the latter case, the final period can have the double function of delimiting the abbreviation and the sentence.

We implemented recognition of 18 multicharacter name initials that previously were interpreted as the last word in a sentence, such as Chr. and Joh., and we improved the handling of ellipsis ("…"). Moreover, the tool can now handle TEI elements that can occur inside sentences or sometimes even inside words: <add>, <app>, <del>, <ex>, <lem> and <rdg>. We added the correct handling of many TEI elements that contain sentences, such as <u>, <s>, <seg>.

dapipe[5] is a UDPipe tool for Danish. Its default behaviour is to take plain text input and to segment, tokenize, POS-tag, lemmatize, and annotate it with syntactic dependencies. We wished to apply segmentation and tokenization according to our rules, since the ParlaMint data was already segmented to some extent, namely in utterances (<u> elements), and dapipe did not correctly identify many of these utterances. We tried to pass each utterance separately to dapipe, but this was a slow process due to the overhead of starting, initialising, and terminating dapipe for each utterance. Then, we run dapipe with CoNLL-U formatted input instead of plain text. For each day in the parliamentary year, we created a single CoNLL-U file, which was sent to dapipe. This tremendously speeded up the process.

Evaluating the output from dapipe, we discovered two disadvantages of applying its all-in-one philosophy. The first was the low quality of lemmatization compared to that obtained by our lemmatizer (CSTlemma[6]). The second was that the lemmas produced by dapipe in many cases were inconsistent with the tokens' POS tags. So it seems that lemmatization and POS tagging are performed by independent processes. We decided to solve these issues by replacing the lemma outputs of dapipe with outputs from CSTlemma, feeding CSTlemma with input words and their POS tags computed by dapipe. In this set-up, a new type of error appeared: if the POS-tag was wrong, the lemma might be wrong as well, but we decided that accepting these errors was less serious than accepting inconsistent POS and lemmas.

Since dapipe-processing could take tens of seconds for a single file, the wrapper sent a 'HTTP 202 Accepted' return code to the TT allowing the result to be delivered at a later time. This enabled the TT to activate other workflow steps while dapipe was busy, e.g. starting to process the next ParlaMint file. This way, batches of about 110 files (one parliamentary year) could be processed in a couple of hours, depending on the amount of available CPUs. All data could then be processed in a working day. This was extremely useful, since we spent several days in re-processing all data, after we found issues that had to be corrected. As a result of this process, not only the Danish ParlaMint data were annotated, but also several tools in the TT were improved (debugged and refined), to the advantage of future projects.

CSTlemma: We added support for input POS-tagged in accordance with the Universal tag set, which is the tag set utilized by dapipe.

CSTNER: This is a classical NER program. CSTNER uses handwritten heuristics and tables with names, organizations and locations. We made sure that the names of all members of the parliament and of all political parties were entered in these tables. We found and fixed a few bugs in the program.

5 The annotations

In the Danish ParlaMint corpus 1,406,001 sentences were identified and 586,708 name entities. They were distributed into four categories as shown in Table 1.

PERSON	LOCATION	ORGANIZATION	MISCELLANEOUS
156,198	144,074	273,480	12,952

Table 1: The distribution of NE types in the Danish ParlaMint corpus

[5] https://github.com/ITUnlp/dapipe
[6] https://github.com/kuhumcst/cstlemma

5.1 Pilot evaluation

A first manual evaluation of part of the linguistic annotations (sentence and token identification, PoS, lemma and NE annotations) was performed on the first 5058 tokens of one of the Danish ParlaMint speeches from 2019 by one of the authors. The speech was randomly chosen. There were two cases in which a full stop following a number, was incorrectly considered to be part of the numeral. We found 216 erroneously tagged tokens (4,427% of the data). The most frequent errors being wrong tagging of auxiliary and main verb occurrences of the verbs *være* (be), *have* (have) and *blive* (be, become), the two readings of *det* (it, the), of *som* (like, that/which) and *der* (there, that/which). In addition, some title abbreviations were identified as proper names and proper names were tagged as nouns. The majority of lemmatization errors were due to incorrect PoS, but we also found few cases of wrong lemmas with correct PoS tags, e.g. some forms of the verbs *være* (be) and *synes* (think). Since the CSTNER was fed with lists of proper names and organizations, we did not find any errors in its classification (155 NE), apart from missing annotations of four organizations, due to wrong PoS tags (*DCE*, *Århus Universitet*, *Miljøstyrelsen*, *Dansk Industri*). The PoS-tags and the syntactic features produced by the dapipe are not always consistent, as it was the case for the PoS and lemmas, but we have not yet evaluated these errors. In the future, the parsing results will also be evaluated.

6 Conclusion

In this paper, we have described the process of constructing and annotating the Danish ParlaMint corpus as part of the CLARIN ERIC funded ParlaMint project. The corpus is partially overlapping with the Danish Parliament Corpus 2009–2017, v1. The work for annotating the corpus resulted in an enhancement of the CLARIN-DK tools and the TT workflow management system. The evaluation during the construction of the corpus and the pilot evaluation show that it would be useful to separate all the different tools involved in the annotations in order to improve each tool's performance as well as tune the tools to various domains. In the future, we will evaluate more extensively all annotations. We also wish to retrain the LINDAT/CLARIN UDpipe[7] with data annotated with the CST and CLARIN-DK tools in order to avoid inconsistencies between the various annotation levels, and improve where possible the results of the automatic annotations.

References

Erjavec, T., Ogrodniczuk, M., Osenova, P. et al. 2021. ParlaMint: Comparable Corpora of European Parliamentary Data. In *Proceedings of CLARIN 2021*.

Fišer, D., Eskevich, M. and de Jong, F. (eds.). 2018. *Proceedings of LREC 2018 Workshop ParlaCLARIN: Creating and Using Parliamentary Corpora*, ELRA.

Fišer, D., Lenardič, J. and Erjavec, T. 2018. Meet CLARIN's Key Resource Families. In *Proceedings of LREC 2018*, May, Japan, ELRA, pp. 1320-1325.

Fišer, D., Eskevich, M. and de Jong, F. (eds.). 2020. *Proceedings of LREC2020 Workshop: Creating, Using and Linking of Parliamentary Corpora with Other Types of Political Discourse (ParlaCLARIN II)*, ELRA, May.

Hansen, D. H., Navarretta, C., and Offersgaard, L. 2018. A Pilot Gender Study of the Danish Parliament Corpus. In Fišer et al., editors, *Proceedings of LREC 2018 Workshop ParlaCLARIN*, May, Japan, ELRA.

Hansen, D.H., Navarretta, C., Offersgaard, L., Wedekind, J. 2019. Towards the Automatic Classification of Speech Subjects in the Danish Parliament Corpus. In *DHN 2019 Digital Humanities in the Nordic Countries Proceedings*, Copenhagen 5-8 March CEUR Workshop Proceedings, Vol. 2364, 2019, pp. 166-174.

Jongejan, B. 2020. The CLARIN-DK Text Tonsorium. In *Proceedings of CLARIN 2020*, pp. 93-97.

Navarretta C. and Haltrup Hansen, D. 2020. Identifying Parties in Manifestos and Parliament Speeches. In *LREC2020 Workshop ParlaII*, pp. 51-57, May 2020, ELRA.

[7] https://lindat.mff.cuni.cz/services/udpipe/.

Annotation Management Tool: a Requirement for Corpus Construction

Yousuf Ali Mohammed, Arild Matsson, Elena Volodina
University of Gothenburg, Sweden
name.surname1.surname2@svenska.gu.se

Abstract

We present an annotation management tool, *SweLL portal*, that has been developed for the purposes of the SweLL infrastructure project for building a learner corpus of Swedish (Volodina et al., 2020). The *SweLL portal* has been used for supervised access to the database, for data versioning, export and import of data and metadata, statistical overview, administration of annotation tasks, monitoring of annotation tasks and reliability controls. The portal was developed driven by visions of longitudinal sustainable data storage and was partially shaped by situational needs reported by the portal users, including project managers, researchers, and annotators.

1 Introduction

An infrastructure project dealing with construction and annotation of an electronic learner corpus entails collection of data and metadata, followed by meticulous selection of essays for manual annotation to ensure balance and representativity of various metadata (e.g. balance between texts of different genres and topics, between writer's gender and education level, etc.) and the annotation itself. There are four pillars that tend to be named in connection to digital infrastructures: data, tools for data annotation, tools for data exploration, and expertise (Volodina et al., 2016).[1] What is usually overlooked is some project management environment.

Fort (2016) and Hovy and Lavid (2010) emphasize the need for an annotation management software that would ensure reliability of manual annotations. There are several reasons for that: First, a corpus of good quality must boast representativeness of the language it embodies and balance of the samples that characterize the language. This requires monitoring of collected text instances with regards to various types of metadata. Second, the data as such is only the first step, the most interesting research can be done when data is annotated for one or another text- or language-related feature, and this annotation should better be good. "Tools decay, data stay" - is only true when data is reliably annotated.

The debate on the quality of data annotation often goes in the direction of (1) *tag sets* - their size and ambiguity, (2) *guidelines* - their clarity and deegree of detail, and (3) *tools used for annotation* - their user-friendliness and support in annotation. All too often, the annotation management as such - database handling, statistical overviews, inter-annotator agreement controls, etc. - are overlooked or simply not considered in time. This often ends in a project annotation being managed using excel files, which leads to misses, disbalance, loss of annotation or information, and ultimately - to reduction of annotation quality (Stemle et al., 2019).

Given that the *SweLL-gold corpus* we have been constructing over the past several years would be offered to researchers, developers and teachers to promote the fields of Second Language Acquisition (SLA), Language Assessment (LA), Intelligent Computer-Assisted Language Learning (ICALL) and Language Technology approaches to those – predominantly within Clarin and other European (due to the GDPR restrictions) user groups, we aimed at high annotation standards. The *SweLL portal* came as

[1] https://spraakbanken.gu.se/projekt/swell

one of the steps to ensure those standards. Looking back at our experiences and analyzing the benefits of an annotation management tool, we can say that its use has helped in more ways than just the corpus preparation. Among others, we have tested uploading other (bonus) learner corpora to the portal, and exporting them from the portal applying a unified set of metadata attributes and values (using "N/A" as a value for absent attributes). This step has helped us make several Swedish learner corpora interoperable between each other, *interoperability* being a known challenge in Clarin-related context (König et al., 2021; Stemle et al., 2019; Volodina et al., 2018).

2 Data management

The *SweLL portal* is a user-friendly tool for metadata entry and annotation management. The three modules seen in Figure 1 are loosely connected in code, so that another hypothetical project might replace only the *datacollection* and/or *annotation* modules with their custom implementations.

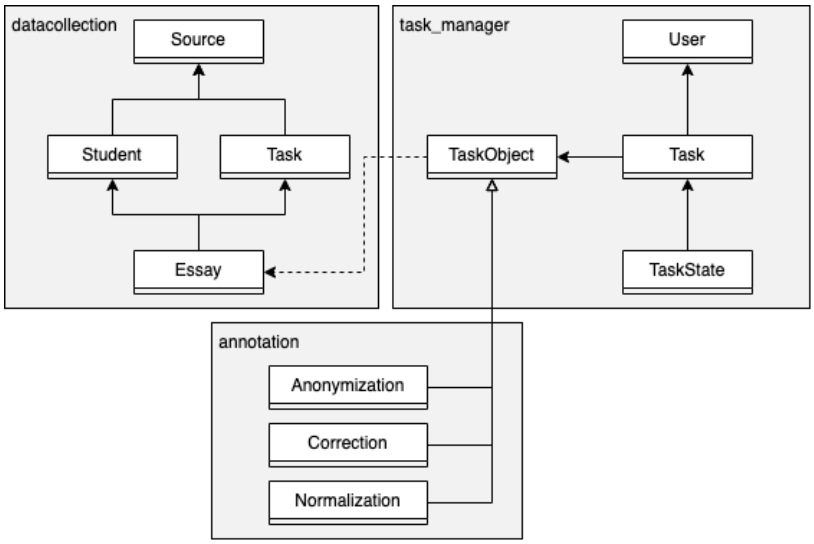


Figure 1: A partial class diagram of the application data model

The *datacollection* module contains the SweLL metadata model:

1. *Source* stores information about the school where the essays have been collected. A *Source* is represented by a school ID, type of education and course type.

2. *Student* stores information about students, including student ID, and structured socio-demographic information about gender, mother tongue(s), education, etc.

3. *Task* stores information about the task learners have received for essay writing, including descriptive information about genre, topic, grading system, allowed time, etc.

4. *Essay* metadata is a record organized as a response by a *Student* to a *Task*, which stores information about the individual performance on an essay. (Essay texts are not stored here, but in *TaskObject* and *TaskState* objects, see below.)

In the user interface, under the *Metadata* tab, it is possible to get an overview of all items in each class (e.g. Figure 2)[2], filter them, open them for editing or add new records (e.g. Figure 3). Access to the portal is password-protected.

3 Annotation task management

The *annotation* module defines the three types of annotation tasks: Anonymization, Normalization and Correction Annotation. Each task type is defined along with workflow control and configuration for the separate annotation tool.

[2]Text in the Figures is predominantly in Swedish

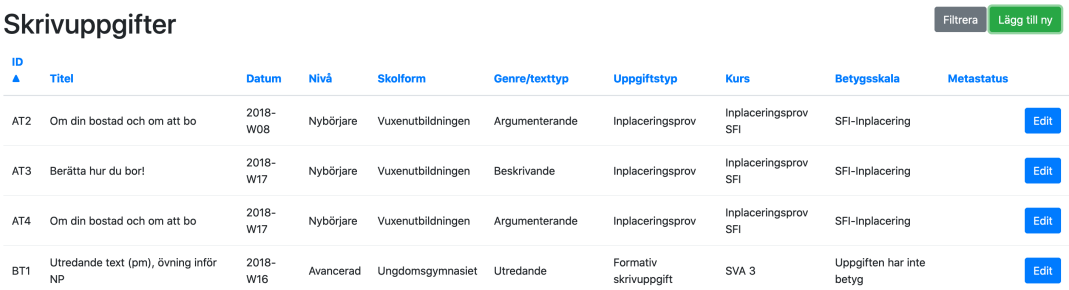


Figure 2: A list of *Task* metadata records

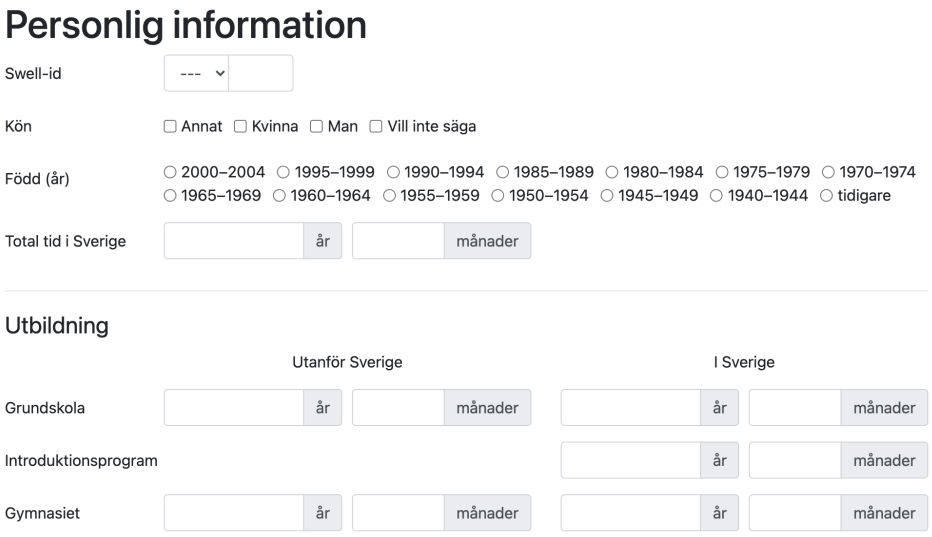


Figure 3: A record for *Student* personal metadata

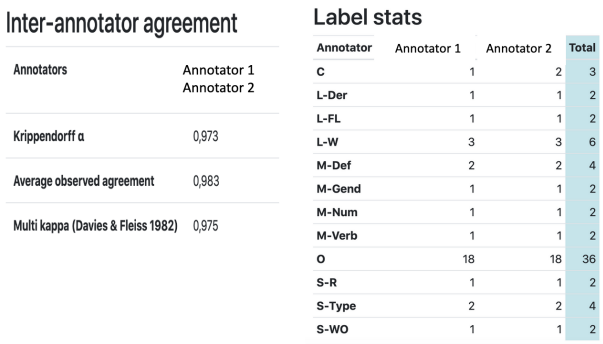
The *task_manager* module, in turn, provides the ability for users to create, assign and work on annotation tasks.

1. A *TaskObject* pairs a task type with a specific essay, e.g. *anonymization of essay A1AT2*. It is implemented with a generic reference to the essay metadata object, ensuring a loose module dependency.
2. A *Task* tracks a specific user's work on a *TaskObject*. If more than one user work on the same annotation task, they each have a separate *Task* with separate progress.
3. Work in the annotation tool generates a sequence of *TaskStates*, each a snapshot version of the text plus annotations.

When an annotation task is started, an external annotation tool SVALA (Wirén et al., 2019)[3] is opened, changes being versioned as *TaskStates* on every introduced change. From the annotation tool, the task can be marked as *Done*, in which case this is reflected in the portal as well.

The functionality of the portal allows to assign the same Correction Annotation task to several users, and to run Inter-Annotator Agreement once at least two versions of annotations are available, see Figure 4 for an example. There is a possibility to click on the EssayID to view the essay and to monitor the progress of its annotation.

[3]SVALA demo version: `https://spraakbanken.gu.se/swell/dev/`



Inter-annotator agreement

Annotators	Annotator 1 Annotator 2
Krippendorff α	0,973
Average observed agreement	0,983
Multi kappa (Davies & Fleiss 1982)	0,975

Label stats

Annotator	Annotator 1	Annotator 2	Total
C	1	2	3
L-Der	1	1	2
L-FL	1	1	2
L-W	3	3	6
M-Def	2	2	4
M-Gend	1	1	2
M-Num	1	1	2
M-Verb	1	1	2
O	18	18	36
S-R	1	1	2
S-Type	2	2	4
S-WO	1	1	2

Figure 4: Inter-annotator agreement for a particular essay and a pair of annotators

4 Statistics

Statistics is used to get an overview of the metadata and its frequencies, as well as frequencies over tokens and sentences. Several views are possible:

1. *Running statistics* (see e.g. in Figure 5), which can be exported in an csv file format by objects described in Section 2

	Average observed agreement	0,962
Korrigeringsannoterade (statistik)	Antal meningar	8481
	Antal meningar/uppsats	17
	Antal meningar med fel	6657
	Antal korrigeringar/uppsatser	53
	Antal korrigeringar/mening	3
Antal tokens (μ - mean, σ - standard deviation)	Anonymiserade uppsatser, källtext	211744 (μ = 317.0, σ = 241.4)
	Normaliserade uppsatser, källtext	148905 (μ = 296.6, σ = 246.8)
	Normaliserade uppsatser, måltext	152518 (μ = 303.8, σ = 250.9)
	Korrigeringsannoterade uppsatser, måltext	152518 (μ = 303.8, σ = 250.9)

Figure 5: Running statistics over all material in the portal (excerpt)

2. *Summary* by filtered values (see Figure 6) can help navigate among the various metadata available for Students, Tasks and Schools. Apart from the table view of the statistics, also graphs are available for inspection of the material.

The Summary view can visualize the progress of the annotation process, which was used for monitoring the progress of the project as a whole. A decision has been made in the project that advanced searches and filters (a lá Excel) should not be implemented. Instead, a possibility is provided to download files with statistics that can be opened using excel or processed through other programs.

5 Data export and import

Finally, the *SweLL portal* offers a possibility to *import* new essays or new subcorpora. Metadata records can either be created manually for those, or generated automatically from xml-headers according to a protocol. For *export*, several data formats are available: xml, json, raw text. The work on documenting the SweLL portal is ongoing with some documentation available from the SweLL project webpage.[4]

[4] `https://spraakbanken.gu.se/en/projects/swell/swell-docs`

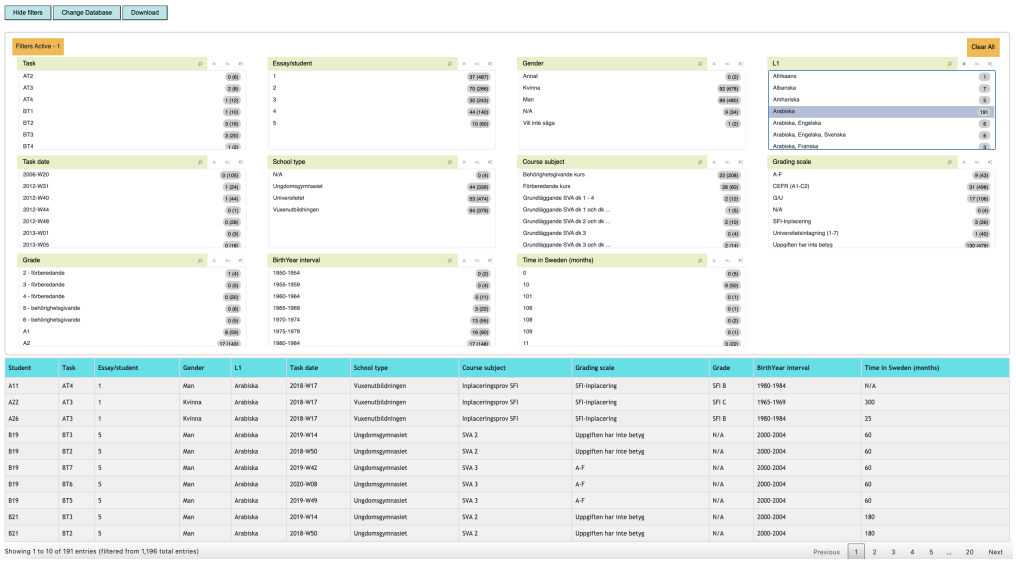


Figure 6: Statistics with a filter function

6 Concluding remarks

In the current project, the aspects of data storage and annotation management have been taken seriously following the arguments outlined in Fort (2016) and Hovy and Lavid (2010) that stable annotation management has been shown to be one of the important prerequisites for creation of well-balanced and reliably annotated corpora. The portal development was incremental, with changes introduced in responce to the needs of the project. Overall, both project researchers and project assistants were facilitated in their work through the *SweLL portal* functionalities. The *SweLL portal* will continue to be used for new learner corpus annotation projects as well as for statistical exploration of the material as a part of a newly developed SweLL infrastructure for second language research.

References

Karën Fort. 2016. *Collaborative Annotation for Reliable Natural Language Processing: Technical and Sociological Aspects*. John Wiley & Sons.

Eduard Hovy and Julia Lavid. 2010. Towards a 'science'of corpus annotation: a new methodological challenge for corpus linguistics. *International journal of translation*, 22(1):13–36.

Alexander König, Jennifer-Carmen Frey, and Egon W Stemle. 2021. Exploring reusability and reproducibility for a research infrastructure for l1 and l2 learner corpora. *Information*, 12(5):199.

Egon W Stemle, Adriane Boyd, Maarten Jansen, Therese Lindström Tiedemann, Nives Mikelić Preradović, Alexandr Rosen, Dan Rosén, Elena Volodina, et al. 2019. Working together towards an ideal infrastructure for language learner corpora. In *Widening the Scope of Learner Corpus Research Selected Papers from the Fourth Learner Corpus Research Conference*. Presses universitaires de Louvain.

Elena Volodina, Beáta Megyesi, Mats Wirén, Lena Granstedt, Julia Prentice, Monica Reichenberg, and Gunlög Sundberg. 2016. A Friend in Need? Research agenda for electronic Second Language infrastructure. In *Proceedings of Swedish Language Technology Conference (SLTC) 2016, Umeå, Sweden*.

Elena Volodina, Maarten Jansen, Egon W. Stemle, Therese Lindström Tiedemann, Nives Mikelić Preradović, Silje Karin Ragnhildstveit, Kari Tenfjord, and Desmedt Koenraad. 2018. Interoperability of second language resources and tools. In *Proceedings of the CLARIN Annual Conference 2018, Pisa, Italy*, page 90–94.

Elena Volodina, Lena Granstedt, Arild Matsson, Beáta Megyesi, Ildikó Pilán, Julia Prentice, Dan Rosén, Lisa Rudebeck, Carl-Johan Schenström, Gunlög Sundberg, and Mats Wirén. 2020. The SweLL Language Learner Corpus: From Design to Annotation. *Northern European Journal of Language Technology, Special Issue*.

Mats Wirén, Arild Matsson, Dan Rosén, and Elena Volodina. 2019. SVALA: Annotation of Second-Language Learner Text Based on Mostly Automatic Alignment of Parallel Corpora. *Proceedings of CLARIN 2018*.

A method for building non-English corpora for abstractive text summarization

Julius Monsen
Computer and Information Science
Linköping University
Linköping, Sweden
julmo634@student.liu.se

Arne Jönsson
Computer and Information Science
Linköping University
Linköping, Sweden
arne.jonsson@liu.se

Abstract

We present a method for building corpora for training, and testing, abstractive text summarizers for languages other than English. The method builds on the widely used English CNN/Daily Mail corpus and the assumption that corpora for other languages can be built by filtering language-specific news corpora to have similar properties as the CNN/Daily Mail corpus. In the paper, we show how to achieve this by removing texts from the target corpus that do not adhere to the characteristics of the CNN/DaiyMail corpus. Models are trained on these filtered subsets of the corpus and compared to results from training a model on the CNN/DaiyMail corpus. The results show that the method can be used to build corpora for training abstractive text summarisers for languages other than English that have properties on par with those trained using the CNN/Daily Mail corpus.

1 Introduction

When building, and assessing, abstractive text summarizers the English CNN/Daily Mail corpus (Nallapati et al., 2016; Hermann et al., 2015) is the currently most used benchmark corpus[1]. For languages other than English the MLSUM corpus (Scialom et al., 2020) is a corpus built on the same principles as the CNN/Daily Mail corpus. From publicly available news articles that corpus was built by filtering out articles with certain properties. This gives a multilingual extension of the CNN/Daily Mail corpus for French, German, Spanish, Russian, and Turkish.

In this paper, we present a method, similar to the one used to build the MLSUM corpus, for building a corpus in Swedish for training abstractive text summarizers. The main difference is that we go further than MLSUM and apply additional filters based on semantic textual similarity and the abstractness of the summaries to even more resemble the CNN/Daily Mail corpus. To assess the method these subsets of our corpus are used to train abstractive text summarizers and the results are compared to results achieved using the CNN/Daily Mail corpus. The corpus will be freely available as a SweCLARIN resource[2].

2 The corpora

The basis for our method is two news articles corpora, one from the Swedish newspaper Dagens Nyheter (DN) and the CNN/Daily Mail corpus. The original DN corpus comprises $1,963,576$ news articles published during the years 2000-2020. We use the preamble as a summary. Unfortunately, the preamble is not always a good summary of an article, which is one of the main problems that the proposed method handles by filtering out those article/summary-pairs that are not useful for building abstractive summarizers. Scialom et al. (2020) filter out articles shorter than 50 words or those with summaries shorter than 10 words. As we intend to apply further

[1] See e.g. https://paperswithcode.com/sota/document-summarization-on-cnn-daily-mail

[2] https://spraakbanken.gu.se/resurser#corpora

filtering mechanisms we filter out article/summary pairs with articles less than 25 words or summaries with less than 10 words. We further filter out pairs with a compression ratio (summary length/article length)[3] higher than 0.4, since very few article/summary-pairs in the CNN/Daily Mail corpus has a compression ratio higher than that.

This initial filtering and removal of duplicates yielded a corpus of 802.405 article/summary-pairs. 42% of these were categorised as domestic news, 22% as sports, 16% as economy, 10% culture and 9% other (consisting of very small categories). To further characterise the corpora we measure compression rate and novelty. Novelty (n-gram) is the fraction of n-grams in the summary that was not in the paired article (Scialom et al., 2020). In calculating novelty stop words were removed and words were stemmed. Semantic similarity was calculated as the cosine similarity between embeddings yielded from Sentence BERT (Reimers and Gurevych, 2019). Both embedding similarity on the document level (doc/doc) and between the complete summary and the most similar sentence in the article (doc/sent) are calculated. The characteristics of this corpus (DN-LC) are presented in Table 1 together with the characteristics for the CNN/Daily Mail corpus.

	DN-LC	CNN/DailyMail
Corpus size	802,405	311,971
Vocabulary size	2,575,130	803,487
Occurring 10+ times	387,370	161,820
Article words	370.41	677.21
Article sentences	24.00	28.52
Summary words	29.31	48.34
Summary sentences	2.19	3.70
Compression ratio	0.13	0.09
Novelty (uni-gram)	0.42	0.14
Novelty (bi-gram)	0.80	0.57
Novelty (tri-gram)	0.93	0.77
Semantic similarity (doc/doc)	0.49	0.65
Semantic similarity (doc/sent)	0.52	0.67

Table 1: Statistics for the different corpora. Corpus size is the number of article/summary pairs. Vocabulary size is the total number of different words in the corpus and Occurring 10+ times, the total number of words occurring 10+ times. Article/Summary words/sentences are the number of words/sentences in the articles/summaries. Compression rate, novelty, and semantic similarity are as presented in the text. All values except Corpus size, Vocabulary size and Occurring 10+ times are mean values across all article/summary-pairs.

As can be seen, the Swedish DN-LC corpus is much larger than the CNN/Daily Mail corpus. It also shows a larger variation in terms of article and summary lengths with significantly shorter articles and summaries. The Swedish corpus also has a much higher novelty. This indicates that it contains a lot of lower-quality article/summary-pairs (alternatively that the summaries are very abstract). In terms of semantic textual similarity, the Swedish corpus contains less semantically similarly article/summary-pairs. This also points to the current problem of low-quality article/summary-pairs.

3 Filtering the Swedish corpus and fine-tuning models for summarization

The goal is to create a corpus with properties similar to the CNN/Daily Mail corpus by filtering out those article/summary pairs that are not semantically similar or concrete enough, based on novelty.

[3]The reason for not measuring compression ratio as article/summary, c.f. Grusky et al. (2020) and Scialom et al. (2020) is that when filtering on compression ratio we prefer a number between 0 and 1.

We build three different corpora that are compared to the characteristics of the CNN/Daily Mail corpus. DN-S was filtered on semantic similarity first using the doc/doc similarity measure then the doc/sent similarity measure. The threshold values for filtering were iteratively adjusted to obtain a similar distribution as the CNN/Daily Mail corpus. DN-N was filtered to have similar distributions as the CNN/Daily Mail corpus with regards to novelty measures based on uni-grams, bi-grams and tri-grams. It turns, however, out that filtering on uni-grams also provides similar bi-gram and tri-gram values as well (probably since they are strongly correlated). DN-SN, finally, used both semantic similarity and novelty in a similar manner.

The results of this filtering are shown in Table 2, with the characteristics of the CNN/Daily Mail corpus as reference. The distribution among news categories was approximately the same as the DN-LC corpus for all subsets, except for DN-N which had 40% domestic news, 18% other, 17% culture, 13% economy, 12% sports.

	DN-S	DN-N	DN-SN	CNN/DailyMail
Corpus size	122,419	124,105	38,151	311,971
Vocabulary size	727,406	1,070,351	435,412	803,487
Occurring 10+ times	118,661	171,100	70,450	161,820
Article words	362.52	630.36	512.43	677.21
Article sentences	22.51	40.27	31.71	28.52
Summary words	32.15	33.16	35.67	48.34
Summary sentences	2.38	2.51	2.62	3.70
Compression ratio	0.13	0.07	0.10	0.09
Novelty (uni-gram)	0.32	0.14	0.14	0.14
Novelty (bi-gram)	0.73	0.57	0.57	0.57
Novelty (tri-gram)	0.89	0.78	0.78	0.77
Semantic similarity (doc/doc)	0.65	0.52	0.65	0.65
Semantic similarity (doc/sent)	0.67	0.60	0.67	0.67

Table 2: Statistics for the filtered corpora.

As can be seen in Table 2 the DN-SN corpus is much smaller than the CNN/Daily Mail corpus but it has the same compression ratio, novelty and semantic similarity.

4 Evaluating the corpus for abstractive summarization model building

We trained Swedish abstractive summarizers on the different corpora using a state-of-the-art approach (Rothe et al., 2020), in which an encoder-decoder model is warm-started with a pre-trained model, in our case a Swedish pre-trained BERT model (Malmsten et al., 2020). The training settings were adapted depending on the size of the corpora, but they were all trained until convergence. These models were then evaluated on two small subsets containing 9000 article/summary-pairs each, that had been picked out before the filtering. One of these subsets (test-SN) had similar semantic similarity and novelty as the CNN/Daily Mail corpus and the other (test-LC) had similar properties as the DN-LC corpus filtered on length and compression ratio. We used ROUGE scores as metrics for evaluating the model generated summaries.

In Table 3 the results for all models trained on the different corpora are presented. As can be seen, the results differ significantly between the different test sets with higher scores on test-SN for all models. Furthermore, the best result is achieved when using the rather small corpus filtered on both semantic similarity and novelty, DN-SN, with ROUGE scores almost on par with those achieved with the CNN/Daily Mail corpus. On test-SN, we also note that larger corpora, DN-S and DN-N, performs slightly worse and that the DN-LC corpus, only filtered by length and compression ratio, c.f. Scialom et al. (2020), which is even larger than the CNN/Daily Mail corpus, performs the worst. On the other hand, the model trained on the larger DN-LC corpus performs best on the test-LC set. All this highlights the importance of having high-quality test

data when evaluating models and that more training data does not necessarily produce a better model, if the data is of lower quality.

	ROUGE-1		ROUGE-2		ROUGE-L	
CNN/DailyMail	39.89		18.18		27.54	
	test-LC	test-SN	test-LC	test-SN	test-LC	test-SN
DN-LC	29.08	35.46	9.67	14.71	20.23	24.71
DN-S	28.44	36.97	8.98	15.79	19.48	25.71
DN-N	27.42	36.96	8.42	16.17	18.74	25.95
DN-SN	26.48	37.22	7.44	16.32	17.84	26.11

Table 3: Evaluation results on test data measured with ROUGE F-scores.

5 Conclusion

We have presented a method for building corpora to be used when training abstractive text summarizers for Swedish. From a comparatively large corpus of Swedish summary/article pairs we use a number of filtering techniques to achieve properties similar to an English state-of-the art corpus in terms of compression ration, semantic similarity, and novelty rather than size. We show that using such a corpus to train a state-of-the art text summarizer gives results almost on par with results using the English corpus, even though the Swedish corpus is much smaller. We believe that the method can be used for other languages to build high-quality corpora that can be as useful as English corpora and extend the CLARIN infrastructure.

Acknowledgements

This research is financed by the Swedish CLARIN node SweCLARIN and the Swedish Research Council. We are indebted to Dagens Nyheter for providing the news articles.

References

Max Grusky, Mor Naaman, and Yoav Artzi. 2020. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of NAACL-HLT 2018*, pages 708–719.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *NIPS'15: Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, Cambridge, MA, USA. MIT Press.

Martin Malmsten, Love Börjeson, and Chris Haffenden. 2020. Playing with words at the national library of Sweden – making a Swedish BERT.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar GuÌ‡lçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany, August. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November. Association for Computational Linguistics.

Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2020. Leveraging Pre-trained Checkpoints for Sequence Generation Tasks. *Transactions of the Association for Computational Linguistics*, 8:264–280.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. MLSUM: The multilingual summarization corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8051–8067, Online, November. Association for Computational Linguistics.

CLARIN Flanders: new prospects

Vincent Vandeghinste
Instituut voor Nederlandse Taal, Netherlands
vincent.vandeghinste@ivdnt.org

Els Lefever
Ghent University, Belgium
els.lefever@ugent.be

Walter Daelemans
University of Antwerp, Belgium
walter.daelemans@uantwerpen.be

Tim Van de Cruys
University of Leuven, Belgium
tim.vandecruys@kuleuven.be

Sally Chambers
Ghent University, Belgium
sally.chambers@ugent.be

Abstract

We describe the creation of CLARIN Belgium (CLARIN-BE) and, associated with that, the plans of the CLARIN-VL consortium within the CLARIAH-VL infrastructure for which funding was secured for the period 2021-2025.

1 Introduction

We describe the efforts that have been undertaken to ensure the re-entry of Flanders, the Dutch-speaking community in Belgium, into the world of the CLARIN ERIC, in section 2. We also describe the new linguistic tools and services that are planned to be developed within the second phase of the CLARIAH-VL project, which has recently started, in section 3.

2 CLARIN-VL and CLARIN-BE

Given that, in Belgium, most funding of scientific research happens at the level of the communities, of which there are three: the *Vlaamse Gemeenschap* (the Flemish Community – Dutch speaking), the *Fédération Wallonie-Bruxelles (FWB)* (the Federation Wallonia-Brussels – French speaking) and the very small *Deutschsprachige Gemeinschaft* (the German-speaking Community), and given that members or observers of CLARIN ERIC have to be countries or intergovernmental organizations,[1] it follows that Flanders cannot be a member of CLARIN directly.

In the past, Flanders participated in CLARIN through the international organization *Nederlandse Taalunie* (Dutch Language Union), but such a construction was no longer possible after 2018, resulting in a *Flexit* from CLARIN.

The only possible way to become a member of CLARIN ERIC was to apply for political support from Flanders (without funding) for the formal founding of CLARIN Belgium and payment of the CLARIN ERIC membership fees by the Belgian Science Organization BELSPO,[2] in a similar construction as the DARIAH infrastructure.[3] Such political support was granted and membership of Belgium should become a fact in 2021.

The Flemish CLARIN consortium consists of several research groups from three Flemish universities, and the *Instituut voor de Nederlandse Taal* (INT – Dutch Language Institute)[4] as third party, and is open for more research groups. INT is located in the Netherlands but is partly funded by Flanders, and is the *de facto* CLARIN-B centre for Flanders, serving as a data depositing centre. Currently different data sets and tools developed in Flanders have been integrated into the CLARIN infrastructure already, and more will follow, see section 3. CLARIN-VL also focuses on user involvement, through the organisation of CLARIN information sessions and lectures during different courses.

[1] cf. https://www.clarin.eu/content/participating-consortia
[2] https://www.belspo.be
[3] https://be.dariah.eu/
[4] http://www.ivdnt.org

In the meantime, efforts are undertaken to involve users from the French-speaking community in Belgium to take part actively in CLARIN Belgium, and we expect contributions to CLARIN from several of these research groups once Belgian membership has been formally established.

3 CLARIAH-VL

CLARIN-VL decided to join forces with DARIAH Flanders in the FWO International Research Infrastructure project *CLARIAH-VL*, and has secured funding until early 2025, for the development of several infrastructural services. This section presents our plans and ongoing work, so that CLARIN users know what to expect, and other CLARIN members know what we are working on, in order to promote cooperation and avoid parallel development of similar tools and resources.

We aim to develop a *Digital Text Analysis Dashboard and Pipeline* for processing both Dutch texts and parallel texts. There are already several pipeline approaches available (Bel, 2010; Hinrichs et al., 2010; Zinn, 2018; van der Sloot et al., 2018), but these are often limited to linguistic analysis — tokenization, pos-tagging, lemmatization, named entity labeling, dependency parsing. We aim at an approach which re-uses existing (CLARIN) tools and pipelines, but is extended with a variation of models and several natural language understanding analysis layers.

The user-friendliness of the design will be ensured through a user-centred approach involving also non NLP-users. To this end, a list of dedicated use cases will be defined from both the NLP and digital humanities research communities in Flanders and will be worked out in detail. Users will also be consulted for personalization of the dashboard and adapting it to their own needs.

An example of such a use case could be the Spoken Academic Belgian Dutch corpus, which is currently under development and which needs to be speech recognized, manually corrected and (automatically) linguistically annotated. Another use case is the processing of parallel data, with sentence and word alignment tools and extended corpus search functions to allow searching in parallel data with Blacklab (de Does et al., 2017), a cooperation with CLARIAH-NL, and with an example-based query engine similar to PolyGrETEL (Augustinus et al., 2016). A final digital humanities use case could be pipelines for the text and data mining of sub-corpora of digitised newspapers from KBR, the Royal Library of Belgium's BelgicaPress.[5]

For a number of NLP tasks, different alternatives are available in different forms and programming languages. We will benchmark existing tools and new models. The results allow users to curate which alternatives to integrate in the pipeline. This includes (re)training tools on existing resources, such as creating state-of-the art methods for language modelling of historical Dutch or specialized text corpora (e.g., medical text, legal text, etc).

A first set of tools that will be benchmarked are **linguistic processing tools**, which enrich corpora in plain text format with linguistic information, such as part-of-speech tags, lemmas (basic form as found in a dictionary), named entity information and chunk information. Existing and new implementations for English, French, Dutch and German will be tested. Amongst these we will test LeTs Preprocess: the Multilingual LT3 Linguistic Preprocessing Toolkit (Van de Kauter et al., 2013), which provides linguistic annotation for 4 languages (English, Dutch, French, German): tokenisation, part-of-speech tagging, lemmatisation, chunking and named entity recognition. Other tools that will be evaluated are the memory-based linguistic analysis tool Frog (van der Sloot et al., 2018), and its successor in the deep learning paradigm DeepFrog[6], which was developed in CLARIAH-NL, and which provides neural network models for Dutch NLP, part-of-speech and named entity tagging.

The previous two pipelines will be compared to other state-of-the-art open source libraries, such as for instance the Spacy libraries for text processing[7], or the Stanza Python NLP package[8].

Secondly, we focus on the benchmarking of NLP tools for **natural language understanding**. Recent machine learning methods based on neural transformer architectures have greatly improved the state of

[5] `https://www.belgicapress.be`
[6] `https://github.com/proycon/deepfrog`
[7] `https://spacy.io/usage/linguistic-features`
[8] `https://stanfordnlp.github.io/stanza/`

the art for a wide range of natural language understanding (NLU) tasks. In order to provide an extensive testbed for language-specific NLU models, we will develop a suite of NLU evaluation tasks, similar to the well-known GLUE (Wang et al., 2018) and SuperGlue (Wang et al., 2019) evaluation frameworks for English. Specifically, we will focus on an evaluation suite for Dutch. Due to the laborious nature of manual labeling, we will mainly focus on the semi-automatic construction of evaluation tasks (construction of datasets from web forums and resources, prediction of discourse markers, ...), as well as the compilation of existing evaluation sets within one overarching suite. In a second stage, the construction may be complemented by manually labeled evaluation instances, gathered by means of a voluntary crowdsourcing setup.

Additionally, we will explore the application of neural network models for the **search and extraction of linguistic structures**. There is corroborating evidence that self-supervised transformer architectures implicitly encode a wide range of linguistic knowledge, from part of speech information over syntactic structure to co-reference information (Peters et al., 2018; Hewitt and Manning, 2019; Clark et al., 2019). We will investigate to what extent such implicit linguistic representations might be exploited as a tool for linguistic analysis. More specifically we will investigate whether the linguistic information present in the models might be distilled for the purpose of similarity computations. Such a process would allow to automatically harvest a corpus of linguistically similar structures, in order to support linguistic analysis. Moreover, as transformer architectures simultaneously encode syntactic and semantic information in their contextualized representations, this would allow to automatically harvest syntactically disparate realization of similar semantic content, providing an adequate means for a linguistic analysis of the syntax-semantics interface.

The second batch of tools that will be integrated in the pipeline are tools aiming at solving specific natural language processing and understanding tasks, amongst which:

- **Sentiment analysis**: a pipeline annotating text strings of varying length (e.g., words, chunks, sentences, reviews, documents, etc.) with polarity information (positive, negative, neutral) and unsupervised learning techniques for adapting dictionary-based sentiment analysis tools to new domains.
- **Emotion detection**: a pipeline annotating text strings with more fine-grained emotion information. Two types of emotion detection approaches will be evaluated: (1) classification approaches providing categorical labels (e.g. anger, disgust, fear, joy, sadness, surprise), and (2) dimensional models representing emotions as vectors in a multidimensional space, defined by three axes: valence (unhappiness/happiness), arousal (calmness/excitement) and dominance (submission/dominance). Every emotional state is then described by the combination of the values on these three axes.
- **Document similarity clustering**: a pipeline which allows uploading a set of documents and provides document clusters based on their similarity according to different models.
- **Topic modelling**: a pipeline extracting topics from text corpora using traditional (LDA, NMF) and more recent (top2vec) embedding based approaches. Visualisation in terms of topic maps and timelines.
- **Stylometry**: a pipeline for unsupervised and supervised machine learning based stylometry (authorship attribution, age, gender personality profiling) allowing the combination of various linguistic and stylometric information sources and adding new information sources, e.g., figurative language detection.

The third batch of tools that will be integrated are tools aiming at processing multilingual data:

- **Sentence alignment**: integration of a tool that 'aligns' (makes relations explicit) between the sentences of two texts that are literal translations.
- **Word alignment**: integration of a tool that identifies relationships among the words in a bitext, 'aligning' words that are translations of one another. Word alignment typically starts from pairs of sentences that have been sentence-aligned before.

Public datasets that have been processed with certain tools will be made available from within the dashboard, allowing users to search within these datasets.

Finally, a Help Desk will be developed to help users by advising users and tailoring tools to specific use cases or domains, as well as deal with feedback on annotation and analysis errors, leading to improved models. This help desk can be contacted through `servicedesk@ivdnt.org`. This Help Desk will provide information similarly to K-Dutch, the CLARIN Knowledge Centre for Dutch, which has been recognized by CLARIN-ERIC this summer.[9]

4 Conclusions

We are very pleased to announce the re-entry of Flanders in CLARIN through Belgian membership, and have presented our work plan for the next funding period.

References

L. Augustinus, V. Vandeghinste, and T. Vanallemeersch. 2016. Poly-GrETEL: Cross-lingual example-based querying of syntactic constructions. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3549–3554, Portorož, Slovenia, May. European Language Resources Association (ELRA).

N. Bel. 2010. Platform for Automatic, Normalized Annotation and Cost-Effective Acquisition of Language Resources for Human Language Technologies. Panacea. *Procesamiento del Lenguaje Natural*, 45:327–328.

K. Clark, U. Khandelwal, O. Levy, and C.D. Manning. 2019. What does bert look at? an analysis of bert's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286.

J. de Does, J. Niestadt, and K. Depuydt. 2017. Creating Research Environments with BlackLab. In Jan Odijk and Arjan van Hessen, editors, *CLARIN in the Low Countries*, chapter 20. Ubiquity Press, London, Dec.

J. Hewitt and C.D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota, June. Association for Computational Linguistics.

E.W. Hinrichs, M. Hinrichs, and T. Zastrow. 2010. WebLicht: Web-Based LRT Services for German. In *Proceedings of the ACL 2010 System Demonstrations*, pages 25–29.

M. Peters, M. Neumann, L. Zettlemoyer, and W. Yih. 2018. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509.

M. Van de Kauter, G. Coorman, E. Lefever, B. Desmet, L. Macken, and V. Hoste. 2013. LeTs Preprocess: the Multilingual LT3 Linguistic Preprocessing Toolkit. *Computational Linguistics in the Netherlands Journal*, pages 103–120.

K. van der Sloot, I. Hendrickx, M. van Gompel, A. van den Bosch, and W. Daelemans. 2018. Frog, a natural language processing suite for dutch. Reference Guide. Technical Report Language and Speech Technology Technical Report Series 18-02, Radboud University.

A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium, November. Association for Computational Linguistics.

A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

C. Zinn. 2018. Squib: The language resource switchboard. *Computational Linguistics*, 44(4):631–639, December.

[9] `https://kdutch.ivdnt.org/`

Reliability of automatic linguistic annotation: native vs non-native texts

Elena Volodina, David Alfter
University of Gothenburg, Sweden
name.surname@gu.se

Therese Lindström Tiedemann, Maisa Lauriala, Daniela Piipponen
University of Helsinki, Finland
name.surname@helsinki.fi

Abstract

We summarize the results of a manual evaluation of the performance of automatic annotation on three different datasets: (1) texts written by native speakers, (2) essays written by second language (L2) learners of Swedish in the original form and (3) the normalized versions of the same essays. The focus of the evaluation is on lemmatization, PoS-tagging, dependency annotation, word sense disambiguation and multi-word detection.

1 Introduction

In the current project, *Development of grammatical and lexical competences in immigrant Swedish*,[1] we explore profiling of lexical and grammatical competences among second language (L2) learners of Swedish based on two corpora. The course-book corpus, COCTAILL, and the L2 Swedish learner corpus, SweLL-pilot, (introduced further below), are used for qualitative and quantitative analysis of lexical and grammatical categories that L2 learners are exposed to or produce themselves. The texts in the two corpora have been automatically annotated with linguistic information using the Sparv-pipeline (Borin et al., 2016) which is an essential part of the CLARIN infrastructure for the Swedish language. Sparv, in turn, relies on the annotation standards from the Stockholm Umeå Corpus (SUC) (Ejerhed et al., 1997) and on the theoretical framework in the Saldo lexicon (Borin et al., 2013). Since the process of linguistic annotation is performed automatically, we need to evaluate to which degree we can expect the results of the annotation to be reliable, so that our theoretical generalizations and conclusions about language learning can factor that in. For this reason, we have performed a manual "annotation quality check" of Part-of-Speech (PoS) tagging, lemmatization, dependency annotation, identification of multi-word expressions (MWE) and word sense disambiguation (WSD) which we report in this paper.

Previous work suggests that the performance of automatic pipelines trained on native language models is non-optimal on L2 language due to a large number of non-words, deviating syntactic patterns and statistical distributions in L2 production (Štindlová et al., 2012). Thus, we need to assess to which degree we can trust the conclusions based on the Sparv-annotated L2 data, and evaluate whether there is a need to develop a specific pipeline (or modular additions to the standard pipeline) for L2 language varieties. This experiment complements and extends several investigations of the Sparv pipeline where Sparv performance was *automatically* evaluated and only for native language (L1) varieties (Berdicevskis, 2020a; Berdicevskis, 2020b), whereas we examine the reliability of annotations *manually* and on several types of language – L1, L2 original and L2 corrected.

Despite Swedish being the focus of this experiment, we expect our findings to be generalizable to other languages and to the performance of other pipelines on non-native language samples. We also believe it is an important study for CLARIN in that it evaluates how well part of the CLARIN infrastructure works for both L1 and L2 Swedish, hence assessing the need for improvements to the current pipeline for Swedish.

[1] Riksbankens jubileumsfond P17-0716:1, project homepage: <https://spraakbanken.gu.se/en/projects/l2profiles>

2 Experiment setup

Sparv pipeline The Sparv pipeline v.3.0[2] (Borin et al., 2016), analyzed by us, consists of several modules, sequentially applied to the input Swedish data. For *lemmatization*, the Saldo lexicon (Borin et al., 2013) returns lemgrams including potential MWEs and a list of associated senses. *Senses* are disambiguated using an algorithm developed by Nieto Piña (2019). For *PoS tagging*, Sparv uses HunPos (Halácsy et al., 2007) trained on the SUC 3.0 corpus (Ejerhed et al., 1997). For *syntactic annotation*, the MaltParser (Nivre et al., 2007) is used, trained on the Swedish Talbanken (Nilsson et al., 2005).

Problem definition We perform a manual evaluation of the Sparv-pipeline (v.3.0) for Swedish in three different contexts – native speaker texts (L1 Coctaill), original second language texts (L2 orig) and normalized essays written by L2 Swedish speakers (L2 norm) – with a focus on lemmatization, word sense disambiguation, MWE identification, PoS-tagging and syntactic dependency annotation. There are two hypotheses which we explore:
1. Pipelines trained on a standard language (L1) do not perform as well on non-standard language varieties such as learner language (e.g. L2 learner production).
2. Normalization of non-standard language, e.g. through error correction, improves tool performance.

Datasets We have selected 15 texts per language variety which we are interested in – native language used in L2 Swedish course books (*L1 Coctaill*), L2 essays (*L2 orig*) and corrected L2 essays (*L2 norm*) – so that they represent five levels of proficiency with three texts per level for each dataset. The levels are defined in accordance with CEFR, the Common European Framework of Reference (Council of Europe, 2001), from A1 (beginner) to C1 (C2 is not represented). Care was taken to select texts of different genres and topics. Only texts containing at least one MWE according to the Sparv annotation were selected.

L1 Coctaill is an L1 dataset that contains 2 190 tokens (punctuation included) per 15 texts and comprises various genres (narrations, facts, evaluation, dialogues, letters, poems) and different topical domains (traveling, languages, culture and traditions, relations with other people, etc.). It is based on *COCTAILL* – a corpus of course books (Volodina et al., 2014), where each chapter/lesson has been marked for the level of proficiency at which it could be used in the teaching of L2 Swedish.

L2 orig is a dataset which contains 4 012 tokens per 15 essays covering several genres (narration, evaluation, argumentation, etc.) and topical domains (personal identification, daily life, travel, house and home, culture and traditions, etc.). *L2 orig* is a subset of the *SweLL-pilot* – a corpus of learner-written essays (Volodina et al., 2016) also marked with CEFR levels.

L2 norm is a dataset containing 3 955 tokens per 15 essays and consists of the same essays as in *L2 orig*, but normalized for errors and deviations to reflect the current norms of the target language. The normalization was done by an L1 Swedish speaker following the guidelines from the SweLL-project.[3]

Design Two linguistically trained assistants, one a L1 Swedish speaker and one an advanced L2 Swedish speaker (L1 Finnish), manually analyzed the automatic tags of the three datasets, introducing corrections where necessary. Assistants were equiped with guidelines[4] and performed the check using excel files. On completion of the check, we[5] analyzed the results using accuracy, F-score and LAS measures (averaged over the two annotators for all tasks except syntactic parsing/DepRel annotation), and inspected the Inter-Annotator Agreement.

[2]The recently released version of Sparv (4.0.0) uses Stanza for PoS and dependency tagging, reporting improved accuracy.

[3]The language of the L2 essays has been made more like standard L1 Swedish through a process of "normalization" where an L1 Swedish speaker has "corrected" them, e.g. correcting spelling mistakes. The target hypothesis produced by the L1 speaker has followed the guidelines for normalization used within the SweLL project, see `https://spraakbanken.github.io/swell-project/Normalization_guidelines`

[4]`https://docs.google.com/document/d/1W9gcwRwFJ7-DsAC6cf6BHUoEivt73r-XWCV1oKS6xV8/edit?ts=5f3518d7#`

[5]The authors of this paper are the two annotators and three researchers. As stated above the L1s of the annotators are Swedish and Finnish respectively. Both have master degrees in the Swedish language. The L1s of the researchers are Russian, Luxembourgish and Swedish.

3 Results

Inter-annotator agreement To put the reported results into perspective, we calculated inter-annotator agreement (IAA) for the two annotators using Krippendorff's alpha, see Table 1. IAA is calculated on a token basis, and we count only whether a change has been made to the original annotation or not. We can see that for some of the linguistic features the reliability is relatively low, notably for PoS tagging of *L2 norm*. This is especially consistent in that one annotator corrected the suggested PoS-tag to the word class "determiner" (DT) several times, whereas the other rarely did. Inspection of the annotations further reveals that one of the annotators is stricter and in this respect deviates from the second annotator. However, the intersection of corrections introduced by both is high. Furthermore, we would like to note that the accuracy is high in all datasets. To increase our understandning of the low IAA and why it was lowest for the normalized dataset, analyses of the disagreements per linguistic feature, proficiency level and word class are necessary but they need to be left outside this abstract. There is no sign that disagreement is due to the L1 of the annotator, but appears more related to the differences in how PoS are defined and used in different frameworks of grammar.

Corpus	Lemma	PoS	MWE	Sense
L1 Coctaill	0.70	0.55	0.85	0.58
L2 orig	0.71	0.56	0.74	0.59
L2 norm	0.66	0.46	0.89	0.66

Table 1: Krippendorff's alpha for interannotator agreement per dataset and ling. category

Corpus	Lemma	PoS	DepRel
L1 Coctaill	0.93±0.0	0.98±0.0	74.89
L2 orig	0.90±0.02	0.95±0.0	63.01
L2 norm	0.93±0.02	0.97±0.0	69.02

Table 2: Lemmatization and PoS tagging: accuracy and standard deviation; Dependency: LAS

Automatic lemmatization, PoS-tagging and dependency annotation We present accuracy of the automatic annotation (i.e. how often the two annotators have corrected automatically assigned tags)[6] in Table 2 for lemmatization and PoS tagging. Dependency annotation is evaluated using micro-averaged (i.e. token-based) Labeled Attachment Score (LAS).[7]

We see that our general assumptions are confirmed: the performance of the automatic annotation on learner essays (*L2 orig*) has lower accuracy than on native (*L1 Coctaill*) or standardized (*L2 norm*) texts, even though only marginally for lemmatization and PoS tagging. This echoes the results obtained in the automatic evaluation of the Sparv PoS-tagging on in-domain L1 texts versus out-of-domain Internet texts (accuracy 0.98 vs 0.93) (Berdicevskis, 2020b). Dependency annotation shows less reliable results for native language, with a preserved tendency of quality loss on *L2 orig*. These results are in line with previous results reporting a LAS score of 78.39 on L1 text (Berdicevskis, 2020a) in automatic evaluation of dependency-relation annotation with Sparv (v.3.0). More detailed (qualitative) analysis of error types by the Sparv pipeline is outside this abstract. We can, however, summarize that automatic lemmatization and PoS tagging are reliable enough to base further generalizations about learner language on them.

Automatic detection of MWEs The purpose of the MWE detection check was to find out, for each text, whether MWEs: (1) were correctly identified; (2) failed to be identified; (3) were incompletely identified; or (4) were incorrectly identified. Table 3 shows precision, recall and F1 score. These values are calculated relative to the number of automatically and manually identified MWEs ($\approx$ the total correct number of MWEs) and not on a token basis.

It seems that the *L2 orig* dataset contains fewer MWEs (or attempts at MWEs) and those that are used are simpler in nature and therefore easier to detect automatically than in *L1 Coctaill* or *L2 norm*. The results, in general, indicate that we can expect that out of 10 MWEs, 7 are correctly captured, 2–3 are missed and a small percentage of noise is introduced in the form of suggestions of MWEs that are not actually in the text or incomplete MWEs. In nine of the missed cases (45% of the missed cases) an MWE entry is missing in the Saldo lexicon as well. However, there are also cases where the MWEs did exist in

[6] Results per various levels of learner essays are omitted in the abstract for need of space.

[7] Note that the dependency annotation was only checked by one assistant, while the rest of the annotation was checked by two.

	#tokens excl punct	MWE Precision	MWE Recall	MWE F1	WSD Accuracy (corr/tot)
L1 Coctaill	1900	0.80	0.71	0.75	0.84±0.03
L2 orig	3635	0.90	0.72	0.80	0.82±0.07
L2 norm	3565	0.85	0.78	0.81	0.83±0.04

Table 3: Number of correctly identified MWEs including precision, recall and F1 score and automatic word sense disambiguation (WSD)

Saldo but were still missed. All in all, results of this evaluation suggest that we can trust the automatic MWE identification, even though we need to be aware of possible misses.

Automatic word sense disambiguation (WSD) The goal of this check was to find out how often: (1) sense was correctly identified; (2) no sense was assigned at all; (3) a lemgram for the correct sense was missing in Saldo; and (4) the correct sense was missing in Saldo.

Table 3 shows the results of the WSD annotation checks. In the three datasets the accuracy of sense disambiguation is high, with very slight fluctuations between the datasets. Counter to our expectations, we do not see any radical improvement in performance following normalization of L2 errors in essays. On inspection, we could see that some senses are missing in Saldo; sometimes even lemgrams are missing. Most challenging are function words, like *som, mången, än* (Eng. as, much, yet), that have very few (sense-based) entries in Saldo, and often in combination with a PoS that does not match PoS tagging based on SUC, which then leads to mismatches and failure on the word sense disambiguation task.

Checking the quality of the automatic word sense disambiguation on a small subset of our data has shown that we can expect that in 80–90 percent the word sense is correctly assigned. Despite the fact that the word sense disambiguation in Sparv is not bullet-proof, we consider it reliable enough to build our vocabulary resource (L2 lexical profile) on the sense level using lemma+PoS+sense as our main entry.

4 Conclusions

Further analysis of our results is required to reach a deeper understanding of the effect of annotation on different types of texts, for different linguistic features and for different levels of proficiency. Comparing a non-standard text to a standard text is complicated, and such an evaluation is affected by the type of texts which are used in the evaluation, including levels of text complexity and levels of proficiency of essay writers. One complication in evaluating learner texts is that mistakes can be on many different levels. A word might have been used in the wrong context but annotated correctly taken out of context based on the morphological principles, disregarding semantic and syntactic principles. If a learner writes, for instance, `Vi gick *snabb` (lit. *We walked rapid.*[8]), Sparv would annotate `snabb` as an adjective, even though syntax and the meaning of the sentence would indicate that it is an adverb. If we normalize it to `Vi gick snabbt` (*We walked rapidly.*) Sparv will instead annotate it as an adverb, following the morphological clue, suffix -t, in connection with syntax. How should we then evaluate the annotation of `snabb` as an adjective in the L2 original text – as a correct or as an incorrect PoS tag? What is right or wrong will depend on the aim of the evaluation. The best solution would be to do two–three annotation checks – one based on the morphological form of lemmas and one including semantic and/or syntactic context – and to specify in guidelines how this should be done. In addition, this also emphasizes the need to know something about how the PoS-tagger works, what it takes as its primary input (morphology or syntax). In the current experiment and check, annotators were told to evaluate the PoS-tagging based on the use of the word in a specific context. as well as this latter issue.

Despite the challenges and varying results per linguistic features, we find that the hypotheses that we started with have been generally confirmed:

[8]Inspired by an English example originally described by Alexandr Rosen (CzeSL project) in a private conversation.

1. Pipelines trained on standard language do not perform equally well on non-standard deviating language. We have seen, though, that the performance drop is different for different linguistic features, and in certain cases relatively negligible (e.g lemmatization and PoS tagging). In certain cases, notably for MWE identification, the tendency is contrary to our expectations, with higher precision and F1 scores on learner language. The nature of that behaviour is left for analysis outside this abstract.

2. Normalization of the learner language improves the performance of the automatic pipeline for all linguistic features, even though sometimes very marginally. We need to go deeper into the analysis to inspect the influence of the level of proficiency on the pipeline accuracy.

All in all, the results of our evaluation are very encouraging, especially with regards to lemmatization, PoS tagging, and word sense disambiguation. MWE identification seems to be a more challenging task. For instance, one annotator viewed some MWEs as incomplete when missing the preposition, e.g. *få reda (på)*, whereas the other did not. The latter view is consistent with the theoretical approach within Saldo, which sees it not as part of the MWE but rather as part of the valency of the MWE in the same way as a preposition can be part of the valency of a single word. This means that IAA would also probably be improved if guidelines more strictly defined what is to be included in a MWE. Further, we have strong indications that automatic dependency relation annotation is unreliable, which can influence our chances to use those annotations as a basis for grammatical profiling of L2 Swedish. However, the new version of the Sparv pipeline may perform reliably enough for our purposes which is left for future assessment.

Acknowledgements

This work has been supported by a research grant from the Swedish Riksbankens Jubileumsfond (Development of lexical and grammatical competences in immigrant Swedish, P17-0716:1) and Språkbanken Text. We also wish to thank the anonymous reviewers for their valuable comments on a previous version.

References

Aleksandrs Berdicevskis. 2020a. Choosing a new dependency parser for Sparv. Technical report, University of Gothenburg, Department of Swedish, 2020-06-03.

Aleksandrs Berdicevskis. 2020b. Choosing a new POS-tagger for Sparv: Update. Technical report, University of Gothenburg, Department of Swedish, 2020-05-12.

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

Lars Borin, Markus Forsberg, Martin Hammarstedt, Dan Rosén, Roland Schäfer, and Anne Schumacher. 2016. Sparv: Språkbanken's corpus annotation pipeline infrastructure. In *Proceedings of Swedish Language Technology Cconference (SLTC)*. Umeå University.

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

Eva Ejerhed, Gunnel Källgren, and Benny Brodda. 1997. Stockholm Umeå Corpus version 1.0, SUC 1.0. *Department of Linguistics, Umeå University*.

Péter Halácsy, András Kornai, and Csaba Oravecz. 2007. HunPos-an open source trigram tagger. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, United States. Association for Computational Linguistics.

Luis Nieto Piña. 2019. *Splitting rocks: Learning word sense representations from corpora and lexica*. PhD Thesis, Data Linguistica 30.

Jens Nilsson, Johan Hall, and Joakim Nivre. 2005. MAMBA meets TIGER: Reconstructing a Swedish treebank from antiquity. *Proceedings from the special session on treebanks at NoDaLiDa 2005*, pages 119–132.

Joakim Nivre, J. Hall, J. Nilsson, A. Chanev, G. Eryigit, S. Kübler, S. Marinov, and E. Marsi. 2007. Maltparser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2):95.

Barbora Štindlová, Alexandr Rosen, Jirka Hana, and Svatava Škodová. 2012. CzeSL–an error tagged corpus of Czech as a second language. In *Corpus data across languages and disciplines*. Peter Lang.

Elena Volodina, Ildikó Pilán, Stian Rødven Eide, and Hannes Heidarsson. 2014. You get what you annotate: a pedagogically annotated corpus of coursebooks for Swedish as a Second Language. In *Proceedings of the 3rd Workshop on NLP for Computer-Assisted Language Learning*. Linköping University Press.

Elena Volodina, Ildikó Pilán, I. Enström, L. Llozhi, P. Lundkvist, G. Sundberg, and M. Sandell. 2016. SweLL on the rise: Swedish learner language corpus for European reference level studies. In *Proceedings of the 10th Language Resources and Evaluation Conference (LREC)*, Portorož, Slovenia.

Seamless Integration of Continuous Quality Control and Research Data Management for Indigenous Language Resources

Anne Ferger
University of Paderborn, Germany
anne.ferger@upb.de

Daniel Jettka
University of Paderborn, Germany
daniel.jettka@upb.de

Abstract

This paper reports on further substantial developments of the continuous quality control framework proposed by Hedeland and Ferger (2020) for assuring and enhancing the quality of linguistic research data, esp. for indigenous language resources in the project INEL (Arkhipov and Däbritz, 2018). The focus lies on the seamless integration of continuous quality control into data creation workflows as well as the induction and improvement of automated monitoring, reporting, and documentation mechanisms. Best practices as well as enhanced and new open access tools for projects intending to optimize their research data management are provided.

1 Introduction

This paper describes the implementation and further enhancement of the continuous quality control framework proposed in Hedeland and Ferger (2020). The developments are based on conceptual and practical experiences made in the project INEL[1] ("Grammatical Descriptions, Corpora and Language Technology for Indigenous Northern Eurasian Languages") which is part of the CLARIN Knowledge-Centre for linguistic diversity and language documentation (CKLD).

Collaborative work on (linguistic) corpora following high-quality standards demands a great deal of administrative and technical organization. Aspects and inter-dependencies of various components of Research Data Management (RDM) come into play, e.g. concerning the used Version Control System (VCS), Continuous Quality Control (CQC) mechanisms, and Project Management System (PMS).

For linguistic corpora, which contain multiple manually compiled transcription and translation layers, annotations, and complex metadata, considerations regarding the quality and consistency of the inter-dependent information play a crucial role in creating scientifically relevant, long-lasting resources. Especially in the context of endangered and less-resourced languages, which come with a very restricted amount of source material, the production of high-quality resources following up-to-date research data standards has to be a central goal of the research process.

The enhancements to the existing framework are designed and intended to be used by further research projects in the spirit of Open Access and Open Science. The related tools Corpus Services (Ferger et al., 2020) and LAMA (Ferger and Jettka, 2021) are published according to Open Access standards.

After laying out the initial situation and the special nature of the data used and created in the project INEL, the specifics of the further developments regarding research data management and quality assurance will be presented.

2 Continuous Quality Control revisited

The continuous quality control framework proposed in Hedeland and Ferger (2020) was already implemented in the project. The creation and publication process of research data in the initial phase relied on a solution based on a standard VCS.[2] A central data repository was installed on a virtual machine,

[1] https://inel.corpora.uni-hamburg.de/

[2] Git: https://git-scm.com/

working directories were implemented on client machines and the clients were free to establish connections via SSH or through mapping the bare repository to a network drive on their local machines and this way simulate a local connection. While the connection via network drive on the one hand made the support for problems in the individual working directories much faster (direct access by technical team), on the other hand it proved to be much slower regarding load times which became ever more evident with growing amounts of research data.

The VCS was integrated with a PMS[3] which allowed for browsing the contents of the central repository, making them generally more accessible. Furthermore, it opened up the possibility to connect tickets in the PMS with data revisions in the VCS and vice versa, and thus facilitate actual enhancements in the collaboration process.

The minimally extended standard VCS solution described in Hedeland and Ferger (2020) poses a number of practical challenges. Firstly, for some corpus creators learning and using VCS procedures originating from software development is too far outside of the scope of their research focus. There is a proposal to overcome this issue by introducing a "silent Git workflow" (Hedeland and Ferger, 2020, p. 9) which, however, results in a striking rise of necessary human supervision and maintenance. Secondly, by using standard VCS mechanisms and interfaces, the unrestricted permission for everyone to perform any action in the VCS will eventually lead to major maintenance cases and poses the imminent risk of data loss. Attempts to pre-filter the actually performed actions by providing precise instructions on what should be done in the VCS unfortunately cannot prevent all problems and risks. Lastly, the standard VCS approach has shortcomings in dealing with (large) binary files and semi-structured files (e.g. XML), which leads to more necessary maintenance the larger the amount and size of files grows. While keeping (static) binary files separately from the VCS (Hedeland and Ferger, 2020, p. 9) solves severe practical issues, it also leads to parallel storage structures, which can be avoided with the advancement of technical infrastructure, i.e. a VCS component that has support for large file handling (e.g. Git LFS).

The CQC framework consisting of a code base with consistent implementation of data checks and fixes described in (Hedeland and Ferger, 2020, p. 10) was applied manually and sporadically whenever needed. Further functionality was needed to achieve a higher and measurable level of data quality and to simplify the publication of this high-quality data. Thus further automation of the quality control operations was needed to accomplish this goal.

3 Advances in the Integration of CQC in linguistic RDM

Since the workflows of the data creation were already in place and using new software (e.g. Git GUI for Windows) yielded various problems, no change of data creation workflows and no additional work for the linguists working on data creation was desirable, obtaining a seamless integration of the existing CQC into the data creation and VCS workflows. The work on the extensions continued over some time, a preceding concept was presented at the CLARIN Bazaar 2019 by (Ferger et al., 2019).

3.1 Managed Version Control with LAMA

With the VCS challenges outlined in section 2 in mind and without other applicable alternatives available the Linguistic Automation Management Assistant – LAMA[4] (Ferger and Jettka, 2021) was created.[5] It is based on a simplistic design, completely text-based, follows a minimal-dependencies approach and facilitates the easy user interaction with a potentially quite complex VCS.[6] There have been first approaches to implement a GUI for LAMA, but it became apparent that a GUI may not necessarily solve any of the basic problems in version control, or differently stated: "Graphical user interfaces (GUIs) have made software user-friendly and arguably fostered the popularization of software in general. However,

[3] Redmine: `https://www.redmine.org/`

[4] `https://doi.org/10.5281/zenodo.4725651`

[5] For an overview of graphical and programming interfaces for Git see `https://git.wiki.kernel.org/index.php/InterfacesFrontendsAndTools`

[6] While LAMA currently can only be used with an underlying Git installation, it could supposedly also be extended to be used on top of other VCSs

scientific software might require alternatives that, if not more intuitive, are more appropriate and efficient depending on the user's needs" (Queiroz et al., 2017).

LAMA basically consists of a script that provides a textual interface with seven options for basic user interaction. By piping the users' interaction with the VCS through this additional presentation layer, several advantages can be obtained. Firstly, the complexity of the VCS is reduced significantly from the users' perspective, which also extensively limits the possibilities and the risk of using data-compromising functions (e.g. deletion of data, history re-writing). Secondly, it is possible to interpose routines for CQC, for instance in the form of data cleaning and harmonisation or functionality that anticipates and prevents potential problems, e.g. by capturing defective updates and automatically sending out notifications to VCS administrators. It proved very useful to send the notifications directly into a PMS[7] where they can be escalated depending on the level of expertise that is necessary for resolving the issues. The documentation and evaluation of tickets in the PMS in addition provides a good source for optimisation of the problem handling. This has already been implemented and used successfully. To overcome the problems with the VCS and binary files, changing to Git LFS and GitLab for hosting was necessary, which made no difference for the use with LAMA, which shows another advantage of the additional presentation layer.

3.2 Seamless Integration of VCS, CQC, and PMS with Cubo

To achieve a seamless integration of the quality control mechanism into the version control workflows, the Corpus Services (Ferger et al., 2020) tools are used. Additional checks and further options for automatic fixing of inconsistencies were implemented by adding to the Corpus Services code base.[8] While some checks were more project-specific (like comparing names and coordinates of settlements added to a map and added to the metadata file; ComaKmlForLocations), many checks are usable for a wide range of projects (such as a spell-checker for various languages; LanguageToolChecker).

The Corpus Services framework is intended to be usable in a wide variety of contexts and designed with an architecture as open as possible for contributions by other developers and projects.[9] A list of all available functions with a short description is available.[10]

In combination with the Corpus Services extensions a collection of scripts called Cubo were established.[11]By automating the CQC to run at a specified time (while no one is working on the data) and directly integrating the changes and reports into the VCS no changes to the creation workflow are needed.

The cubo scripts accomplish different tasks during their runtime. Mirrors of the linguistic corpora are used to run the CQC routines by checking out the most recent version of the data edited by the researchers. Then the corpus services tool is used to run CQC routines customized to the data, consisting of changes to the data (fixing) and reporting on inconsistencies that need to be mended manually. During this step messages are sent to a messenger used for project communication[12] updating on the current state of inconsistencies in the data. Central monitoring using reports and charts created with the corpus services tool[13] are added to the messages. After the CQC procedures the changes on the data and reports are added to the VCS, keeping the fixing and checking steps apart from each other to allow for precise rollback of changes anytime. In case of errors or a very high number of inconsistencies the cubo scripts additionally send reports and messages to the PMS[14] and a messenger[15]. The cubo scripts download the most recent version of the corpus services tool pre-build on GitLab using CI functionality.[16]

[7] Redmine for instance —like other PMSs— provides a REST API that can be used for automated issue creation

[8] `https://gitlab.rrz.uni-hamburg.de/corpus-services/corpus-services`

[9] Originating from developments at the HZSK (Hamburger Zentrum für Sprachkorpora), substantial developments have been made in INEL and the project QUEST "Quality - Established: Erprobung und Anwendung von Kurationskriterien und Qualitätsstandards für audiovisuelle, annotierte Sprachdaten" also started to contribute.

[10] see `https://gitlab.rrz.uni-hamburg.de/corpus-services/corpus-services/-/blob/develop/doc/List_of_corpus_functions.md`

[11] "Curation Bot", the scripts are included in the code base of Corpus Services, see `https://gitlab.rrz.uni-hamburg.de/corpus-services/corpus-services/-/tree/develop/cubo`

[12] Mattermost: `https://mattermost.com/`

[13] `https://inel.corpora.uni-hamburg.de/curation/`

[14] Redmine: `https://www.redmine.org/`

[15] Mattermost: `https://mattermost.com/`

[16] see `https://gitlab.rrz.uni-hamburg.de/corpus-services/corpus-services`

4 Discussion

The integration of continuous quality control measures into the corpus creation and version control workflows accomplished the goal of achieving a higher and measurable level of data quality and a considerable simplification of the publication of the produced high-quality data. Advancements to the version control setup and the implementation of further functionality in the Corpus Services tool were needed for the integration to be successful. As intended, no changes had to be applied to the corpus creation workflow, preventing additional effort for the implementation of the new setup.

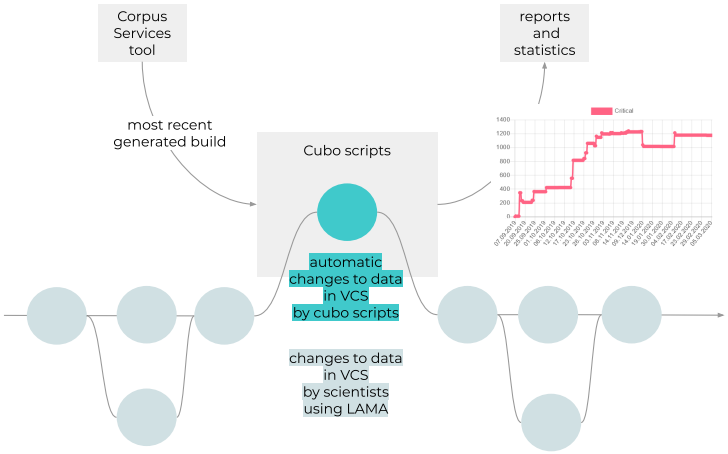


Figure 1: Cubo Scripts.

Additionally, further advantages of the setup showed, including the possibility to reset any changes done automatically by separating the fixes with the Corpus Services tool in the version control system. The automated application of the continuous quality control measures allows for the generation of meaningful statistics on the inconsistencies in the data and simplifies the manual mending by supplying reports together with the corpora.

The setup along with the software and scripts can be used and expanded in other projects and contexts. The project QUEST (Arkhangelskiy et al., 2020) for instance already aims at further development of the Corpus Services tool.

Acknowledgements

Parts of this paper have been produced in the context of the joint research funding of the German Federal Government and Federal States in the Academies' Programme, with funding from the Federal Ministry of Education and Research and the Free and Hanseatic City of Hamburg. The Academies' Programme is coordinated by the Union of the German Academies of Sciences and Humanities.

References

Timofey Arkhangelskiy, Hanna Hedeland, and Aleksandr Riaposov. 2020. Evaluating and assuring research data quality for audiovisual annotated language data. In *Selected papers from the CLARIN Annual Conference2020*, page 180 1–7. Linköping Electronic Conference Proceedings 180.

Alexandre Arkhipov and Chris Lasse Däbritz. 2018. Hamburg corpora for indigenous northern eurasian languages. *Tomsk Journal of Linguistics and Anthropology*, 21(3):9–18.

Anne Ferger and Daniel Jettka. 2021. LAMA – your friendly and easy git script (version 3.0), https://doi.org/10.5281/zenodo.4725651.

Anne Ferger, Daniel Jettka, and Timm Lehmberg. 2019. Tools and methods for continuous collaboration and curation, https://www.clarin.eu/sites/default/files/clarin2019_bazaar_ferger.pdf.

Anne Ferger, Hanna Hedeland, Daniel Jettka, and Tommi Pirinen. 2020. Corpus Services (version 1.0), https://doi.org/10.5281/zenodo.4725655.

Hanna Hedeland and Anne Ferger. 2020. Towards Continuous Quality Control for Spoken Language Corpora. *International Journal of Digital Curation*, 15(1), https://doi.org/10.2218/ijdc.v15i1.601.

Francisco Queiroz, Raniere Silva, Jonah Miller, Sandor Brockhauser, and Hans Fangohr. 2017. Track 1 Paper: Good Usability Practices in Scientific Software Development. In *Proceedings of the Workshop on Sustainable Software for Science: Practice and Experiences (WSSSPE5.1)*, 8.

The TEI-based ISO Standard "Transcription of Spoken Language" as an Exchange Format within CLARIN and beyond

Hanna Hedeland
Leibniz-Institut für Deutsche Sprache
Mannheim, Germany
hedeland@ids-mannheim.de

Thomas Schmidt
Research and Infrastructure Support
Universität Basel, Switzerland
th.schmidt@unibas.ch

Abstract

This paper describes the TEI-based ISO standard 2462:2016 "Transcription of spoken language" and other formats used within CLARIN for spoken language resources. It assesses the current state of support for the standard and the interoperability between these formats and with relevant tools and services. The main idea behind the paper is that a digital infrastructure providing language resources and services to researchers should also allow the combined use of resources and/or services from different contexts. This requires syntactic and semantic interoperability. We propose a solution based on the ISO/TEI format and describe the necessary steps for this format to work as an exchange format with basic semantic interoperability for spoken language resources across the CLARIN infrastructure and beyond.

1 Introduction

Today, the CLARIN infrastructure is well established across Europe, comprising a network of centres providing a vast number of digital resources and services. Since an increasing number of funders require researchers in the humanities and social sciences to deposit their data for reuse, the collections of digital resources hosted within CLARIN are growing steadily. Following the digital turn, the use of CLARIN's tools and services for manual and automatic analysis has also become a relevant option for research projects from various disciplines. An ideal scenario would allow researchers to use and freely combine data and tools or services from different CLARIN centres and contexts across the infrastructure. This, however, is still possible only for smaller sets of resources – large scale interoperability remains a desideratum. Unlike early digital corpora created by pioneering corpus linguists, digital language resources today seldom fit into the traditional view of language data as "natural running text" or "a single stream of tokens". This is particularly true for spoken or multi-modal resources, which are at the same time no longer a rare exception in the resource landscape.

2 A standard for spoken language transcription?

One reason for the heterogeneity of spoken language corpora is the existence of several widely used tool formats. ELAN (Sloetjes, 2014), Praat (Boersma, 2001), CLAN (MacWhinney, 2000) and EXMARaLDA (Schmidt and Wörner, 2014) all come with their individual formats, which are, apart from Praat's TextGrid format, XML-based. These formats are mainly based on similar tier-/time-based data models and to a sufficient extent interoperable – from the syntactic perspective. A file in one format can usually be converted into a file with a representation of the data using another format. There are undoubtedly some limitations regarding conversion scenarios, depending on the varying complexity of data models, where e.g. certain tier hierarchies or associations between annotation elements in ELAN's EAF format cannot be modelled by the more restrictive data model for Basic Transcriptions (EXB) in the EXMARaLDA system, but in these rather rare cases, customized workarounds are still possible.

From a semantic perspective however, interoperability is not that straightforward. As an example, the CHAT format of the CLAN software exactly defines the set of transcription and annotation conventions to be used for common spoken language phenomena, which makes the data easy to process and understand. But researchers are at the same time required to subscribe to theoretical concepts implemented by these conventions, and this is not a good basis for a standard to be used across disciplinary boundaries. On the other side of the spectrum, the EAF format of the ELAN software hardly imposes any restrictions on the individual researcher who is free to define the structure and content of the data format according to her needs. While this promises a perfect fit for the individual research context, data modelling is not trivial and not all variation is semantically relevant. It should be noted that ELAN provides means for defining the semantics of tiers and annotations using references to ISOcat, but this has hardly been adopted as a practice by researchers (cf. von Prince and Nordhoff (2020)) and ISOcat had its own issues.

The idea behind the ISO standard for Transcription of spoken language (ISO/TC 37/SC 4, 2016; Schmidt, 2011) is a solution which differentiates between general information that is shared across different research methods and disciplines on the one hand, and information that is theory-dependent (cf. Ochs (1979)) and therefore cannot be standardized, on the other. Standardization can be applied to aspects of the shared reality of spoken conversation, which includes e.g. the modelling of participants and the temporal alignment of their contributions, referred to here as macro-structure. The ISO/TEI format is not a tier-/time-based format, but instead models speaker contributions as a common list of `<u>` elements, possibly containing one or more `<seg>` element for the linguistic units defined by the relevant transcriptions system via `@type` and `@subtype` attributes, e.g. `@type="intonation-phrase" @subtype="falling"`. Annotations are by default modelled by `<span>`s in `<spanGrp>`s with an additional element `<annotationBlock>` to group the speaker contribution `<u>` with all relevant annotations. References to defined speakers and time points are modelled by the attributes `@who`, `@start` and `@end`, with the option to use `<anchor>`s for additional alignment in any position.

While some aspects of speaker contributions can be standardized, such as the existence of pauses and (possibly) non-verbal behaviour, the detailed choices regarding e.g. a set of relevant different pause durations or the descriptions of non-verbal behaviour are not part of the standard but of the transcription system currently in use. The same is true for the details of the segmentation into linguistic units in `<seg>`s, which usually differs according to the linguistic level used as the basis, e.g. intonation phrases for interactional prosody or utterances for pragmatics. Allowing for controlled variation within this area, the micro-structure, which defines the precise form of representation for spoken material, makes it possible to represent data created with different transcription systems using the same standard format.

3 Support for the ISO/TEI format in CLARIN

Within CLARIN, centres are not bound to accept or support particular formats, but several lists and overviews of standards and recommendations have been available over the years. Many centres refer to these resources[1] to define the formats they accept as deposits, e.g. for the Core Trust Seal, and thus include TEI as a general recommendation without further specifying any specific variants. The CLARIN Standards Committee has been gathering information on the recommendations on standards and formats actively issued by individual (mainly B) centres and made this information available on their web page[2]. A brief assessment of this information can provide insights into the current and potential support for the ISO/TEI standard within CLARIN. For this paper, the recommendations given by individual centres were revisited to allow for a more detailed picture. As not all (B) centres provide this information yet, the picture is however not complete. Since there is also no consistent and reliable information on the general types of resources a centre accepts nor on specific restrictions e.g. regarding languages or time periods, negative results cannot really be interpreted.

Nevertheless, of the centres that provide their own preferences and recommendations, three groups

[1] Such resources are e.g. `https://www.clarin.eu/faq/what-standards-are-recommended-clarin` or `https://www.clarin.eu/sites/default/files/Standards\%20for\%20LRT-v6.pdf`

[2] `https://www.clarin.eu/content/standards`, this abstract is based on the Release 0.1 from January 2021 (the list at `https://www.clarin.eu/content/standards-and-formats` includes links to centres' original published documents in English)

with respect to ISO/TEI support can be distinguished. Three B centres already recommend ISO/TEI explicitly: the CLARIN.SI Language Technology Centre, the Hamburg Centre for Language Corpora (HZSK) and the Leibniz-Institut für Deutsche Sprache (IDS). The second group recommends TEI, but not explicitly ISO/TEI (or other variants). Among these are the Austrian Centre for Digital Humanities and Cultural Heritage - A Resource Centre for the HumanitiEs (ACDH-ARCHE), Eberhard Karls Universität Tübingen (EKUT), Meertens Instituut/HuC (MI) (which only includes XML in the list, but refers to TEI as an example). And as noted above, all centres referring to existing CLARIN documents also in effect recommend TEI without further restrictions. The third group is the most interesting, since these centres explicitly recommend other widely used formats and not ISO/TEI. The CMU-TalkBank (CMU) recommends CHAT (only), MPI for Psycholinguistics (MPI-PL) recommends CHAT too, though in addition to EAF and Praat, which are in turn also recommended by The Language Bank of Finland (FIN-CLARIN) and the Bayerisches Archiv für Sprachsignale (BAS). Both Praat and EAF can be converted into the ISO/TEI format with dedicated software as described in (Schmidt et al., 2017), and this also applies to CHAT data that passes the data quality and consistency tests in CLAN. Still, the ISO/TEI format seems to be of little relevance to these four centres, presumably because of strong traditions and eco-systems around specific formats for specific types of resources and research areas. On the other hand, in addition to the information from certified B centres, there is information on accepted and recommended formats from the centres Open Resources and TOols for LANGuage (ORTOLANG) and Language Archive Cologne (LAC), which are both participating in knowledge centres and aiming for B Centre status: both recommend the ISO/TEI format for deposits. Furthermore, the LINDAT/CLARIAH-CZ centre, which does not give explicit recommendations on formats to depositors, now hosts the TEI-based TEITOK system (Janssen, 2016; Janssen, 2021), which includes both a search engine, visualization and editing functionality and has many features for spoken language. Since it is interoperable with e.g. EXMARaLDA and EAF through a set of scripts, interoperability between the TEITOK and ISO/TEI formats should not be difficult to establish.

4 Tools and Services for ISO/TEI within and beyond CLARIN

For a standard to be useful to researchers and operators of research infrastructures, there need to be sufficient relevant use cases and software solutions that are compatible with existing tools and methods. For data creation, thanks to the existing conversion functionality described above (Schmidt et al., 2017), widely established tools can continue to be used. The EXMARaLDA transcription and annotation editor can not only export the ISO/TEI format, but also import these files e.g. after further enrichment outside of the EXMARaLDA environment.

Since the creation of the ISO/TEI standard, the format has been used as the basis for enhanced interoperability with several existing tools and services. In many cases, this was software created on the basis of data models or notions of written language. Since the ISO/TEI standard is a TEI-based format, it shares a common core with TEI variants used for written language data and thus facilitates interoperability across the spoken and written modality. For instance, the development of WebAnno-MM (Remus et al., 2019) as an extension for audiovisual and transcription data in the ISO/TEI format allows manual annotation with a wider textual focus than transcription tools offer, and also more complex types of annotations such as tree or chain annotations.

For automatic annotation, the converters described above were integrated into the WebLicht SOA (Hinrichs et al., 2010) of CLARIN-D, thus enabling the use of various services from all German centres. Initially, this meant another mapping to formats and services for written data (internally, TCF, see Schmidt et al. (2017)), but services adapted to spoken language data based directly on the ISO/TEI format have now also been developed (Fisseni and Schmidt, 2020) and can improve results where the linguistic characteristics of spoken and written language differ to a great extent. The phonetics web services provided by the BAS (Kisler et al., 2017) have been able to import and export ISO/TEI data since version 2.36 of January 2020.

Based on the ISO/TEI format, the project ZuMult has developed new web-based functionality for both visualization and browsing of spoken language corpora within qualitative approaches and for complex

querying and analysis[3] based on an extension of the MTAS system (Brouwer et al., 2017) using CQP and a highly efficient query engine (Frick and Schmidt, 2020). Another corpus analysis platform that now supports the ISO/TEI format is Tsakorpus (Arkhangelskiy et al., 2019), which is one use case for ISO/TEI within the long-term project INEL (Arkhipov and Däbritz, 2018; Ferger and Jettka, 2020). Another project in the field of language documentation, the DoReCo project (Paschen et al., 2020), developed the Multitool to generate ISO/TEI as a distribution format for resources in various tool formats. The use of the ISO/TEI standard as a pivot format for various language resources and different tool formats has also been implemented as a proof-of-concept workflow (Parisse et al., 2018).

5 Discussion

The development of interfaces between the ISO/TEI standard and various existing tools and services has shown that this is not only feasible, but also efficient using the ISO/TEI standard as a pivot format. This is important since software development and maintenance is usually the bottleneck in the development of the infrastructure. By using a TEI-based format for spoken data, apart from the proximity to more familiar written language data models on the textual level, interoperability on the metadata level could also be facilitated. With the TEI header, there is also a common structure for a core set of relevant contextual information on the setting and the participants, e.g. for analyses within virtual collections. Since TEI is used and extended in many contexts, there are also existing conventions for basic token-based linguistic annotation (Bański et al., 2018) and a common approach for the integration of the W3C standard RDFa is being developed (Chiarcos and Ionov, 2019) to tackle the issue of strict linked data requirements.

Though conversion is already possible for widely used tool formats, as pointed out above, only features of the macro-structure are defined by the ISO/TEI standard, and only syntactic interoperability is to some extent simple to achieve. For semantic interoperability, the tier structure, the annotation levels and schemas and the conventions for transcription – the micro-structure – also need to be made explicit and machine processable to allow for tokenization and structural mark-up. This means that a conversion into the ISO/TEI format is not only a question of interoperability with a standard, but at the same time a process of FAIRification, of defining the semantic model of the data, making it more transparent and increasing the number and types of possible re-use scenarios. Creating digital language resources that are FAIR according to the well-known principles (Wilkinson and others, 2016) is a great, and often somewhat abstract, challenge for CLARIN and its users. We suggest that the adoption of the ISO/TEI standard with its basic semantics and the corresponding conversion scenarios as a way of assessing digital language resources could not only improve interoperability across resources, but also increase their general FAIRness and help foster a culture of data documentation required for truly FAIR infrastructures for both humans and machines.

References

Timofey Arkhangelskiy, Anne Ferger, and Hanna Hedeland. 2019. Uralic multimedia corpora: ISO/TEI corpus data in the project INEL. In *Proceedings of the Fifth International Workshop on Computational Linguistics for Uralic Languages*, pages 115–124, Tartu, Estonia, January. Association for Computational Linguistics.

Alexander Arkhipov and Chris Lasse Däbritz. 2018. Hamburg corpora for indigenous Northern Eurasian languages. *Tomsk Journal of Linguistics and Anthropology*, (3):9–18.

Piotr Bański, Susanne Haaf, and Martin Mueller. 2018. Lightweight grammatical annotation in the TEI: New perspectives. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018), 7-12 May 2018, Miyazaki, Japan*, pages 1795–1802, Paris, France. European language resources association (ELRA).

Paul Boersma. 2001. Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10):341–345.

Mathhijs Brouwer, Hennie Brugman, and Marc Kemps-Snijders. 2017. MTAS: a solr/lucene based multi tier an-notation search solution. In *Selected papers from the CLARIN Annual Conference*, pages 19–37, Aix-en-Provence, France. Linköping University Electronic Press, Linköpings Universitet.

[3] `http://zumult.ids-mannheim.de/ProtoZumult/index.jsp`

Christian Chiarcos and Max Ionov. 2019. Linking the TEI: Approaches, Limitations, Use Cases. In *Digital Humanities Conference 2019 (DH2019)*, Utrecht University, July.

Anne Ferger and Daniel Jettka. 2020. Use cases of the ISO standard for Transcription of spoken language in the project INEL. In *Proceedings of the CLARIN Annual Conference 2020*. CLARIN ERIC.

Bernhard Fisseni and Thomas Schmidt. 2020. CLARIN web services for TEI-annotated transcripts of spoken language. Selected Papers from the CLARIN Annual Conference 2019. Leipzig, 30 September–2 October 2019, pages 12–22. Linköping University Electronic Press, Linköping.

Elena Frick and Thomas Schmidt. 2020. Using full text indices for querying spoken language data. In *Proceedings of the 8th Workshop on Challenges in the Management of Large Corpora*, pages 40–46, Marseille, France, May. European Language Ressources Association.

Erhard W. Hinrichs, Marie Hinrichs, and Thomas Zastrow. 2010. WebLicht: Web-Based LRT Services for German. In *Proceedings of the ACL 2010 System Demonstrations*, pages 25–29.

ISO/TC 37/SC 4. 2016. Language resource management – Transcription of spoken language. Standard ISO 2462:2016, International Organization for Standardization, Geneva, Switzerland.

Maarten Janssen. 2016. TEITOK: text-faithful annotated corpora. In Nicoletta Calzolari et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23–28, 2016*. European Language Resources Association (ELRA).

Maarten Janssen. 2021. A corpus with wavesurfer and TEI: Speech and video in TEITOK. In Kamil Ekštein, František Pártl, and Miloslav Konopík, editors, *Text, Speech, and Dialogue*, pages 261–268, Cham. Springer International Publishing.

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

Brian MacWhinney. 2000. *The CHILDES project: Tools for Analyzing Talk, Third edition. Volume I.* Lawrence Erlbaum, Mahwah, NJ u.a., 3rd edition.

Elinor Ochs. 1979. Transcription as theory. In E. Ochs and B.B. Schieffelin, editors, *Developmental pragmatics*, pages 43–72. Academic Press, New York.

Christophe Parisse, Céline Poudat, Ciara R. Wigham, Michel Jacobson, and Loïc Liégeois. 2018. CORLI: A linguistic consortium for corpus, language, and interaction. In *Selected papers from the CLARIN Annual Conference 2017, Budapest, 18–20 September 2017*, pages 15–24, Budapest, Hungary. Linköping University Electronic Press, Linköpings Universitet.

Ludger Paschen, François Delafontaine, Christoph Draxler, Susanne Fuchs, Matthew Stave, and Frank Seifart. 2020. Building a time-aligned cross-linguistic reference corpus from language documentation data (DoReCo). In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2657–2666, Marseille, France, May. European Language Resources Association.

Steffen Remus, Hanna Hedeland, Anne Ferger, Kristin Bührig, and Chris Biemann. 2019. WebAnno-MM: EXMARaLDA meets WebAnno. In *Selected papers from the CLARIN Annual Conference*, Pisa. Linköping University Electronic Press, Linköpings Universitet.

Thomas Schmidt and Kai Wörner. 2014. EXMARaLDA. In Jacques Durand, Ulrike Gut, and Gjert Kristoffersen, editors, *Handbook on Corpus Phonology*, pages 402–419. Oxford University Press.

Thomas Schmidt, Hanna Hedeland, and Daniel Jettka. 2017. Conversion and annotation web services for spoken language data in clarin. In *Selected papers from the CLARIN Annual Conference*, pages 113–130, Aix-en-Provence, France. Linköping University Electronic Press, Linköpings Universitet.

Thomas Schmidt. 2011. A TEI-based Approach to Standardising Spoken Language Transcription. *Journal of the Text Encoding Initiative*, 1, 06.

Han Sloetjes. 2014. ELAN: Multimedia annotation application. In Jacques Durand, Ulrike Gut, and Gjert Kristoffersen, editors, *Handbook on Corpus Phonology*, pages 305–320. Oxford University Press.

Kilu von Prince and Sebastian Nordhoff. 2020. An empirical evaluation of annotation practices in corpora from language documentation. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2778–2787, Marseille, France, May. European Language Resources Association.

Mark D. Wilkinson et al. 2016. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3:160018–, March.

Curation criteria for multimodal and multilingual data: a mixed study within the Quest Project

Amy Isard
IFUU/IDGS
University of Hamburg, Germany
amy.isard@uni-hamburg.de

Elena Arestau
IFUU
University of Hamburg, Germany
elena.arestau@uni-hamburg.de

Abstract

We conducted a user survey and expert interviews within the ongoing Quest project to get an impression of the needs of users and researchers who are working with multimodal and multilingual linguistic corpora. This contribution describes the design and results of the mixed study, whose main goal is to improve the reuse potential of these resources, and to identify concrete topics which are important for the curation of such data.

1 Introduction

The research described in this paper was conducted during the ongoing Quest project[1] (Arkhangelskiy et al., 2020), which has the aim of enhancing research data quality and re-use for audiovisual annotated language data, and improving adherence to the FAIR principles (Wilkinson et al., 2016). The Quest project will create a web portal for users who intend to deposit or create a corpus, which will complement the existing CLARIN services for corpus access and reuse. The website will cater both to users who are in the process of designing and creating a corpus and those who have already completed their corpus collection and/or annotation. For the former it will provide a knowledge base and walk-throughs on various topics, such as metadata and anonymization, offering for example wizard forms like the CLARIN Data Management Plan[2] and DARIAH Consent Form[3]wizards. Once a corpus is completed, the portal will provide tools which allow users to check for conformity against various criteria, for example whether the structure of the corpus conforms to what is specified in the metadata. Once checks have been carried out, the corpus can then be deposited.

The studies reported here relate specifically to the project's work on curation criteria for multimodal data and for the linguistic secondary use of multilingual data. We set out to get an impression of the needs of corpus researchers, and the obstacles which they currently encounter in re-using or creating such data. Based on this, we have identified concrete topics which are important for the curation of such data, which have informed our development of the tools and knowledge-base in the Quest portal.

It was decided that the most effective method for designing this study would be a mixed approach (Rubin and Rubin, 2005) which involves 1) a quantitative user survey and 2) qualitative interviews with researchers and experts as data providers, users and creators. The results of the study highlighted the need for transparent and consistent criteria for the documentation, annotation and metadata in the areas of multilingual and multimodal corpora.

2 Survey

2.1 Design and Creation

The target groups of our survey were researchers who were involved in projects dealing with multimodal or multilingual data. The survey was open between July 2020 and March 2021. During this time it was

[1] https://www.slm.uni-hamburg.de/en/ifuu/forschung/forschungsprojekte/quest.html
[2] https://www.clarin-d.net/en/preparation/data-management-plan
[3] https://consent.dariah.eu

advertised a number of times via twitter, DhD-blog, corpora-list, linguistlist, internal mailing lists, and professional associations.

For the conceptualisation of the survey we were informed by several studies dealing with the curation, management and reuse of research data (Ferus et al., 2015; Fandrych et al., 2016; Arndt et al., 2018). Based on these studies and on the preliminary criteria we developed a catalogue of questions. We conducted a pilot survey with 5 participants prior to the survey release and then finalised the questions together with other project members.

The survey was created using the open source-software LimeSurvey (LimeSurvey GmbH, 2021) and was available in German and English. The survey contained a maximum of 74 questions, but was designed so that later questions were presented depending on the answers to earlier ones, to avoid participants having to see and respond to questions which were not relevant for them. Data from the survey were handled anonymously, to ensure that there would not be any privacy concerns and that participants would feel free to make negative comments if necessary.

The questionnaire consisted of seven subject blocks covering the following topics relating to the corpus used by the survey participant. The subjects were chosen based on the FAIR principles and the objectives of the Quest project. In all cases, questions which might lead to loss of anonymity, such as the name of the corpus, were optional. Some questions were multiple choice, and others allowed free text input. At the end of each section, there was a text field where participants could add any extra comments. The questionnaire blocks were as follows:

1. **Corpus General Information** - which format the corpus was in, which primary data it contained, what questions the participant was researching.
2. **Languages** - the languages present in the corpus, including primary data and translations.
3. **Transcription and Annotation** - which transcriptions and annotations were already present, and which were added by the participant.
4. **Anonymization** - what type of anonymization was present, if any, and whether it affected the research.
5. **Metadata** - which metadata and/or bias statements were included, if any, and whether they were sufficient.
6. **Access** - How the participant accessed and worked with the corpus, and any problems which they encountered.
7. **Participant General Information** - the country, type of institution and research area the participant works in.

2.2 Survey Results

The survey was fully completed by 44 participants, and we include only completed results in our analysis. Although this number of responses does not allow us to draw firm quantitative conclusions, we were able to observe some trends and received useful feedback in the free-form comment fields. The number of questions answered by each participant ranged between 23 and 53, with an average of 35. The participants currently work in 13 different countries: Germany, Italy, Australia, France, Brazil, Ireland, USA, Hungary, Canada, Czech Republic, UK, Tunisia and Austria, with the majority in Germany (62%). They are active in a wide range of research areas: Linguistics, Corpus Linguistics, Computational Linguistics, Historical Linguistics, Multilingualism, Language Acquisition, Sociolinguistics, Translation and Interpreting, Computer Science, Virtual Agents, Multimodal Behaviour, Finance and Sociology. They are employed by universities (72%), data centres, companies and archives.

The corpora described by the participants contained more than 30 different languages as primary data. Audio recordings were present in 84% of the corpora and video in 41%. Translations were present in 36% of the corpora, and 30% of participants stated that their research questions were in the area of

multilinguality. A variety of tools and formats were used including EXMARaLDA[4] (21.4%), FOLKER[5] (14.3%), PRAAT[6] (9.5%) and ELAN[7] (7.1%).

Eighty-eight percent of the participants stated that their corpus provided metadata, 7% that it did not, and 5% did not know. Ninety-four percent of those who had metadata stated that it was sufficient for their research needs. Where this was not the case, examples of what was missing included in one case detailed information of the recording location, the people present, and the position of the recording equipment, and in another full information about the languages spoken by subjects in a learner corpus. We also asked whether the corpora had documentation of potential biases in the data, for example in the form of a Data Statement (Bender and Friedman, 2018) or Data Sheet (Gebru et al., 2018); this was only confirmed in 7% of cases, while 49% stated that there was none present, and the remainder that they did not know.

Most of the participants found it easy to locate and use the corpora, but when asked about barriers to access and reuse, several mentioned the issue of funding - some corpora have expensive licences, and if a university or department does not already have a licence, the funds must be obtained from an individual project or research grant. Funding for the storage of corpora is also an issue, mentioned by the experts interviewed about multimodal corpora, detailed in the next section.

3 Interviews

3.1 Structure

We carried out qualitative semi-structured interviews to gather deeper insights into the experiences and needs of the experts as data providers and users. We conducted 20 interviews with experts in the areas of multilingual and/or multimodal corpora, and each interview lasted between 45 and 60 minutes. The interview topics were based on the survey, and each interview consisted of three key sessions: 1) Transcription and Annotation, 2) Formats, Standards and Metadata, and 3) Obstacles, Wishes, Suggestions and Challenges. The interviews were carried out in the areas of both multimodal and multilingual corpora. The experts in multimodal corpora worked in areas including the documentation of endangered languages, the semiotics of multimodal signed and spoken language interaction, multi-party interaction, non-verbal communication and the socio-linguistic contexts of communication, sign language corpora, and the interface between spoken language and gestural behaviour. The experts in multilingual corpora worked on interpreter-mediated interaction within the study of community interpreting, on the analysis of learner languages and errors, on first language attrition and second language acquisition and contrastive research.

3.2 Interview Results

Based on the interviews the following key conclusions can be made. The experts all agreed that corpus documentation is crucial for the reuse of resources and that detailed descriptions should be kept of every stage of corpus creation. The collection and publishing of metadata, including information on how the metadata were collected, is essential. Reference to standards or best practices in the field can ensure comparability of the data. Concerning transcription and annotation of multilingual resources, standardised transcription formats are necessary (for example CHAT (MacWhinney, 2000), HIAT (Rehbein et al., 2004)) and the use of editors like CLAN, Praat or EXMARaLDA can facilitate the transcription and annotation process. In addition, in the context of multilingualism, translations into a widely known language are necessary. The experts in multimodal corpora pointed out two major challenges in corpus creation. Firstly, the difficulty of discovering and following different national and international rules for data protection. Secondly, the storage of large amounts of video and audio data can be problematic, both in terms of cost and in terms of maintaining long-term storage options.

[4] https://www.exmaralda.org
[5] http://agd.ids-mannheim.de/folker_en.shtml
[6] htto://www.praat.org
[7] https://archive.mpi.nl/tla/elan

4 Conclusions

This study has provided information about the experiences of multimodal and multilingual corpus users and creators, which we have used to inform the design and content of the Quest project portal. We heard from corpus users from numerous countries and research areas, and were able to create an overall picture of the current needs and challenges of the research community and how they might be met. The most common issues raised as potential barriers to corpus reuse were those of storage for multimodal data, and of the need for extensive documentation at every stage of corpus creation.

References

Timofey Arkhangelskiy, Hanna Hedeland, and Aleksandr Riaposov. 2020. Evaluating and Assuring Research Data Quality for Audiovisual Annotated Language Data. In *Proceedings of CLARIN Annual Conference 2020*, pages 131–135, October.

Oleksandra Arndt, Laura Glatz, Benedikt Hummel, Magdalena Porst, Wassili Schabalowski, and Sophia Skubatz. 2018. Umfrage zum Forschungsdatenmanagement an der FH Potsdam : Projektbericht. Zenodo. DOI: 10.5281/zenodo.1161792.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Christian Fandrych, Elena Frick, Hanna Hedeland, Anna Iliash, Daniel Jettka, Cordula Meißner, Thomas Schmidt, Franziska Wallner, Kathrin Weigert, and Swantje Westpfahl. 2016. User, who art thou? User Profiling for Oral Corpus Platforms. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 280–287, Portorož, Slovenia, May. European Language Resources Association (ELRA). https://www.aclweb.org/anthology/L16-1043.

Andreas Ferus, Juan Gorraiz, Veronika Gründhammer, Christian Gumpenberger, Nikolaus Maly, Johannes Michael Mühlegger, José Luis Preza, Barbara Sánchez Solís, Nora Schmidt, and Christian Steineder. 2015. Researchers and their data. Results of an Austria survey–Report 2015. Zenodo. DOI: 10.5281/zenodo.34005.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*.

LimeSurvey GmbH. 2021. LimeSurvey: An Open Source survey tool. Hamburg, Germany. http://www.limesurvey.org.

Brian MacWhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk. Volume 1: Transcription format and programs*. NJ: Lawrence Erlbaum Associates.

Jochen Rehbein, Thomas Schmidt, Bernd Meyer, Franziska Watzke, and Annette Herkenrath. 2004. Handbuch für das computergestützte Transkribieren nach HIAT.

H. J. Rubin and I. S. Rubin. 2005. *Qualitative Interviewing: The Art of Hearing Data*. Thousand Oaks: Sage.

Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. 2016. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9.

Flexible metadata schemes for research data repositories
The Common Framework in Dataverse and the CMDI use case

Jerry de Vries
DANS-KNAW
The Netherlands
jerry.de.vries@dans.knaw.nl

Vyacheslav Tykhonov
DANS-KNAW
The Netherlands
Vyacheslav.tykhonov@dans.knaw.nl

Andrea Scharnhorst
DANS-KNAW
The Netherlands
andrea.scharnhorst@dans.knaw.nl

Eko Indarto
DANS-KNAW
The Netherlands
eko.indarto@dans.knaw.nl

Femmy Admiraal
DANS-KNAW
The Netherlands
femmy.admiraal@dans.knaw.nl

Abstract

This paper presents how DANS, which participates in the CLARIAH+ project, works on a Common Framework which makes it possible to expose CMDI metadata via a DANS discovery service. The Common Framework refers to discussions in CLARIN about integrating standards in Dataverse. This paper informs CLARIAH+ about the explorations of the envisioned use of the Common Framework and reports about the possibilities and challenges of the interoperability of these metadata schemes. The challenges faced are: First, a proposal of a core set of CMDI metadata as recommendation. Second, the extraction of CMDI metadata and transform and load the metadata fields into the Dataverse core set of metadata. Third, a workflow for prediction and linking concepts from external controlled vocabularies to CMDI metadata values. Fourth, the extension of the Common Framework with support for FAIR controlled vocabularies to create FAIR metadata. Fifth, the extension of the export functionality of Dataverse to export deposited CMDI metadata back to the original CMDI format

1 Introduction

Research data repositories are increasingly expected to operate together. Standardization and alignment of metadata schemes used to describe datasets are a precondition for any platform to work (see as an example https://datacite.org). At the same time, data repositories usually serve specific knowledge domains, and have tailored their indexing practices towards those communities. In short, there is a tension between serving one or few communities in a very detailed manner and being integratable into a cross-domain platform. The Dataverse community responded to this natural tension by offering both a standard, common core set of metadata called Citation Block and the possibility to extend this core set with custom fields defined as a discipline specific metadata block.

This paper discusses in detail challenges and solutions when it comes to implement Common Framework principles into a very concrete Dataverse instance and a very concrete community. (Conzett et al., 2020) More specifically, this paper reports how the Data Archive and Network Services institute (DANS)[1], which participates in the CLARIAH+[2] project, works on a Common Framework which makes

[1] https://dans.knaw.nl/en
[2] https://www.clariah.nl/

it possible to expose CMDI[3] metadata via a DANS discovery service. With Common Framework we refer here to discussions about standards in CLARIN integrated in Dataverse. The envisioned use of the exploration reported in this paper is two-fold: primarily, it informs CLARIAH+ about possibilities and challenges when it comes to the interoperability of metadata schemes; secondly, it informs DANS as service provider of a long-term archive in its use of a technological backbone. DANS is currently migrating its research data archiving service from a Fedora-based platform (DANS-EASY) to Dataverse[4] and introduced so-called DANS Datastations as specific Dataverse instances for designated communities. (Wals, 2021). The exploration described in this paper bases its analytic part on the current production system while at the same time, also informs the on-going migration process.

In the current DANS-EASY archive, CLARIN datasets are tagged as part of a specific collection containing 29 datasets. But, much more datasets use CMDI metadata, as search for 'CMDI' reveals (1096 datasets). The use of CMDI is often noted in either the Description metadata field or the Form metadata field of the Dublin Core Standard the current EASY is using. While the use of CMDI is noted, one cannot search in those CMDI notations. The CMDI based indexes are delivered as specific, additional files of the dataset, and hence not automatically searchable (in short called CMDI files). The core of the exploration of this paper, is the development of a so-called Extract, Transform and Load-pipeline (ETL). This pipeline extracts all metadata fields from CMDI files archived in CLARIN datasets at DANS-EASY[5] archive and automatically transform this metadata to the defined core set of CMDI metadata and load this metadata at a DANS Datastation. This results in findable and harvestable CLARIN metadata which is interoperable with CLARIN discovery services.

2 Five challenges

In this paper we detail challenges (and partly envisioned solutions) we are facing in our work on the Common Framework to expose CMDI metadata (ISO 24622-1:2015) (ISO 24622-2:2019) via a DANS discovery service and our work during the migration of the DANS archiving service to the DANS Datastation. Both are still ongoing processes.

2.1 Challenge 1: A proposal of a core set of CMDI metadata as recommendation

The first challenge is the fact, that CMDI itself acts as a recommended standard, but that there is (not yet) a defined core set of CMDI metadata. (Goosen et al., 2014) The CLARIN taskforce CMDI[6], in which we are participating, is currently working on a proposal and acceptance of a core set of CMDI metadata as a recommendation for all CLARIN centers. This core set of metadata will be the basis for an extension of the Dataverse core set of metadata for describing CLARIN datasets when depositing them in the corresponding DANS Datastation

2.2 Challenge 2: Extraction of CMDI metadata and transform and load the metadata fields into the Dataverse core set of metadata

The second challenge concerns the Dataverse software itself, and how to best extend the core Dataverse metadata schema with a set of metadata whereby each metadata field should be part of some CMDI component and linked to the CMDI component registry. The creation of a pipeline to Extract, Transform and Load CMDI metadata fields into the Dataverse Core Set is part of this challenge. This ETL-pipeline uses the Dataverse DDI Converter tool[7] which so far only supports customized XSLT mappings for xml input. We are extending this functionality with Jinja2[8] templating in combination with key-value mapping for csv input. In this case the DDI Converter tool will not only be depending on xml input, to guarantee a broader use of the converter tool in the ETL-pipeline.

[3] CMDI stands for Component Metadata Infrastructure; for the CLARIAH+ use case see: https://github.com/CLARIAH/usecases/blob/master/cases/dans-cmdi.md
[4] https://dataverse.org/about
[5] https://easy.dans.knaw.nl/ui/home
[6] TF CMDI members: Goosen, T, Windhouwer, M, Conzett, Ph, Uytvanck, D, Tykhonov, V
[7] https://github.com/IQSS/dataverse-ddi-converter-tool
[8] https://jinja.palletsprojects.com/en/2.11.x/

2.3 Challenge 3: Workflow for prediction and linking concepts from external controlled vocabularies to the CMDI metadata values

The third challenge consist in the extension of such a Common Framework for Dataverse beyond the CMDI case. Beyond the extension of the Citation Core set, what is also envisioned is to support a link between other 'indexing' metadata fields to the other Knowledge Organization System Providers. In particular, we think here of recommended FAIR[9] controlled vocabularies and ontologies which potentially become part of the set of metadata fields. (Wilkinson et al., 2016) (Broeder et al., 2021), (Wang et al., 2021) Coming back to the CMDI case, this could lead to linking a or any CMDI metadata value to a recommended ontology or controlled vocabulary.

2.4 Challenge 4: Extension of the Common Framework with support for FAIR controlled vocabularies to create FAIR metadata

The fourth challenge is the support of FAIR controlled vocabularies. We use the SKOSMOS[10] framework developed at the Finnish National Library. A semi-automatic workflow, which uses a SKOSMOS API, is developed to query any SKOSMOS representation of the recommended external controlled vocabularies. The NDE's Network of Terms GraphQL endpoint[11] is used to make linkage to the appropriate controlled vocabularies for the terms extracted from the CMDI fields. These metadata fields link to the CMDI component registry in the CMDI metadata schema.

2.5 Challenge 5: Extension of the export functionality of Dataverse to export deposited CMDI metadata back to the original CMDI format

The fifth and last challenge is to extend the export functionality of Dataverse to export deposited CMDI metadata back to the original CMDI format, following the used specification.

To support all these workflows in the Common Framework, the Apache Airflow[12] is investigated as a proper solution to implement a reliable deposit pipeline. All insights and workflows will be shared with the CLARIN and CLARIAH community and we're looking for the collaboration on semantic mappings that should be used to get an appropriate ontology linkage not only on value level but also between fields available in CMDI Component Registry. Another task is to link CMDI component fields to the common ontologies like DCAT2[13] and Dublin Core using RDF Modelling Language[14] (RML). The ultimate goal is to leverage the metadata with a search engine developed as a part of the ODISSEI[15] portal that will allow users to search across linked concepts from the different disciplines available in the Linked Open Data Cloud (LoD) including linguistics sources (CLARIN). The same semantic mappings could be reused on global scale to get CMDI datasets disseminated in the FAIR Data Points[16] developed in the FAIRsFAIR[17] project.

[9] FAIR stands for Findable, Accessible, Interoperable and Re-usuable.
[10] https://skosmos.org/
[11] https://termennetwerk.netwerkdigitaalerfgoed.nl/
[12] https://airflow.apache.org/
[13] https://www.w3.org/TR/vocab-dcat-2/
[14] https://rml.io/specs/rml/
[15] https://odissei-data.nl/en/
[16] https://www.fairdatapoint.org/
[17] https://www.fairsfair.eu/

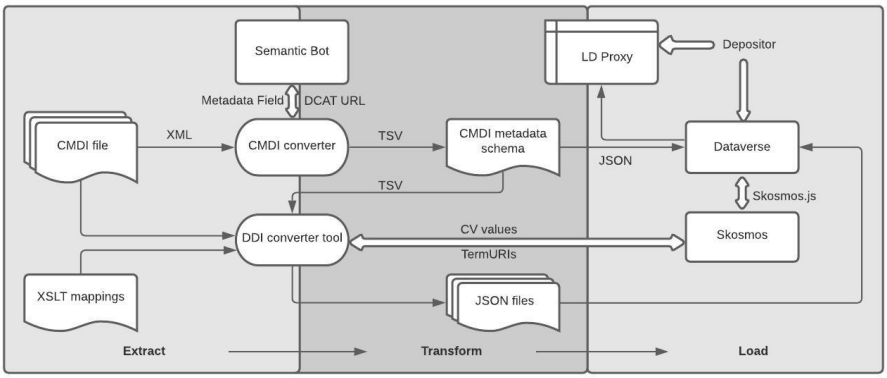


Figure 1. CMDI pipeline

3 Future work

By supporting the enrichment of metadata, we help to make CLARIN datasets Findable and Accessible and make them Interoperable with other CLARIN datasets, and so ultimately also support Re-usability. FAIR compliance automatic of assessment tools, like F-UJI[18] can be included in the Common Framework to evaluate the quality of the metadata. (Devaraju et al., 2020, 2021)

Future work will incorporate the application of Artificial Intelligence to get an automatic linkage of relations between CMDI values and relevant ontologies and controlled vocabularies, and create a semi-automatic mapping tool to generate RDF mappings for CMDI fields.

4 Conclusion

This paper elaborates on our (experimental) work of building a Common Framework to expose CMDI metadata via a DANS discovery service. This work relates to the migration of the DANS archive service to (a) newly to build DANS Datastation(s), which will serve as a basis for the discovery service.

However, the work is still ongoing and the challenges we reflect about when addressed unavoidably are leading to new challenges. For instance, we have been able to extend the Dataverse metadata model with a proposed core set of CMDI metadata which serves the needs of DANS as a basis for the discovery service. This resulted in a flexible solution which is easy to adjust in case the core set of CMDI metadata will be changed in the future. Its implementation in production services is still a challenge ahead.

To get to the proposed core set of CMDI metadata, we have analyzed all CMDI metadata stored in the DANS-EASY archive with the CMDI exploration tool[19]. With the same tool we are able to transform each CMDI metadata file to the proposed core set.

To make the new metadata FAIR, we explored the possibilities of enriching the metadata with recommended external controlled vocabularies. This exploration has led to a flexible and generic solution to add custom external controlled vocabularies to Dataverse beyond the immediate CMDI case. A semi-automatic workflow, which uses a SKOSMOS API, is developed to query any SKOSMOS representation of the recommended external controlled vocabularies. The NDE's Network of Terms GraphQL endpoint is used to make linkage to the appropriate controlled vocabularies for the terms extracted from the CMDI fields.

To extend the semi-automatic workflow we have started to explore the possibilities of semantic gateway. We have started a proof of concept with a semantic gateway lookup API. This API is able to return a list of standardized concepts based on the selected vocabulary and a term. This will help to link each field in the proposed core set of metadata to the appropriate controlled vocabulary.

To make the circle round again we are in the phase of investigating the export of the Dataverse metadata back to the original CMDI format. The basic requirement for this should be that the Dataverse metadata schema must have CMDI metadata that can be extended with custom components which are

[18] https://f-uji.net/
[19] https://github.com/Dans-labs/CLARIAH_CMDI

used by the different CLARIN centers. Second, the original relationships between fields and concepts should be kept whereby the custom components should be added to a SKOS schema. If this will be possible, we should be able to reproduce the original CMDI metadata, which could be offered for download to any user without losing of the quality of the original metadata.

This work has taught us that looking to the future and setting ourselves for big challenges is leading to new challenges. But these challenges are motivating us to build proper solutions with and for the community.

References

Broeder, D., Budrononi, P., Degl'Innocenti, E., Le Franc, Y., Hugo, W., Weiland, C., Wittenberg, P. and Zwolf, C. M. 2021. SEMAF: *A Proposal for a Flexible Semantic Mapping Framework (version 1.0).* Zenodo. http://doi.org/10.5281/zenodo.4651421

Conzett, P., Goosen, T., Scharnhorst, A., Tykhonov, V., Van Uytvanck, D., de Vries, J. and Wittenberg, M. 2020, *How to weave domain specific information sources into a large, FAIR data fabric for the Digital Humanities? The use of the Dataverse platform.* In S Barbiers, A Fokkens & C Olesen (eds), Proceedings of the DH Benelux 2020. Zenodo. https://doi.org/10.5281/zenodo.3879031

Devaraju, A., Huber, R., Mokrane, M., Herterich, P., Cepinskas, L., de Vries, J., L'Hours, H., Davidson, J. and White, A. 2020. *FAIRsFAIR Data Object Assessment Metrics.* https://doi.org/10.5281/zenodo.4081213

Devaraju, A., Mokrane, M., Cepinskas, L., Huber, R., Herterich, P., de Vries, J., Åkerman, V., L'Hours, H., Davidson, J., and Diepenbroek, M. 2021. *From Conceptualization to Implementation: FAIR Assessment of Research Data Objects*. Data Science Journal, vol. 20, no. 1. https://doi.org/10.5334/dsj-2021-004

Goosen, T., Windhouwer, M., Ohren, O., Herold, A., Eckart, T., Durco, M. and Schonefeld, O. 2015. *CMDI 1.2: Improvements in the CLARIN Component Metadata Infrastructure.* in J Odijk (ed.), Selected Papers from the CLARIN 2014 Conference, October 24-25, 2014, Soesterberg, The Netherlands, 116:004, Linköping Electronic Conference Proceedings, Linköping University Electronic Press, Linköpings universitet, Linköping, pp. 36-53. http://www.ep.liu.se/ecp/116/004/ecp15116004.pdf

ISO 24622-1:2015. (2015). *Language resource management – Component metadata infrastructure (CMDI) – Part 1: The Component metadata model*. Standard, International Organization for Standardization, Geneva, CH

ISO 24622-2:2019. (2019). *Language resource management – Component metadata infrastructure (CMDI) – Part 2: The Component metadata specification language.* Standard, International Organization for Standardization, Geneva, CH

Wals, H. 2021. Focus on FAIR. DANS: Dutch national centre of expertise and repository for research data. DANS Strategy 2021-2025. The Hague, https://dans.knaw.nl/en/about/organisation-and-policy/policy-and-strategy/dans-2021-2025/UK_DANS20212025.pdf

Wang, M., Qiu, L. and Wang, X. 2021, *A Survey on Knowledge Graph Embeddings for Link Prediction*. Symmetry, 13, 458. https://doi.org/10.3390/sym13030485

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J., Da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., ... Mons, B. 2016. The FAIR Guiding Principles for scientific data management and stewardship. Nature, 3, 160018. https://doi.org/10.1038/sdata.2016.18

Citation tracking and versioning for linguistic examples

Tobias Weber
Graduate School Language & Literature
Ludwig-Maximilians-Universität München
Munich, Germany
weber.tobias@campus.lmu.de

Abstract

This paper outlines the possible implementation of a data citation tracking method within the CLARIN services, based on Weber (2019), which has not been developed yet. The goal is to create collections of subsets of data, displaying the variation in their cited forms in the literature. This creates a citation infrastructure to increase transparency of scientific workflows, enrich data sets administered by CLARIN, and highlight their relevance.

1 Introduction

The CLARIN infrastructure provides access to a wide range of language data and tools hosted by many affiliated institutions, making their precise usage hard to track. With increasing demand for *FAIR* data (Wilkinson et al., 2016; De Jong et al., 2018), these language data must not only be internally consistent and compliant to international standards but should also allow for sustainable storage and use. The focus of this paper lies on two procedures which can support data management and scientometric research, if they were integrated into the CLARIN infrastructure: data citation tracking and a branching graph of versions. The basic design was outlined in two earlier articles (Weber, 2019; Weber, 2020a), yet without reference to a particular structure where it could be implemented. Furthermore, I refer the readers to these publications for the discussion of meta-scientific implications from the linguistic viewpoint.

Through its services for archiving and data presentation, CLARIN possesses three features which facilitate the implementation of the desired processes. First, data is stored with persistent identifiers, enabling the citation and consistent identification of data sets. Second, there is the provision for version control for the stored data sets, rendering the infrastructure flexible enough to deal with differing layers of data. Third, the virtual collection functionality allows for the curation and collation of data sets which are spread across databases and archives but form a meaningful and coherent unit. These features allow CLARIN to fulfil the promises of earlier attempts to create overarching database infrastructures for linguistic data, like the ODIN project (Lewis, n d), which did not live up to the expectations. Consequently, I consider CLARIN the ideal platform for piloting and trialling the envisioned workflow on the language data managed by the centres, before extending the scope to other resources.

2 Motivation and envisioned use

The envisioned workflow builds on experience in curating legacy materials, gained in a combination of philological editing of linguistic data and the reconstruction of a metadocumentation. The result is an inventory of occurrences and mentions of the underlying data sets, including a comparison of cited forms, across different physical and digital archives and publications. This output could also be contained in a virtual collection, as my manually compiled lists follow the same ideas as envisioned for this feature (Váradi et al., 2008). The workflow addresses predominantly issues outside the core motivation of data citation named by Silvello (2018) – data connection, data discovery, and data sharing through the infrastructure – while providing tools to address the issues of fixity/variation and citation identity.

The first desired feature is, thus, the automatic collection of linguistic examples in a virtual collection, as discussed in Arnold et al. (2020). These instances need to be collated and compared, which is also possible under the concept of the virtual collection. Instead of refined layers of data compiled from different sources (Bański et al., 2013), the desired output contains versions of the same linguistic examples in its various renderings and annotated formats, thereby addressing the fixity issue of citation tracking. At the same time, this might move the data sets into the direction of linguistic linked open data (van Erp, 2012). Parsed metadata can be used to enrich the collection, whereby no information is lost or overwritten in the process, allowing for nesting of entries or vertical Metadata Inheritance (Weber, 2020a). Importantly, the horizontal or temporal component of metadata must not be ignored, allowing the forward and backward search of other versions – in other words, version control. The original data sets can be enriched with additional information on this "citation tracking" for contained language data. An integrated view of different versions across the temporal and the abstractness dimensions helps to achieve two of CLARIN's goals: language data becomes sustainable, as existing annotations or analyses become accessible and comparable through the same platform, while interoperability of different versions can be increased through providing different versions. If treated as subsets of virtual collections, the provision of identifiers for each subset would enable citation of lower levels of data than the file, improving chances for citation and findability.

An important feature for this application is the comparison of transcription, annotation, translation, or similar layers of analysis – something which is reviewed whenever linguists interact with data (Weber and Klee, 2020). On the one hand, it may save time and effort for researchers if they can use existing annotations and facilitate reuse of data, as also desired by Bouda and Cysouw (2012), on the other hand, the comparison allows for a critical assessment of these annotation layers. This does not imply that the rendering of data is faulty, although there are studies showing that linguistic data citation is prone to errors (Engh, 2006), while good data citation metrics can furthermore prevent biases and imbalances in the data (Weber, 2020b). Yet, this tool to evaluate and compare forms and contexts of data citation is not only a means for quality assurance. Following recent trends in linguistic data citation (Berez-Kroeker et al., 2018; Andreassen et al., 2019), the goal is to acknowledge all contributions to data sets – in the understanding of science as evolving discourse, each contribution and additional analysis of an example furthers our knowledge. Gaining access to other annotations and evaluations enables a researcher to interact with these contributions to the scientific discourse, while also permitting credit to be given for each comment about an example (Weber, 2020a).

Given the wide-ranging implications of the envisioned workflow and the number of stakeholders, one can see why this procedure for data citation tracking is not yet implemented. There are too many instances negotiating over the responsibility for instances of data citation: Are depositors and archives responsible for ensuring that data is correctly cited and referenced, with information being collected and stored with the original data sets, or do publishers bear the responsibility for the data cited in their outputs? Could even an external database bridge the gap between original archival data and cited versions? The pragmatic approach advocated for in this paper is to utilise CLARIN's good infrastructure and political influence as an established international cooperation to initiate the adoption of data citation tracking and creation of comparative data sets of data use. This citation infrastructure would demonstrate the relevance of CLARIN in the management of language data and increase the visibility of the administered data sets; at the same time, CLARIN can offer an innovative tool for citation tracking which could also develop into a workflow for other repositories and archives.

3 Technical realisation

The basic set-up follows the implementation of existing citation tracking procedures, although, closer to Altmetrics (Liu and Adie, 2013) than citations of regular publications. This is due to the restriction that the aim is not just to track full data sets but also parts thereof which are specified within the text and not the references. As a result, the software must be capable of reading metadata files as well as source files in text formats which are used for scientific publications (e.g. PDF, HTML). Initially, the selection of publications could be reduced to a few open access journals and book series in linguistics –

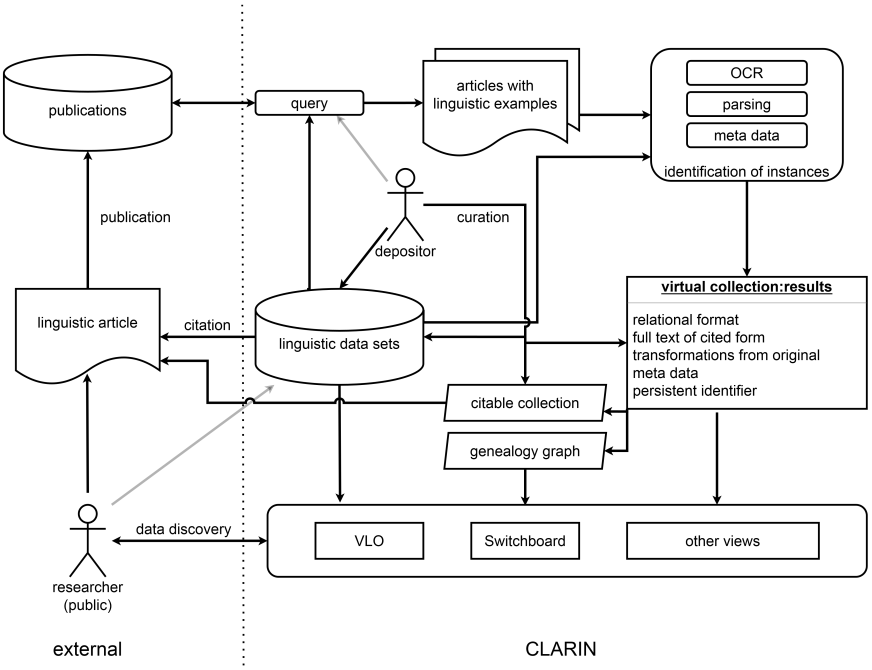


Figure 1. Schematic overview of the proposed workflow

potentially in collaboration with the publishers – which may cite data from the CLARIN resources. The application must be trained to recognise linguistic examples which may be easier for glossed data in a phonetic transcription – the recognised examples are then compared to existing data in (a selection or all) CLARIN corpora. Ideally, an additional check of possible transliterations and translations helps to establish a connection between cited data and the original data set.[1] Metadata and identifiers, if provided in the publication, must be stored and compared to the existing resource descriptions. If a high likelihood of identity, or related similarity measure, is returned by a string-matching algorithm, the form and relevant metadata about the contexts of its occurrence are stored in an additional entry of a "virtual collection" or similar data structure.[2] These collections can be made accessible through the Switchboard or the VLO, where they may be compared or even transformed into a graph displaying similarity like a genealogy of textual artefacts in philology (addressing the issue of citation identity). The location of the citation is added as an entry to the collection and linked to the original data set; the use of a persistent identifier for the collection allows access to the comparative format and enables researchers to cite a combination of different analyses from there. The "virtual collection" format would open the opportunity for manual additions, corrections, or curation of the collection, e.g. in cases where metadata were defective or formatting marks were not completely stripped.

Since CLARIN also holds data and resources from the 20th century which predate the introduction of digital identifiers, the retrospectively assigned identifiers will not be cited in the literature before a certain date. This shows that, ultimately, the data citation tracking must not only deal with forthcoming publications but also contain methods to retroactively track data. This may be further complicated by inconsistent publication formats or print-only publications. Yet, for digitised print publications, an OCR

[1]The human-readable format of linguistic data enables the use of these NLP applications on the cited data, which separates the tracking of linguistic examples from other types of data (e.g. quantitative data, other qualitative data like interviews).

[2]While it may appear sensible to save data storage by collecting only differences in a relational data format, there are instances where a full record may be justified (cf. Silvello (2018))

algorithm may be trained to detect glossed examples which follow the Leipzig Glossing Rules for most of the 21st century (with various similar systems of interlinear glossing already used throughout the 20th century). As a positive side note, many journals and publishers make older issues and editions freely available online or through their web-stores. This allows for extensive data tracking for older records, as access to publications appears to be a bottleneck in the workflow.

4 Conclusion

The outlined procedure for data citation tracking and creation of collections of versioned language examples cited in the literature has not been developed and is not implemented yet. Its addition to the existing CLARIN infrastructure would increase sustainability and accountability of the data sets and enable scientometric or philological research on the resources. A successful trial and implementation for an initially small set of data an publications may encourage other publishers and archives to adopt this method of data citation tracking and lead to standards for interoperable data in citation.

References

Helene N. Andreassen, Andrea L. Berez-Kroeker, Lauren Collister, Philipp Conzett, Christopher Cox, Koenraad De Smedt, Bradley McDonnell, and the Research Data Alliance Linguistic Data Interest Group. 2019. Tromsø recommendations for citation of research data in linguistics.

Denis Arnold, Ben Campbell, Thomas Eckart, Bernhard Fisseni, Thorsten Trippel, and Claus Zinn. 2020. The CMDI Explorer. In Constanza Navarretta and Maria Eskevich, editors, *Proceedings of CLARIN Annual Conference 2020*, pages 157–161. CLARIN, Utrecht.

Piotr Bański, Elena Frick, Michael Hanl, Marc Kupietz, Carsten Schnober, and Andreas Witt. 2013. Robust corpus architecture: a new look at virtual collections and data access. In Andrew Hardie and Robbie Love, editors, *Corpus linguistics 2013. Abstract book*, pages 23 – 25. UCREL, Lancaster.

Andrea L. Berez-Kroeker, Lauren Gawne, Susan Smythe Kung, Barbara F. Kelly, Tyler Heston, Gary Holton, Peter Pulsifer, David I. Beaver, Shobhana Chelliah, Stanley Dubinsky, Richard P. Meier, Nick Thieberger, Keren Rice, and Anthony C. Woodbury. 2018. Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics*, 56(1):1–18.

Peter Bouda and Michael Cysouw. 2012. Treating dictionaries as a linked-data corpus. In Christian Chiarcos, Sebastian Nordhoff, and Sebastian Hellmann, editors, *Linked Data in Linguistics: Representing and Connecting Language Data and Language Metadata*, pages 15–23. Springer Berlin Heidelberg, Berlin, Heidelberg.

Franciska De Jong, Bente Maegaard, Koenraad De Smedt, Darja Fišer, and Dieter Van Uytvanck. 2018. CLARIN: Towards FAIR and Responsible Data Science Using Language Resources. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, pages 3259–3264. European Language Resources Association (ELRA).

Jan Engh. 2006. *Norwegian examples in international linguistics literature. An inventory of defective documentation*. Universitetsbiblioteket i Oslo, Oslo.

William Lewis. n. d. ODIN - The Online Database of Interlinear Text. http://odin.linguistlist.org/.

Jean Liu and Euan Adie. 2013. Five challenges in altmetrics: A toolmaker's perspective. *Bulletin of the American Society for Information Science and Technology*, 39(4):31–34.

Gianmaria Silvello. 2018. Theory and practice of data citation. *Journal of the Association for Information Science and Technology*, 69(1):6–20.

Marieke van Erp. 2012. Reusing linguistic resources: Tasks and goals for a linked data approach. In Christian Chiarcos, Sebastian Nordhoff, and Sebastian Hellmann, editors, *Linked Data in Linguistics: Representing and Connecting Language Data and Language Metadata*, pages 57–64. Springer Berlin Heidelberg, Berlin, Heidelberg.

Tamás Váradi, Peter Wittenburg, Steven Krauwer, Martin Wynne, and Kimmo Koskenniemi. 2008. CLARIN: Common Language Resources and Technology Infrastructure. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. European Language Resources Association, Marrakech.

Tobias Weber and Mia Klee. 2020. Agency in scientific discourse. *Bulletin of the Transilvania University of Braşov Series IV: Philology and Cultural Studies*, 13(1):71–86.

Tobias Weber. 2019. Can Computational Meta-Documentary Linguistics Provide for Accountability and Offer an Alternative to "Reproducibility" in Linguistics? In Maria Eskevich, Gerard de Melo, Christian Fäth, John P. McCrae, Paul Buitelaar, Christian Chiarcos, Bettina Klimek, and Milan Dojchinovski, editors, *2nd Conference on Language, Data and Knowledge (LDK 2019)*, volume 70 of *OpenAccess Series in Informatics (OASIcs)*, pages 26:1–26:8, Dagstuhl, Germany. Schloss Dagstuhl–Leibniz-Zentrum für Informatik.

Tobias Weber. 2020a. Metadata Inheritance: New Research Paper, New Data, New Metadata? In Andrea Mannocci, editor, *Reframing Research Workshop Accepted Papers*. Zenodo.

Tobias Weber. 2020b. A philological perspective on meta-scientific knowledge graphs. In Ladjel Bellatreche, Mária Bieliková, Omar Boussaïd, Barbara Catania, Jérôme Darmont, Elena Demidova, Fabien Duchateau, Mark Hall, Tanja Merčun, Boris Novikov, Christos Papatheodorou, Thomas Risse, Oscar Romero, Lucile Sautot, Guilaine Talens, Robert Wrembel, and Maja Žumer, editors, *ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium*, pages 226–233, Cham. Springer International Publishing.

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. 2016. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3(1), March.

Bagman – A Tool that Supports Researchers Archiving Their Data

Claus Zinn
Seminar für Sprachwissenschaft
Universität of Tübingen, Germany
claus.zinn@uni-tuebingen.de

Abstract

Getting researchers to archive their data properly is hard. Many factors are at play. In this paper, we present *Bagman*, a software that aims at alleviating research data management significantly. Bagman is a web-based software that supports researchers to package their data, assign a minimal set of metadata for their description, define a licence for the data's future distribution, and to submit the entire package in a safe manner to an archive of their choice.

1 Motivation

Research data management is an essential ingredient of good scientific practise. Theories explain the data, and for one researcher to validate another researcher's theoretic models, the inspection of data is central. Nevertheless, many researchers regard the management of research data as a necessary evil. Although one clearly acknowledges the benefits of proper research data management, it is also perceived a something that is not done with overwhelming desire or pleasure.

Fear of scientific scrutiny and competition aside, proper research data management feels like household chores; one needs to make an inventory of all research data, clean-up the data, iron-out a proper file and directory structure of all data, document the procedures and scripts for data annotation and analysis *etc.* When everything is in order, one needs to describe the data with metadata, and then bundle and safely transfer it to an archive of one's choice, so that eventually – once it is ingested into the archive and published – fellow researchers can find and make proper use of it.

The assignment of metadata is a particular nuisance. For this, researchers have to become familiar with metadata standards, profiles, editors, and best practises. Moreover, researchers are expected to take care of licensing issues, and last but not least, know about archives that are well suited to host their precious data. Our new software, *Bagman*, aims at supporting researchers in all of the aforementioned areas.

2 Background

Getting your research data archived constitutes a workflow that varies across institutions. Details aside, it includes data packaging, metadata description, and transfer. Each step is accompanied by some quality control to minimize mishaps in these processes.

2.1 Packaging

In the worst case, researchers send their archive managers an email where all data is attached to the email. Sometimes data is put into some cloud space, or on portable storage devices for manual delivery. Such worst case scenarios often include data loss, files whose formats do not comply with archiving standards or whose names disobey naming conventions. Moreover, metadata descriptions might be anything from absent, incomplete or invalid XML. To avoid such mishaps, the art of packaging needs appreciation.

There are a number of tools that help researchers to bundle their research data into a single package. The open source software `docuteam packer` helps users bundling research data into a single package that can then be transfered to archives (Docuteam, 2018). A software called `Bagger` was created for the U.S. Library of Congress as a tool to produce a package of data files according to the BagIt specification (Kunze et al., 2018). The specification is a set of hierarchical file layout conventions for storage and transfer of arbitrary digital content. Simply speaking, it can be seen as a shopping cart (bag) together with a shopping bill that lists each of the items with its location (path) and its price (an MD5 or SHA

checksum). Those who receive the bag can use the inventory to check whether all goods were received in a complete and correct manner. – Our software, Bagman, makes use of this idea to help CLARIN researchers packing-up their research data so that it can be transfered to an archive in a safe manner.

2.2 Metadata

In the CLARIN community, for researchers to assign metadata to research data, they need to make use of the CMDI metadata framework (Broeder et al., 2012). For many researchers, this exercise feels like taming multi-headed monsters in a landscape that feels rough and bracketed from every angle. Researchers need to consult the CMDI component registry[1] to find a metadata profile that best fits their research data, and once they have identified a profile, they have to instantiate it to the best of their knowledge. This is not a trivial matter given that there are hundreds of profiles to choose from, but not a single metadata editor that gives intelligent help with instantiating the numerous different metadata fields.

No wonder, most CMDI-based descriptions have a rather poor descriptive power, taming the beast is exhaustive, and at some point one rather leaves it alone. As a result, researchers must be supported by dedicated archive management staff that is knowledgeable about the CMDI zoo of beasts, and that is armed with XML magic, best practises, and metadata processing tools to keep them at bay. – In Bagman, users are kept away from editing CMDI content directly. Information is gathered via simple forms, and information stemming from bagged resources is automatically added to the CMDI description.

2.3 Archiving

The German CLARIN website offers a "find your archive" service that helps researchers identifying the archive that is best suited to host their data.[2] Users are requested to answer questions about the modality of their research data (spoken language, written language, multi-modal language, sign language), its lingual type (German, multi-lingual, historical *etc*), the type of their resource (*e.g.*, lexicon, corpus, treebank), and whether they choose a public licence or not. As a result, the centres that fit the answers best are returned, together with the contact details of the respective archive managers. – Bagman will use the information submitted by the user to suggest archives that are suitable for hosting the user's research data. Once the user selected the archive, the bag will be safely transfered to a neutral place; the archive manager can download the bag from there, inspect the package, and then contact the user to proceed with the archiving procedure. Bagman hence acts as a broker between researcher and archive manager.

3 Bagman

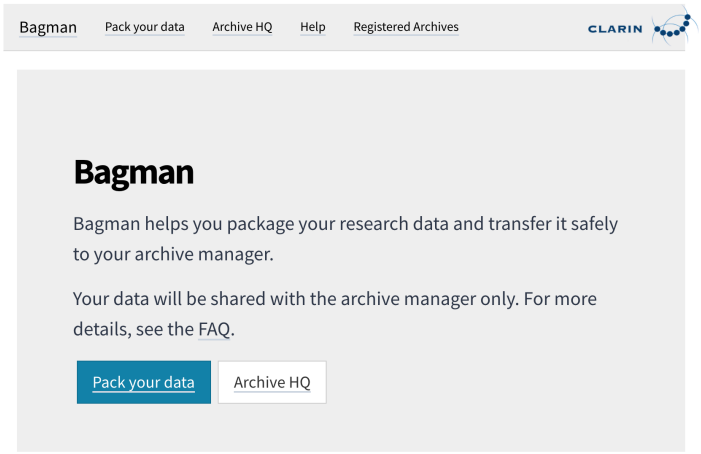


Figure 1: Bagman - Welcome Page.

[1] See `https://catalog.clarin.eu/ds/ComponentRegistry/#/`
[2] See `https://www.clarin-d.net/en/preparation/find-a-clarin-centre`.

Bagman aims at supporting researchers and archive managers alike. The software uses `Java` for the back-end and `react-js` for the front-end. Fig. 1 depicts the welcome page of Bagman; it gives access to its two core functionalities: "Pack your data" and "Archive HQ". The first functionality is targeted at researchers who want to archive their research data; the second one is aimed at archive managers to get access to the research data packages submitted by users. In this paper, we will focus on the first aspect. Fig. 2 depicts Bagman's user interface for collecting data from its users via simple forms.

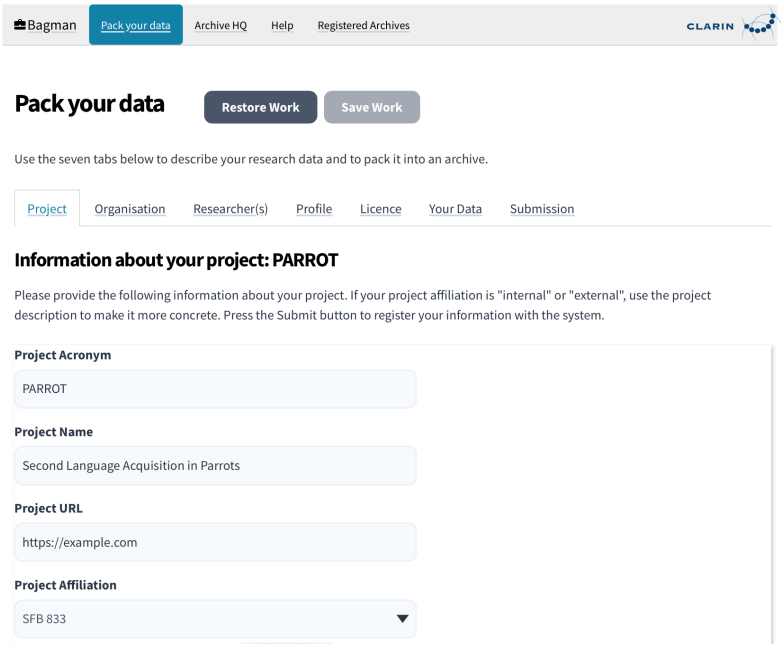


Figure 2: Bagman - Requesting Metadata.

Researchers are requested to describe their research data with respect to the project where it has been collected and the researchers and their organisations that were involved. Users then classify their data in terms of a resource type and by answering a number of targeted follow-up questions about the chosen type. In the fifth step, users can select a licence for their research data. In the sixth step, users can upload their data by selecting a directory from their file system, see top-left part of Fig. 3. Note that some icons in the resulting tree are highlighted in red to signal file formats not suitable for archiving. Here, users are encouraged to convert, say, proprietary file formats to non-proprietary ones, or to delete superfluous ones. Note that Bagman delegates the main task for organising directory structures to users' existing tools such as Finder (Mac OS), Explorer (Windows), or Files (Ubuntu), and file conversion software, say, Numbers, Excel, or OpenOffice. Once users have post-processed the directory tree, they can prepare the submission process (last step). Preparation includes the *automatic* generation of a CMDI file from known inputs as well as the submission package, the bag where all files all listed together with their checksums (see top-right and bottom part of Fig. 3). The back-end of Bagman takes care of all storage of research data, and it also implements basic functionality for CMDI generation. In detail, the back-end implements an API for (i) the generation of XML-based CMDI from JSON input, which is passed on from the client; (ii) the transferal of bags in ZIP format from client to server as well as methods for getting and deleting bags for archive management. Bagman also implements functionality for matching a bag with an archive that is best suited for hosting it.

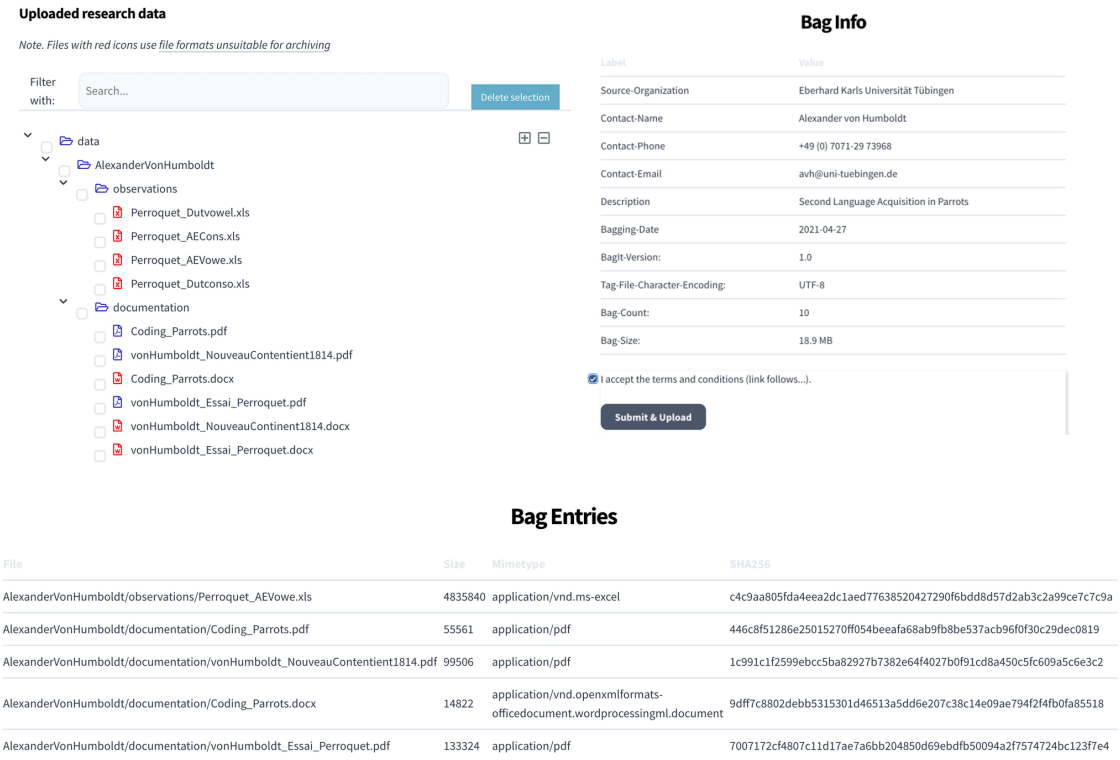


Figure 3: Bagman -Various Screenshots.

4 Current State and Future Work

We have built a prototype of Bagman that implements its core functionality and which is now open for beta testing at `https://weblicht.sfs.uni-tuebingen.de/bagman/`. We invite readers to explore the tool and encourage their feedback. At the time of writing, only a single archive has been connected to Bagman to test and validate the transfer of data between researchers and archive managers. With research data temporarily stored on Bagman's back-end, adding new archives to Bagman means giving their managers a login so that they can get access to the bags submitted to them. At the time of writing, Bagman supports the major resource types hosted by TALAR[3], the Tübingen Archive of Language Resources; our software hence allows the automatic instantiation of CMDI profiles for the description of lexical resources (LexicalResourceProfile), text corpora (TextCorpusProfile), speech corpora (SpeechCorpusProfile), tools (ToolProfile), and experiments (ExperimentProfile), all identifiable via the Group Name "NaLiDa" in the CMDI component registry. The use of these profiles ensures that the corresponding resources can be easily found using faceted browsing in the CLARIN Virtual Language Observatory, say, by searching the facets for language, collection, resource type, modality, or availability.[4]

The design of Bagman walks a fine line between researchers (often taking research data management as a necessary evil) and archive managers (taking it for something absolutely necessary, with an emphasis on "the more metadata the better"). When Bagman users, for instance, identify their data as a lexical resource, they are given the opportunity to specify the type of the lexicon (*e.g.*, dictionary, glossary, thesaurus), the type of the headword, and the subject language, but they may skip the step if they want to. Also, they can put more information about their resource in an open-ended lexicon description field

[3] See `https://talar.sfb833.uni-tuebingen.de`.
[4] See `https://vlo.clarin.eu`.

when they feel that more information need to be put somewhere. Note, however, that Bagman delegates any metadata-related issues to a subsequent one-to-one communication between researcher and archive manager. Metadata fields left open during a Bagman session can often be filled at a later stage when archive managers feel they require more information than researchers provided.

Bagman is browser-based software, and hence, special care needs to be taken to ensure that users can provide their input in a flexible, peace-wise manner. At any time, users can save the current session, that is, write-out all metadata that has been entered to their file system. At a later time, when users like to resume their work, they can then easily restore their session.

At the time of writing, Bagman is only connected to TALAR, but it supports all the archive's profiles. For TALAR users, Bagman has entered production mode. The feedback we obtain from these real-world users informs the further development of Bagman, strengthening its usability and stability. Once Bagman has matured, we will ask other archives whether they want to be connected to Bagman, and we will investigate how their archiving requirements can be meet with the software. Currently, it is too soon to speculate about the detailed implementation roadmap for the archiving' aspect of Bagman. It is clear that other archives will like to see their metadata profiles and archiving policies supported. Here, Bagman would need to adapt its front-end to collect information specific to the new profiles, and the back-end to generate ready-to-use and valid CMDI instances that other archives are happy to work with.

Bagman does not prescribe any guidelines on the granularity of the research data that needs to be archived. Each set of resources is different, and Bagman *per se* does not attempt to promote a *one-size-fits-all model*. For now, most users are unaware of Bagman. They contact the TALAR archive manager because they would like to have their resources deposited. Once the contact has been established and any open questions between the two parties addressed (*e.g.*, granularity or licence issues), the users are then explicitly directed to Bagman to build, describe, and submit their package to the archive via Bagman.

One important aspect to Bagman's usability is the packaging. When the archive manager is informed of a new bag being submitted via Bagman, he can download the bag from Bagman's "Archive HQ" GUI, unzip the bag and run BagIt software to verify that the package has been transfered in a complete and correct manner.[5] Our TALAR archive managers find this functionality very useful and reassuring indeed, and a necessary first step before looking into the CMDI, and contacting the researchers for any follow-ups, such as resolving metadata issues, or the drafting and signing of data depositing agreements.

Getting users to archive their research data is hard. Bagman offers users a single pit-stop approach to get their data archived with not too much hassle. Bagman helps users with metadata management as it generates a CMDI automatically from the information and research data supplied by the user. Given such data, Bagman then helps users to decide on an archive to host their resource, and then helps ensuring that all data is transfered to the archive in a complete and correct manner. In sum, Bagman fills-in a gap in the CLARIN infrastructure; its ease of use encourages users to get their data archived; and its automatic generation of CMDI from known inputs ensures the generation of expressive and high-quality metadata.

Acknowledgements

Our work was funded by the German Federal Ministry of Education and Research, the German Science Foundation (SFB-833), and CLARIN-D.

References

D. Broeder, M. Windhouwer, D. Van Uytvanck, T. Goosen, and T. Trippel. 2012. CMDI: a component metadata infrastructure. In *Describing LRs with metadata: towards flexibility and interoperability in the documentation of LR workshop programme*, volume 1.

Docuteam. 2018. Software – our tools for digital archives. Available at `https://www.docuteam.ch/en/products/it-for-archives/software/`.

J. Kunze, J. Littman, E. Madden, J. Scancella, and C. Adams. 2018. The bagit file packaging format (v1.0). Technical report, RFC 8493, DOI 10.17487/RFC8493, October. See `https://www.rfc-editor.org/info/rfc8493`.

[5] The command `python3 -m bagit --validate bag` verifies the bag.

Help Yourself from the Buffet: National Language Technology Infrastructure Initiative on CLARIN-IS

Anna Björk Nikulásdóttir[1], Þórunn Arnardóttir[2], Jón Guðnason[3], Þorsteinn Daði Gunnarsson[3], Anton Karl Ingason[2], Haukur Páll Jónsson[4], Hrafn Loftsson[3], Hulda Óladóttir[4], Einar Freyr Sigurðsson[5], Atli Þór Sigurgeirsson[6], Vésteinn Snæbjarnarson[4], and Steinþór Steingrímsson[5]

[1]Grammatek ehf., Iceland, [2]University of Iceland, [3]Reykjavik University, [4]Miðeind ehf., Iceland, [5]The Árni Magnússon Institute for Icelandic Studies, [6]University of Edinburgh
anna@grammatek.com, thar@hi.is, jg@ru.is, thorsteinng@ru.is, antoni@hi.is, haukurpj@mideind.is, hrafn@ru.is, hulda@mideind.is, einar.freyr.sigurdsson@arnastofnun.is, atlisigurgeirsson@gmail.com, vesteinn@mideind.is, steinthor.steingrimsson@arnastofnun.is

Abstract

In this paper, we describe how a fairly new CLARIN member is building a broad collection of national language resources for use in language technology (LT). As a CLARIN C-centre, CLARIN-IS is hosting metadata for various text and speech corpora, lexical resources, software packages and models. The providers of the resources are universities, institutions and private companies working on a national (Icelandic) LT infrastructure initiative.

1 Introduction

With the enormous progress in language technology (LT) in the last decades, the use of LT in research and commercial products has greatly increased. LT tools and resources are now not only used by LT specialists but also by researchers and developers from various fields. Beside the improvement in quality and usability, this development is driven by open access to data and software. For such resources to be of broad use, they need to be easily accessible and thoroughly documented. Thus, the large national LT infrastructure initiative *Language Technology Programme for Icelandic (LTPI) 2019–2023* (Nikulásdóttir et al., 2020) chose CLARIN-IS to be the central hub for all deliverables of the programme.

This paper gives a broad overview of the available repositories and the core publishing guidelines.

2 CLARIN-IS

Iceland became a CLARIN ERIC member on February 1, 2020 after having an observer status since November 1, 2018. The Árni Magnússon Institute for Icelandic Studies is the leading partner in the Icelandic national consortium. A Metadata Providing Centre (CLARIN C-centre[1]) has been established at the institute that hosts metadata for Icelandic language resources and distributes them through a Virtual Language Observatory.

As a new member, CLARIN-IS is in the process of establishing a technical Service Providing Centre (CLARIN B-centre), which will maintain language resources among other tasks. For now, we maintain a Gitlab[2], where all relevant GitHub repositories are mirrored, and deliver all resources to the C-centre.

[1]https://clarin.is/
[2]https://gitlab.com/icelandic-lt

3 Language Technology Programme for Icelandic

In October 2019, a consortium of Icelandic universities, companies and institutions (10 in total) started working on the LTPI. The programme aims at making Icelandic viable in future technologies that rely on LT in one way or another. To build foundations for that goal, the LTPI concentrates on developing language resources and infrastructure software, divided into six core project areas: 1) Language Resources, 2) Support Tools, 3) Machine Translation, 4) Spell and Grammar Checking, 5) Automatic Speech Recognition, and 6) Speech Synthesis.

During the preparation work on the LTPI, other European national programmes for LT were reviewed and information from experienced partners collected. Further information on related programmes and the general structure and execution of the LTPI can be found in (Nikulásdóttir et al., 2020).

All deliverables of the programme are published under open licenses and are freely accessible for research as well as commercial use. Therefore, it is of utmost importance to have a stable hosting platform that can ensure access and availability.

3.1 Language Resources

A variety of language resources are being compiled or extended within the LTPI. The **Icelandic Gigaword Corpus** (IGC) (Steingrímsson et al., 2018) is a large text corpus containing over 1.6B tokens. Within the LTPI, the corpus is being updated yearly with new data sources and updated data from previous years. Furthermore, each new edition is annotated using the latest tools. Two versions of a Gold standard corpus, **MIM-GOLD** and **MIM-GOLD-NER**, have been made available, manually annotated with POS tags and named entities, and within the LTPI enriched with manually checked lemmas. Large lexical resources have been redesigned or are being extended and further processed with focus on use in LT: the **Database of Icelandic Morphology** (Bjarnadóttir et al., 2019) and the **Icelandic Word Web** (Daníelsson et al., 2021). Smaller resources, like a hyphenation word list with a hyphenation tool, and a pronunciation dictionary, have also been made available through CLARIN-IS.

3.2 Support Tools

Several NLP tools have been or are currently being developed or improved upon within the LTPI. Each tool is either used as part of a processing pipeline, or as a stand-alone tool. A **tokenizer** has been developed that converts input text to streams of tokens and also segments the token stream into sentences, considering various cases of abbreviations, dates, etc. to prevent wrong segmentation. During the LTPI, a previously published BiLSTM **PoS tagger** for Icelandic (Steingrímsson et al., 2019) has been improved substantially, e.g. by incorporating contextualized word embeddings, resulting in ABLTagger 2.0 in CLARIN-IS with an accuracy of 96.95%. With resources from Section 3.1, a RNN **lemmatizer** accepting the word-form as well as the corresponding PoS tag to predict the lemma is being developed within the LTPI. Latest experiments show an accuracy of 98.9% on known word-forms and 91.7% on unknown word-forms. A **named entity recognizer** based on a fine-tuned ELECTRA-Base model, trained on the IGC, is in development, newest experiments showing an F_1-score of 91.9%, a dramatic improvement from previous models (Ingólfsdóttir et al., 2020). Two previously published **parsers** have been updated within the LTPI, a *full parser* and a *shallow parser*. The rule-based full-constituency parser relies on a wide-coverage context-free grammar and uses a parsing system based on an enhanced Earley parser (Þorsteinsson et al., 2019). The work on the shallow parser (Loftsson and Rögnvaldsson, 2007) consists of making it accept tagged text according to the new MIM-GOLD tagset (see Section 3.1) and improving individual components. A new **lexicon acquisition tool** is used to find neologisms and older words that jump in frequency due to gaining new word senses. All the above tools are currently available through CLARIN-IS and by the end of the LTPI further tools will have been added, thus ensuring open access to the most important basic support tools for LT.

3.3 Machine Translation

Within the machine translation project, the substantial parts are corpus work, translation methods and infrastructure. A collection of parallel English-Icelandic corpora, **ParIce** (Barkarson and Steingrímsson,

2019), contains texts from various sources, most substantially European Economy Area regulations. **Backtranslations** are synthetic parallel corpora created using existing translation systems that have been shown to be greatly beneficial when training neural translation models. ParIce, including development and test sets, and backtranslated corpora have been published on CLARIN-IS. Three different machine translation methods were tried and tested in the first year (Jónsson et al., 2020) to compare traditional methods to recent advances using the available data – the models have been made available on CLARIN-IS. A **Transformer model** showed best results and thus transformers were chosen as the core method for an open Icelandic-English translation system. A **web-based translation interface** was created and set up online to compare the different models, along with translations provided by Google. This served as a way to compare translations between the participating organizations and allows for open discussion about evaluation. Good **model serving infrastructure** is important when sharing translation models based around different methods. The code for the website as well as the code and configurations to deploy and run translations is made available on CLARIN-IS.

3.4 Spell and Grammar Checking

The work in this core project has focused on developing the necessary data and tools for detecting, categorizing and correcting errors for different user groups. The following resources are currently available through CLARIN-IS: An annotated **general error corpus** with a fine-grained error classification that facilitates performance measurements of the spell and grammar checking software (Arnardóttir et al., 2021). Three **specialized error corpora**, each representing a particular user group, have been annotated and published in order to measure the software's performance on errors particular to the respective user groups. The Icelandic L2 Error Corpus is a collection of texts written by second-language learners of Icelandic (Glišić and Ingason, 2021), the Icelandic Dyslexia Error Corpus is a collection of texts written by native Icelandic speakers with dyslexia, and the Icelandic Child Language Error Corpus is a collection of texts written by native Icelandic speakers aged 10 to 15. **Miscellaneous word lists and language models** include aggregated error data from different sources, a confusion sets database and a trigram language model to help with suggestions for corrections. The **spell and grammar checking software** is a Python package and command line tool for checking and correcting spelling and grammar. The version currently available on CLARIN-IS offers token-level correction and some grammar correction. To get the most usable and complete product for the largest user group, the current focus is on grammar errors, error correction in general, and more detailed guidance tailored to different user groups.

3.5 Automatic Speech Recognition

The emphasis of the Automatic Speech Recognition (ASR) project within the LTPI has been data collection, publication of quality ASR recipes and ultimately a support for commercial applications depending on ASR. The data collection effort has many facets. The prompt-based data collection effort was revived through a new crowd-sourcing system based on Mozilla's Common Voice project[3] called **Samrómur** (Mollberg et al., 2020). The Samrómur project continues to collect voice samples from adults but it also reaches out to children, teenagers and people who speak Icelandic as a second language. Transcriptions of broadcast news and media material as well as university lectures are also being produced to create parallel acoustic-text databases for Icelandic ASR. A system to collect prompted questions for Question Answering systems and conversations for spoken dialogue systems have also been set up. All these **speech data collections** are being prepared for publication on CLARIN-IS. Kaldi[4] recipes have been developed for general-purpose speech recognition and for teenage voices. Furthermore, **punctuation models** have been published on CLARIN-IS.

3.6 Speech Synthesis

For speech synthesis (TTS) it is important to have access to a large corpus of high quality recordings. Currently available on CLARIN-IS is the **Talrómur** corpus which includes 213 hours of speech recordings from eight different speakers. The corpus consists of four male voices and four female voices. The

[3] https://commonvoice.mozilla.org/en

[4] https://github.com/cadia-lvl/samromur-asr

voices range in age, from 26 to 71 years old, and speaking style. In total, the corpus is made up of 122,417 single sentence utterances. The reading script was generated to maximize coverage of diphones in the Icelandic language and consists of sentences from multiple different sources (Sigurgeirsson et al., 2020; Sigurgeirsson et al., 2021). The recordings were conducted in 2020 by Reykjavik University and RÚV, the Icelandic National Broadcasting Service, in a professional studio at the headquarters of the latter. Two of the voices were recruited from the north of Iceland and were recorded in a studio at the University of Akureyri. Later this year, **Talrómur 2** will be available on CLARIN-IS with additional two hours of recordings from each of 40 new speakers. For TTS text pre-processing, data, software packages and models for **text normalization** and **automatic grapheme-to-phoneme** (g2p) conversion are already published on CLARIN-IS or will shortly be available.

4 Standards and Licensing

One of the core pillars of the LTPI is the publication of data and software under open licenses. The guiding licenses are CC BY 4.0[5] for data and Apache 2.0[6] for software. In exceptional cases, data have to be published with more restrictive licenses, but all deliverables of the programme will be available for research and commercial use. An important part of ensuring open licensing is the crafting of agreements and consent statements for various data collection efforts.

All teams work by common standards, defined in guidelines for data deliverables, on the one hand, and for software deliverables, on the other. Wherever possible, the guidelines adhere to international standards, e.g. regarding data format, metadata, or coding guidelines. Published data adhere to the FAIR standard[7]. Naming, versioning and keyword definitions are coordinated throughout the deliverables.

Type of Repository	**Number of Repositories**
General text corpora, incl. test/dev	7
Specialized corpora	8
Parallel corpora	4
Lexical resources	7
NLP-tools	7
Machine translation	4
Spell and grammar checking	8
Speech corpora	2
Speech models and related modules	6
ALL REPOSITORIES	47

Table 1: CLARIN repositories from the LT-Programme for Icelandic. Status as of August 2021

5 Usage Scenarios

The aim of the LTPI is that language resources and infrastructure software will be available for research and commercial use. The aimed-at users are LT-specialists and general software developers that need to integrate LT in their products, as well as researchers from various fields.

There are numerous usage scenarios for the "buffet" of the LTPI deliverables. There are several levels of usage possibilities, reaching from low-level development using corpora and basic tools, to the usage of production-ready models or plugins/applications. For speech synthesis, for example, developers can use the speech corpora and necessary language-specific resources, like the pronunciation dictionary, to train and develop their own TTS models and voices. They can use the delivered TTS voices to integrate into their application, or they can use the web reader plugin directly to connect to their website.

[5] http://creativecommons.org/licenses/by/4.0/
[6] https://www.apache.org/licenses/LICENSE-2.0
[7] https://www.go-fair.org/fair-principles/

Acknowledgements

This project was funded by the Language Technology Programme for Icelandic 2019-2023. The programme, which is managed and coordinated by Almannarómur (https://almannaromur.is/), is funded by the Icelandic Ministry of Education, Science and Culture.

References

Þórunn Arnardóttir, Xindan Xu, Dagbjört Guðmundsdóttir, Lilja Björk Stefánsdóttir, and Anton Karl Ingason. 2021. Creating an Error Corpus: Annotation and Applicability. In *Proceedings of CLARIN*.

Starkaður Barkarson and Steinþór Steingrímsson. 2019. Compiling and filtering ParIce: An English-Icelandic parallel corpus. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 140–145, Turku, Finland.

Kristín Bjarnadóttir, Kristín Ingibjörg Hlynsdóttir, and Steinþór Steingrímsson. 2019. DIM: The Database of Icelandic Morphology. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 146–154, Turku, Finland.

Hjalti Daníelsson, Jón Hilmar Jónsson, Þórður Arnar Árnason, Alec Shaw, Einar Freyr Sigurðsson, and Steinþór Steingrímsson. 2021. The Icelandic Word Web: A language technology focused redesign of a lexicosemantic database. In *Proceedings of NODALIDA 2021*, pages 429–434, Reykjavík.

Isidora Glišić and Anton Karl Ingason. 2021. The nature of Icelandic as a second language: An insight from the learner error corpus for Icelandic. In *Proceedings of CLARIN*.

Svanhvít L. Ingólfsdóttir, Ásmundur A. Guðjónsson, and Hrafn Loftsson. 2020. Named Entity Recognition for Icelandic: Annotated Corpus and Models. In Luis Espinosa-Anke, Carlos Martín-Vide, and Irena Spasić, editors, *Statistical Language and Speech Processing*, pages 46–57, Cham. Springer International Publishing.

Haukur Páll Jónsson, Haukur Barri Símonarson, Vésteinn Snæbjarnarson, Steinþór Steingrímsson, and Hrafn Loftsson. 2020. Experimenting with Different Machine Translation Models in Medium-Resource Settings. In Petr Sojka, Ivan Kopeček, Karel Pala, and Aleš Horák, editors, *Text, Speech, and Dialogue*, pages 95–103, Cham. Springer International Publishing.

Hrafn Loftsson and Eiríkur Rögnvaldsson. 2007. IceParser: An Incremental Finite-State Parser for Icelandic. In *Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA 2007)*, pages 128–135.

David Erik Mollberg, Ólafur Helgi Jónsson, Sunneva Þorsteinsdóttir, Steinþór Steingrímsson, Eydís Huld Magnúsdóttir, and Jón Guðnason. 2020. Samrómur: Crowd-sourcing Data Collection for Icelandic Speech Recognition. In *Proceedings of the 12th Conference on Language Resources and Evaluation*, pages 3463–3467, Marseille, France.

Anna Björk Nikulásdóttir, Jón Guðnason, Anton Karl Ingason, Hrafn Loftsson, Eiríkur Rögnvaldsson, Einar Freyr Sigurðsson, and Steinþór Steingrímsson. 2020. Language Technology Programme for Icelandic 2019–2023. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3414–3422, Marseille, France.

Atli Sigurgeirsson, Gunnar Örnólfsson, and Jón Guðnason. 2020. Manual Speech Synthesis Data Acquisition – From Script Design to Recording Speech. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 316–320.

Atli Sigurgeirsson, Þorsteinn Gunnarsson, Gunnar Örnólfsson, Eydís Huld Magnúsdóttir, Ragnheiður Kr. Þórhallsdóttir, Stefán Jónsson, and Jón Guðnason. 2021. Talrómur: A large Icelandic TTS corpus. In *Proceedings of NODALIDA 2021*, pages 440–444, Reykjavík.

Steinþór Steingrímsson, Sigrún Helgadóttir, Eiríkur Rögnvaldsson, Starkaður Barkarson, and Jón Guðnason. 2018. Risamálheild: A Very Large Icelandic Text Corpus. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*, Miyazaki, Japan.

Steinþór Steingrímsson, Örvar Kárason, and Hrafn Loftsson. 2019. Augmenting a BiLSTM tagger with a morphological lexicon and a lexical category identification step. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1161–1168, Varna, Burgaria.

Vilhjálmur Þorsteinsson, Hulda Óladóttir, and Hrafn Loftsson. 2019. A Wide-Coverage Context-Free Grammar for Icelandic and an Accompanying Parsing System. In *Proceedings of the International Conference on Recent Advances in Natural Language Proceedings (RANLP 2019)*, pages 1397–1404, Varna, Bulgaria.

CLARIN-IT Resources in CLARIN ERIC - a Bird's-Eye View

Dario Del Fante
ILC-CNR - Italy
dario.delfante@ilc.cnr.it

Francesca Frontini
ILC-CNR - Italy
francesca.frontini@ilc.cnr.it

Monica Monachini
ILC-CNR - Italy
monica.monachini@ilc.cnr.it

Valeria Quochi
ILC-CNR - Italy
valeria.quochi@ilc.cnr.it

Abstract

This paper investigates the visibility of CLARIN-IT language resources within the services of the CLARIN ERIC central infrastructure, notably the Virtual Language Observatory, the Switchboard and the Federated Content Search, from a user perspective in order to identify possible issues. While the experiment focused on one national consortium, the ultimate goal is to develop an assessment methodology that can be used by any national consortia aiming to review the accessibility of their resources and tools within the CLARIN central services.

1 Introduction

With a distributed network of over 50 centres, CLARIN ERIC's principal aim is to ensure easy access to their resources and tools by researchers from all over Europe and beyond, independently of their original producers, of the centre or consortium physically hosting them. Ideally, a researcher should not need to know where a given resource is deposited or even be aware of its existence to be able to find, access and use it, thanks to the central functionalities and services available via the CLARIN portal, which is a gateway to the whole network's offerings.

The first and foremost central service, the CLARIN *shop window*, is the Virtual Language Observatory (VLO)[1] (Broeder et al., 2010) which harvests metadata from all the official CLARIN data providing centres and makes them searchable via a unified interface offering faceted search. Other interesting and useful central discovery services are the Federated Content Search (FCS)[2], the Language Resources Switchboard (SB)[3], and the CLARIN Resource Families[4].

In order to ensure visibility from the CLARIN central services, we argue that national consortia should monitor regularly these four "points of access" and analyse them from a user perspective. We are not talking about automated checks, but rather of a more qualitative assessment aimed at ensuring that any researcher/end-user can easily find the resources she needs and use them as intended. As pointed out by Sugimoto (2016), despite the wide array of useful services for digital research in linguistics and the humanities, it is unclear whether the community is thoroughly aware of the status-quo of the growing infrastructure. At the same time, such an analysis could provide useful instruments in the hands of national coordinators and center managers for bringing to the fore strengths and critical issues of their data providing community.

For these reasons we set out to check and analyse the presence of the LRs available in the CLARIN-IT consortium in the central discovery services with the twofold aim of (i) assessing the Italian consortium presence and of (ii) devising a reproducible qualitative methodology from the user perspective. In this paper, we will present a case-study aimed at investigating the visibility, reliability and searchability of CLARIN-IT LRs in the VLO. In doing so, we shall sketch a proposal for generalising this qualitative assessment procedure to any given consortium.

[1] https://vlo.clarin.eu
[2] https://contentsearch.clarin.eu/
[3] https://switchboard.clarin.eu/
[4] https://www.clarin.eu/resource-families

2 CLARIN-IT in the VLO

CLARIN-IT is actively involved in the sectors of documentation, digitization and language technologies for the Humanities and it is focused on Language Resources (LRs) both data and tools (Monachini and Frontini 2016,). Currently the consortium offers two data centres, ILC4CLARIN[5], the national B centre, and the EURAC Research CLARIN centre (ERCC)[6]; both host repositories for the preservation of LRs, which thus become visible from the Virtual Language Observatory (VLO) [7]. The VLO represents the principal means of exploring LRs in CLARIN. By using a facet browser that allows for the filtering of metadata records according to previously specified categories - the facets - the VLO makes it possible to carry out targeted searches.

However, given the variability of the CMDI metadata framework (Haaf et al., 2014), the VLO represents an improvable asset. In this work we rely on and extend a previous attempt to assess the searcheability of tools (Odijk, 2019; Odijk, 2014).

The CLARIN-IT resources in the VLO can be easily extracted thanks to a faceted query that uses the national project as filter[8]. The search query returns 490 different LRs, of which 51 are hidden because of duplicate naming, which leaves 439 distinct resources as shown in Table 1. The presence of duplicate naming represents a quite common occurrence in the VLO search. The VLO browser automatically removes all the duplicates from the search results on the basis of the naming. However, the presence of duplicates is specified under each problematic record.

CLARIN-IT - a birds eye view	
Total Number of LR	439
Monolingual	388
Multilingual resources	46
Format	12
Languages	10
Organisations	8
Collections	7
Resource type	6
Data providers	2

Table 1. CLARIN-IT on VLO

This basic query can be further narrowed down by using other facets, which allows not only for a systematic classification of the national resources by language, resource type, collection, but also for the verification of other metadata such as subject, format, availability, and so on. The aim is to determine the extent to which the resources are correctly described, in particular by verifying the difference between the number of LRs which are expected to occur and the actual number of LRs under each facet. The idea is to explore and test an assessment procedure that may assist repository managers, national coordinators or even the central office in harmonising the content of each repository and consequently of the VLO. In what follows we discuss the results of some of the most interesting filters.

2.1 Languages of CLARIN-IT

Filtering by *Language* identified ten different languages as indicated in Table 2.

These results show that CLARIN-IT offers LRs in a variety of languages, not only Italian. Due to the specialisation of the ILC4CLARIN, Latin and Ancient Greek are particularly represented. However, at a closer look, the over-representation of Latin LRs even with respect to Italian ones is due to the choice of metadata description of the ALIM corpus, which reflects the internal organisation of the original archive.

[5] https://ilc4clarin.ilc.cnr.it/

[6] https://clarin.eurac.edu/

[7] https://www.clarin.eu/content/virtual-language-observatory-vlo

[8] https://vlo.clarin.eu/search?fqType=nationalProject:or&fq=nationalProject:CLARIN-IT

Languages			
Latin	366	Italian	30
English	40	German	8
Arabic	32	Czech	2
Ancient Greek (to 1453)	6	Breton	1
Ancient Greek	8	Basque	1

Table 2. Languages in CLARIN-IT

As described in (Boschetti et al., 2020), every text of that corpus is deposited as a separate resource, while this is not true for other corpora. Such a finding may indicate a need for harmonization of the depositing guidelines for specific resource types. *Thus, the store of ALIM corpus in such a way, might be problematic in terms of availability and search-ability because its structures lacks of interconnections.*

2.2 Organisations and Collections of CLARIN-IT

By checking the filter results of the *Organisation* and *Collections* facets we can easily verify that the organisations and consortium members are actively contributing to the repository and correctly represented in the VLO. Table 3 indicates the number of LRs which each organisation is responsible for:

Organisations			
Archivio della Latinità Italiana del Medioevo (ALIM)	354	CIRCSE - Università Cattolica Sacro Cuore	8
Istituto di Linguistica Computazionale - CNR	39	Ghent Universities	2
Institute for Applied Linguistic Research - EURAC	9	Università di Parma	2
Università di Salerno	8	Basque	1

Table 3. Organisations in CLARIN-IT

This query confirms that almost all of the Latin LRs are indeed items of the ALIM collection, but also that most of the CLARIN-IT LRs are produced by Italian consortium members. Looking at Table 3, only two LRs deposited in CLARIN-IT centres were produced by a foreign institution (Ghent University and Basque).

Collections			
ALIM Literary Sources	344	ILC4CLARIN : OPEN Data and Tools	7
ILC4CLARIN	54	ERCC Learner Corpora	8
Alim Documentary Sources	11	ERCC Web Corpora	4
CIRCSE	8	ERCC	1

Table 4. Collections in CLARIN-IT

2.3 Resource Types and Data Providers of CLARIN-IT

By combining *Resource type* and *Data Provider* it is possible to check the number and types of resources offered by each of the two CLARIN-IT centers, and thus get some information on their specialisation, as appears in Table 5.

2.4 CLARIN-IT Formats and Subjects

A very useful query concerns the available Formats and Subjects for CLARIN-IT resources. By checking this we assess whether all resources are correctly typed and whether they have been further described with suitable and harmonised subject keywords. In the Italian case, the coverage for the latter seems to

ILC4CLARIN		**Eurac Research**	
Corpus	368	Corpus	13
Lexical Resource	43		
Software, webservice	12		
Webservice	2		
Text	1		

Table 5. Resource type for each Data provider

be incomplete (only 18 LRs are mapped onto VLO subjects keywords, whereas many of the keywords present in the national repositories are not visible in the VLO) and harmonisation could be increased by using controlled vocabularies.

2.5 CLARIN-IT LRs availability

One important final check concerns Availability, which indicates "degree to which resources and tools are publicly accessible". In the case of CLARIN-IT, most of the LRs are publicly available; however, the filter also returned 25 resources with unspecified availability. A closer inspection shows that these correspond to corpora from the ERCC repository and webservices from ILC4CLARIN. This finding might lead to amendments of the records.

3 The Methodology

During this work, we assessed the presence of two unexpected issues which gave us the opportunity of learning two relevant lessons in terms sustainability and usability of our methodology.

3.1 Two Lessons

- *Granularity*
  Regarding the CLARIN-IT consortium, we found that there are some cases when LRs are stored in the VLO as single records belonging to the same collection. This is due to the fact that some repository does not support a nested archive. For example, the collection *ALIM Lietrary Sources*, does not appear in the VLO as a single entry. It is stored as a collection of different entries categorised as corpora despite the fact that are composed of only one single texts. As a result, the high level of granularity in relation to data storage might work against the availability and the accessibility of the resources caused by the dispersive collection methods. However, different levels of granularity respectively correspond to different analytical possibilities. A nested archive represent a more solid sample of data which can be representative of a specific language and might give different options for analysis by using the Switchboard. Differently, a non-nested archive represents a more agile option for comparing records from different collections and national consortia by saving these records as virtual collections.

- *Duplicate namings*
  The second issue encountered is related to the presence of duplicates which are automatically removed from the results produced by the search in the VLO. This issue has attracted our attention and after a careful examination, most of them resulted in being false duplicates. Within the *ALIM Literary Sources* for instance, all 50 (out of 394) hidden items, are in fact different critical editions of the same texts by different editors. Having the exact same title, the system considers them as duplicates. For example the *Summa Dictaminis* corresponds to three records, one for each editor (Matteo de'Libri , M. Thumser, Emil Polak). While a possible strategy for avoiding such texts to be treated as duplicates could be to add "by EDITOR" in the 'title' metadata, this conflicts with the collection praxis. This may be another issue for discussion for the Standing Committee on CLARIN

Technical Centres. In order to enhance the availability and the accessibility of the data within the VLO, we propose that greater attention should be payed in relation to these aspects.

3.2 Extending the methodology

Based on the results and observations stemming from the analysis we have just described, a more general methodology for qualitative assessment can be thus generalised. We suggest the following checks should be carried out by new consortia, after the registration of at least one B or C centre, but also, periodically, buy existing national consortia, especially when new centres are registered or large collections are injected.

1. Select in the National Project tab, the project of interest in order to select only the LRs which are provided by CLARIN centre of the national consortium
2. Check which LRs are shown and how are presented in the VLO, filtering for: (a) Languages, (b) Organisations and Collections, (c) Resource type
3. Check the presence of duplicates
4. Check the status of activation for a sample of links to the original place
5. Register all the inconsistencies in terms of accessibility and availability

4 Conclusions

With the growth of the CLARIN-IT consortium, a thorough check of the LRs contributed by various partners across two repositories has become necessary. This exercise of qualitative assessment of the visibility of the consortium resources in the VLO has proven extremely useful and might become a model for other projects. To this end, more checks could be added, including a template search for important resources such as reference corpora and lexicons, to ensure that they correctly appear as expected.

References

Boschetti, F., Del Gratta, R., Monachini, M., Buzzoni, M., Monella, P., and Rosselli Del Turco, R. 2020. "Tea for Two": The Archive of the Italian Latinity of the Middle Ages meets the CLARIN Infrastructure. In C. Navarretta and M.Eskevich (Eds.) *Proceedings of CLARIN Annual Conference 2020.* . Virtual Edition.

Haaf, S., Fankhauser, P., Trippel, T., Eckart, K., Hedeland, H., Herold, A., Knappen, J., Schiel, F., Stegmann, J., and Van Uytvanck, D. 2014. CLARIN's virtual language observatory (VLO) under scrutiny - the VLO taskforce of the CLARIN-D centres. In *CLARIN annual conference 2014*.

Monachini, M. and Frontini, F. 2016 CLARIN, l'infrastruttura europea delle risorse linguistiche per le scienze umane e sociali e il suo network italiano CLARIN-IT. *IJCoL - Italian Journal of Computational Linguistics*, 2(2),11-30.

Odijk, J. 2014. *Discovering Resources in CLARIN: Problems and Suggestions for Solutions.* Unpublished paper.

Odijk, J. 2019. *Discovering software resources in CLARIN*. In *Selected Papers Clarin Conference 2018*, 159: 121-132.

Sugimoto,G. 2016 Number game -Experience of a European research infrastructure (CLARIN) for the analysis of web traffic. In *CLARIN Annual Conference 2016*,ArXiv: abs/1706.05089.

Broeder, D., Kemps-Snijders, M., Van Uytvanck, D., Windhouwer, M., Withers, P., Wittenburg, P., and Zinn, C. 2010. A Data Category Registry and Component-based Metadata Framework. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA).

A data repository for the management of dynamic linguistic datasets

Thomas Gaillat
LIDILE - EA 3874
University of Rennes, France
thomas.gaillat@univ-rennes2.fr

Leonardo Contreras Roa
LIDILE - EA 3874
University of Rennes, France
lcontrerasroa@gmail.com

Juvénal Attoumbre
LIDILE - EA 3874
University of Rennes, France
juvenalak@gmail.com

Abstract

This paper addresses the issue of using Nakala, a dynamic database technology, for the management of language corpora. We present our ongoing attempt at storing and classifying multimedia documents of a corpus of language learner oral and written productions with universal resource identifiers. The architecture supports query APIs compatible with R packages and other tools which will facilitate the generation of linguistically enriched datasets for a more effective corpus-based study of language acquisition.

1 Introduction

For several decades, many corpora have been made public. The open-science initiatives which insist on the importance of making data Findable, Accessible, Interoperable and Reusable (*FAIR* (Wilkinson et al., 2016)) further reinforce this trend. However, even if it is important to make corpora public, it should not be the sole objective of the move towards more versatile and flexible data. It is equally important to think about the way corpora are made accessible. Current trends show that most shared corpora are static (Sérasset et al., 2009) in the sense that once downloaded, the data are not updated and run the risk of obsolescence.

When research experiments are conducted on external data, the first stage usually consists in data download prior to pre-processing (curation, part-of-speech (POS) tagging, etc.) Subsequent stages depend on the initial data and all pre-processing and analytical tasks rely on the same data as initially downloaded. However, if, in the meantime, corpus authors have modified their data by updating or correcting observations, experiments automatically rely on out-of-date data. The problem is to find a method to include the most up-to-date corpora as part of the pre-processing stage of experiments.

One way to approach the problem could consist in designing a data architecture ensuring a seamless workflow from the data ingestion to the querying stage. In this paper we present a use case based on the exploitation of a learner corpus stored in a database.

2 Current corpus sharing practices

Over the last few decades many corpora have been developed and a gradual shift towards public access has been observed. In the last decade, different platforms have been set up to make corpora available to the research community (ELRC, META-SHARE, among others). CLARIN illustrates the case as it includes access to data repositories. These repositories provide resources such as corpora which rely on persistent URLs. For instance, Huma-Num Ortolang (Pierrel, 2014), the French platform for language resources, includes a number of resources which are downloadable thanks to such URLs. As far as we know, most corpora are available under single URLs. Other infrastructures also exist such as ELRA's[1] catalogue. As of 2013, the repository provides open access to two sets of non-commercial language resources [2] classified according to a number of criteria that form metadata. As accessible as they are,

[1] The European Language and Resources Association

[2] http://portal.elda.org/en/catalogues/free-resources/

these resources are provided at corpus level rather than observed subject level. It is not possible to update or retrieve specific items from the datasets.

In parallel, the European Language Equality consortium (Rehm et al., 2020) has launched a project to establish the European Language Grid in which language resources of many types will be interconnected. Data sets, tools, models will be interoperable to allow the construction of natural language processing (NLP) pipelines dedicated to specific language tasks. There is a clear move towards an integrated set of resources and services. To support this infrastructure, one essential point is to make corpus items accessible in their up-to-date version. Persistent data sources will support this process by making corpus items queriable and retrievable.

However, the current type of corpus distribution is done mostly via platforms that provide static recordings of data in the sense that, even if the data is updated in the back end, the downloadable version remains in its initial state until a new static version is ready. We propose a dynamic solution to store and query a corpus whereby the corpus architecture gives controlled access to corpus items via APIs, while a robust online repository – in our case the Nakala database (Huma-Num, 2021)– ensures long-term storage and allows constant updates and maintenance.

The data workflow implemented by Nakala is illustrated in Figure 1 below. This workflow is being tested for the upload of our own multimedia corpus of the written and spoken productions of learners of English and French as foreign languages.

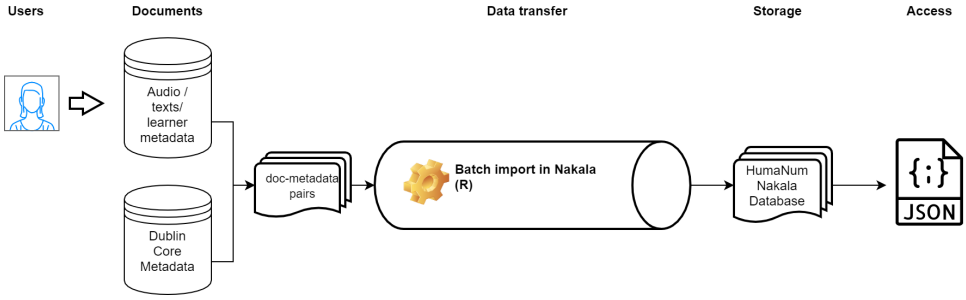


Figure 1: Nakala data workflow

3 Use case: A data repository for a learner corpus

3.1 Storing a learner corpus in a data repository

Making up-to-date corpus data constantly available implies the use of database technologies. This is because, as opposed to static file collections, database management systems allow queries both for augmenting and retrieving data.

We are currently developing a dynamic database management system for the CIL corpus (*Corpus InterLangue*), a collection of recordings and written productions of learners of English and French as a foreign language (L2). The corpus has been collected for the past fifteen years by Master students of the Linguistics and Didactics program of the University of Rennes 2 in France , following roughly the same data collection protocol (Arbach, 2015, p. 239). The outcome of this protocol is a complex set of learner data, composed of the following files for each of the surveyed learners: 1. A recording of the learner being interviewed by a native speaker of the L2 (.WAV) 2. An orthographic transcription of the interview (.CHA/.EAF[3]) 3. A recording of the learner reading a text in the L2 (.WAV) 4. A page-long handwritten production made by the learner in the L2 (.PDF) 5. A digitised transcription of the handwritten production (.TXT) 6. A sociolinguistic questionnaire filled in by the learner (.PDF/.CSV) 7. A consent form signed by the learner (.PDF).

[3]CHA: A proprietary plain-text format for single-tier text-to-audio alignment generated by the software CLAN.
EAF: An XML-based format for multiple-tier, time-aligned transcriptions generated by the software ELAN.

All of these files are stocked in one sub-directory of the corpus, each sub-directory corresponding to a single speaker. As of April 2021, a total of 115 speakers compose the CIL corpus. Each year, a new generation of recordings is added to the corpus, which demands for a system of dynamically updated storage.

The data can be accessed in two different ways. Nakala provides a web interface through which users can browse the data and its directories. Figure 2 illustrates how the user can play an audio file, access the transcription (.CHA file) or download the written production (.PDF file) of a learner whose ID is *fre_al_tr_99_m_20*. The persistent link specific to the audio file is displayed underneath the audio player and the persistent link for the entirety of the data of that specific learner is available on the top-left corner of the browser. The same data can also be accessed through APIs and queried/filtered. A metadata file (.CSV) is included in each sub-directory to keep track of the number and type of files in the corpus, and to serve as search-word tag information for future queries. Once uploaded, Nakala assigns a persistent URL and a DOI to each file or each collection of corpus items, allowing full corpus querying. However, to date, Nakala does not support global-state versioning. This structure allows authors to feed the corpus and users to access it from the same back-end.

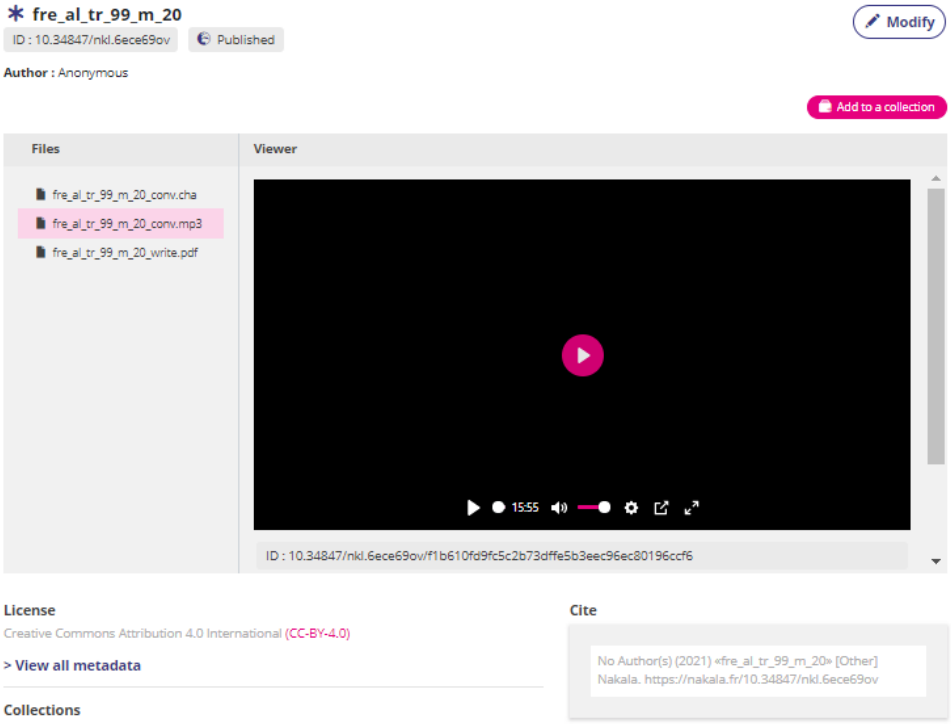


Figure 2: Example of data display on the web layout of the Nakala database

3.2 Generating Linguistic datasets

As well as providing up-to-date data, our purpose is to let researchers select which type of linguistic annotation they want to include in their datasets. Language resource users may not be interested in certain linguistic features while others might be of paramount importance. Here again, flexibility can provide an attractive answer.

We provide modular R programs[4] that make it possible to merge the data ingestion, curation and querying phases with that of linguistic annotation. These programs can be run in an IDE such as R Studio and automate on-demand creation of datasets containing linguistic data as well as learner metadata, i.e.

[4] https://github.com/LIDILE/CIL_query

age of the learner, their level of education, their age of first contact with the L2. Furthermore, in order to analyse aspects of syntax, POS-Tagging adapted to learners can be performed via UDpipe learner model (Wijffels, 2020). Syntactic parsing can also be conducted by UDPipe. In addition, a number of semantic features such as aspect and gender can also be of interest and obtained with this package. By running these scripts, users query Nakala repository via their IDE and generate datasets locally. They can then turn to modelling tasks.

An analogue process could be applied to process oral data from a phonetic/phonological point of view. Automatic text-to-phoneme alignment can be achieved, for example, by attaching the Nakala query pipeline to that of BAS WebMaus (Kisler et al., 2017). Vowel-formant extraction can also be envisaged via the emuR package for speech database management (Winkelmann et al., 2021). Even though the annotation and extraction provided by the two aforementioned methods usually require manual correction, especially if applied to learner speech, they can provide a rich amount of readily available data to automatically obtain a quick first glance at a learner's pronunciation.

4 Discussion and perspectives

This project is a first step in the direction of making a language learner corpus dynamically available. The Human-Num Nakala architecture provides persistent identifiers for single items of the corpus. Researchers can initiate queries via APIs in order to retrieve relevant data and metadata from the corpus.

Further work involves the development of customised R scripts and datasets according to the specific needs of researchers: semantic, morphosyntactic, phonological and phonetic annotations can be performed by branching our query pipeline with specific R scripts and packages aimed at enriching linguistic data. Various state-of-the-art annotation tools may be exploited depending on the research questions. For instance, in the case of morphosyntactic POS tagger comparisons, it could be necessary to use a dataset including tokens which have been POS-tagged with several tools and different tagsets. The resulting enriched linguistic data can be exploited to carry out research on the interlanguage of learners of English and French, which can in turn help develop didactic tools adapted to learners' specific needs. In our use case, we seek to extract the transcripts of interviews conducted with learners of French in order to model their proficiency according to the levels proposed by the Common European Framework of Reference (Gaillat et al., 2021).

Also, it would be beneficial for Nakala to provide a global state versioning functionality, akin to the IDs that Git generates for each user commit. This would allow researchers not only to create datasets relying on the latest version of the corpus collections, but also to query previous versions of the same corpus collections. Such a functionality would support longitudinal comparisons within the same corpora.

Once achieved, the workflow will also be made available in order to allow other systems to extract and enrich linguistic data via Nakala's APIs. With this approach we seek to contribute to a general shift towards corpus interoperability, towards the dynamisation of automatic linguistic annotation of spoken corpora, and ultimately, towards a better understanding of the process of language acquisition.

References

Najib Arbach. 2015. *Constitution d'un corpus oral de FLE : enjeux théoriques et méthodologiques*. Phd thesis, Université Rennes 2.

Thomas Gaillat, Andrew Simpkin, Nicolas Ballier, Bernardo Stearns, Annanda Sousa, Manon Bouyé, and Manel Zarrouk. 2021. Predicting CEFR levels in learners of English: the use of microsystem criterial features in a machine learning approach. *ReCALL*.

TGIR Huma-Num, 2021. *Documentation NAKALA*. Documentation des services de la TGIR Huma-Num.

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

Jean-Marie Pierrel. 2014. Ortolang. Une infrastructure de mutualisation de ressources linguistiques écrites et orales. *Recherches en didactique des langues et des cultures. Les cahiers de l'Acedle*, 11(11-1). Number: 1 Publisher: Acedle (Association des Chercheurs et Enseignants Didacticiens des Langues Étrangères).

Georg Rehm, Kalina Bontcheva, Khalid Choukri, Jan Hajič, Stelios Piperidis, and Andrejs Vasiļjevs, editors. 2020. *Proceedings of the 1st International Workshop on Language Technology Platforms*. European Language Resources Association, Marseille, France.

Gilles Sérasset, Andreas Witt, Ulrich Heid, and Felix Sasaki. 2009. Multilingual language resources and interoperability. *Language Resources and Evaluation*, 43(1):1–14.

Jan Wijffels, 2020. *udpipe: Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the 'UDPipe' 'NLP' Toolkit*. R package version 0.8.5.

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1):1–9. Number: 1 Publisher: Nature Publishing Group.

Raphael Winkelmann, Klaus Jaensch, Steve Cassidy, and Jonathan Harrington, 2021. *emuR: Main Package of the EMU Speech Database Management System*. R package version 2.2.0.

Opening language resource infrastructures to non-research partners: practicalities and challenges

Verena Lyding
Eurac Research, Italy
{verena.lyding,egon.stemle}@eurac.edu

Egon Stemle
Eurac Research, Italy

Alexander König
CLARIN ERIC, The Netherlands
alex@clarin.eu

Abstract

By now, digital infrastructures for language data and tools have become commonplace in the research domain, but their possible benefits are still almost unknown outside of these circles. However, it stands to reason that the data and methods developed there could also be of use to non-research language actors like publishing houses or libraries. In this article, we present a use case within a local language infrastructure project that provides a newspaper portal with modern NLP tools via an API to help them improve their online search. We describe how this use case was implemented with a special focus on the problems that came up during the realisation, specifically those that arose from the interaction between a research and a non-research institution.

1 Introduction

Large scale research infrastructure projects like CLARIN[1] (de Jong et al., 2018), DARIAH[2] (Edmond et al., 2017) or ELG[3] (Rehm et al., 2021) aim at addressing the need for making language resources and tools available, sustainable and easily reusable by their stakeholders. The efforts have proven to create standards and frameworks and have become a reference point for visibility. Yet, up to today, the active involvement of stakeholders and the ambition to attract users to the provided services and tools is challenging. Different User Involvement (UI) events of CLARIN helped to provide specific training to a number of research stakeholders[4] and the CLARIN Resource Families initiative (Fiser et al., 2018) links resources of several research stakeholders. Stakeholders from industry are hardly found among the users and experiences from projects that actively involve commercial partners show that the industrial use of the offered services is far from trivial (Bleichner et al. (2005) report on a cooperation between two German universities and the a part of the archiving division of AIRBUS, and Poesio and Magnini (2009) report on a project with data providers of audio, video and text news from the Trentino Valley).

This might be related to the naturally slow advancement of these large scale, complex and sometimes abstract endeavours. In particular, solutions that aim at encompassing various use cases and demands tend to result in powerful yet generic frameworks, as for example, the Component Metadata Infrastructure[5] (Goosen et al., 2014). Those are not always trivial to adopt because they require knowledge and technical capacity to be realised, and it is not always clear at the beginning whether the actual use case can be implemented. For this reason, and to bridge the gap between research and application projects are created that cover domain-specific use cases with the help of large infrastructures (for example, ELEXIS (Woldrich et al., 2021)).

This paper presents a use case from the local language infrastructure project DI-ÖSS[6] (Lyding et al.,

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

[2] https://www.dariah.eu/

[3] https://www.european-language-grid.eu/

[4] See, for example, https://cmc-corpora2017.eurac.edu/uievent/ and an overview here: https://www.clarin.eu/content/user-involvement-funding#guest-blog-posts

[5] https://www.clarin.eu/cmdi

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

2019). This project bridges the gap between an existing infrastructure project (CLARIN) and a local community. That is, instead of targeting a specific application *domain* (like e.g. lexicography) DI-ÖSS targets a wider set of *local stakeholders* that are working with language in different ways. By doing so DI-ÖSS aims to connect those local actors to ideas, procedures and solutions from the large infrastructure. Among other use cases, DI-ÖSS implements an interaction between an online newspaper portal and an NLP service hosted at a research institution. The NLP service is offered and consulted via an API and has become an integral part of the daily workflow of the news portal. We describe the aims and objectives of this use case and explain the different steps followed for its implementation. In particular, we discuss the challenges that had to be tackled to get the inter-institutional cooperation to work.

2 Project background and finding partners

The given use case was carried out as part of the small local infrastructure project named DI-ÖSS (Lyding et al., 2019) which has been running from 2017 to 2021. The aim of the DI-ÖSS project was to connect various types of language actors on the local level to exploit synergies between their activities and institutional goals and the objectives of Eurac Research's Institute for Applied Linguistics (IAL). The IAL is a member of the European Infrastructure CLARIN and is the initiator and leader of the DI-ÖSS project. The project explicitly aimed at the involvement of non-research partners, which are typically not familiar with infrastructure efforts on the European level. A consortium with four local language actors was established to explore different use cases. The model of cooperation between each of the four local language actors and the IAL was an asymmetric project cooperation, with the major workload on the side of the IAL, and a smaller workload on the side of each of the partners. The contribution requested from the partners was limited to accompanying the use case development with their relevant institutional knowledge. Any active data curating and development work was delegated to subcontractors, which were coordinated by the IAL and paid by the project budget.

In the first phase of the project, was as the lead partner at IAL were searching for cooperation partners outside research as the aim of the project was to widen the idea of *infrastructure for language data* beyond the scope of research where it is well established by now. When looking for partners, it turned out to be surprisingly difficult to convince non-research institutions to join this kind of project. We encountered quite strong reservations regarding whether participating in such a project would be worth the institution's time. Especially institutions organised as a business have to always calculate whether the time (and therefore money) they invest into such a new project will be met by enough revenue, that is, will they get enough out of it or even more general *what* can they get out of it. For partners from the newspaper publishing world, three out of the four institutions we contacted were not interested in any cooperation even though, apart from the possible technological benefits, a small monetary reward was offered.

Another problem that paired with this general reservation was the challenging task of envisioning possible use cases that could be explored in such a cooperation. When meeting with people from the one newspaper portal that agreed to join the project as a partner, it was difficult for them to see beyond their day-to-day work within an established environment and to come up with a use case that could benefit their enterprise, and in the end, the idea for the use case was instead proposed by us as academic partner leading the project. It turned out that it is difficult for a potential industrial partner in such a project to envision a possible use case using NLP tools and other language technology methods because the extent of these methods is not widely known. Therefore, users either have no idea at all what is possible or on the other hand greatly overestimate the power of these tools and come up with ideas that are virtually impossible with today's capabilities. Also, the factor of having to adapt established workflows can pose obstacles for businesses offering professional services. Any adaptation can lead to a possible disruption of a workflow, and it is therefore understandable that non-research partners are particularly wary of committing to 'unnecessary' changes to a running system, even more so if the added value is something abstract like an evaluation metric, a promise of an improved experience, or a functioning prototype with different data or not fully integrated into their usual workflow. While we can envision potential use cases and can explain the expected added value of a project, the cooperation with a research partner and

research tools cannot be guaranteed to be as stable and predictable as commercial services. Insofar, it was a fine line we had to walk in order to entice partners with possible opportunities on the one hand, but also not to promise too much.

3 A use case for a newspaper portal

The use case explored further in this paper is built on the cooperation among the Institute for Applied Linguistics (IAL) at Eurac Research and the local online newspaper portal salto.bz[7] which is publishing news in German and Italian. The use case targets the creation of an improved search service for the text archive of the news portal. The new service combines a full-text search with manually created metadata for each article (i.e., author name, publication date and section of the newspaper) and automatically generated keywords as facets to refine the search results. While news articles and static metadata are created by the editorial team of the newspaper, the IAL worked on the task of creating a tool for the automatic extraction of keywords from text. The full-text search was developed by the partner with only little input from us. The keyword extraction tool extracts keyword and keyword like phrases from an article, written in one language, and uses the Microsoft Bing translation API[8] to translate each extracted keyword into the respective other language. The implementation of the new search interface that uses the general metadata, the full-text index and the keywords as facets was developed by the newspaper portal developer with some technical feed-back from us and the authors' team of the newspaper portal.

The cooperative work on this use case concerned the interaction between the author interface of the newspaper portal and the keyword extraction service offered by the IAL. This includes the user interaction as well as all aspects of the data exchange. In terms of the team effort for implementing this use case, both on our side and on the side of the newspaper portal, several people had to be involved. On the side of the newspaper portal, the cooperation was initially set up by the manager of the newspaper and was later handed over to the programmer of the news portal. Once an advanced prototype of the setup was completed, the editorial office of the newspaper and several authors were involved in discussing and testing usage aspects of the updated author interface. On our side, the different phases of the project were handled by four people covering project planning and communication, infrastructure knowledge and system administration and programming expertise. At the final steps of the collaboration also an external expert in translation and terminology was involved for data curation. Given the challenges for involving non-research partners in the project, it is relevant to observe that a relatively simple and seemingly technical use case required the involvement of people on all operational levels at the newspaper.

The primary goal of the project was to explore as many aspects of a cooperation as possible. In this respect, the project is to be understood as an all-encompassing feasibility study and not as a technical one, i.e. in particular that a functioning prototype was only one partial aspect. In the first place, it focused on creating and trial running a workflow that allows for the inter-institutional exchange and processing of texts, their integration into a running newsportal and the interaction of automatically applied procedures (the keyword extraction) with manual processing and validation tasks by the news authors. On the technical side, we are aware that keyword extraction and translation are in themselves extensive topics within computational linguistics, but ultimately we opted for pragmatic solutions to get the use case up and running within a restricted time frame. This was also made all the easier by the fact that it was desired from the side of the content creators (the news authors) to have full control over all automatically generated output by being able to check and change the automatically generated keywords as 'suggestions' in each case. The quality of the suggestions was one important aspect, but the integration of automatic methods and manual quality control within one workflow was the primary one. Either way, keywords for us are single or multi-word expressions that do not necessarily have to occur in sequence in the text and are extracted by a handful of manually devised rules[9]. Through the rules, we were able to ensure that some peculiarities of typical local naming, for example, names of public administration entities, can be recognised in both languages, German and Italian. The translation was done for each keyword individually

[7] `https://www.salto.bz`

[8] `https://docs.microsoft.com/en-us/azure/cognitive-services/translator/`

[9] See `keyword_extractor_salto.py` in `https://gitlab.inf.unibz.it/commul/di-oss/api-service-salto`

and independently of its context - this is a borderline use of the translation service, but for a working prototype we accepted this shortcoming.

4 Implementation of the use case

Several steps involving different competences had to be carried out for implementing the interaction between the authors' interface of the news portal and the keyword extraction and translation tool.

In a first step, we had to define the **user interaction** of newspaper authors with the keyword extraction tool. The keyword extraction tool receives a text either in Italian or German and returns a list of bilingual keyword pairs in the form *German@@Italian*. Together with the developer of the newspaper portal, we decided on three requirements for the user interaction. The interface should allow to (1) retrieve a list of candidates of keyword pairs on demand, (2) select or deselect candidates according to their relevance, and (3) modify or correct the selected candidates if needed. In order to accommodate these requirements, the author's interface has been enhanced with a section for managing keywords of articles. It allows to retrieve a candidate list on demand and to select or deselect candidates. Selected candidates get imported into a text field, where they can be modified by the user before saving them.

Another core part of the work related to the authors' interface relates to the communication with the newspaper authors. After conceptual and technical questions had been solved with the newspaper management and technician and the interface updates were approved by the chief editor, all newspaper authors who were expected to use the tool had to be brought on board. This required instructing them on how to use the tool and motivating them to effectively use it for each article they write. In order to succeed in motivating authors, the added value of employing the keyword tool had to be communicated (the keywords will serve as additional search facets), and the chief editor had to encourage participation. In terms of instructions, we had to provide (bilingual) guidelines of how to make decisions about selecting keyword candidates and apply corrections if needed. Finally, the authors also were instructed on how to report feedback and bugs, as a functioning feedback loop is essential to maintain and improve the service. Indeed, we encountered several situations in which authors had noticed a bug and stopped using the tool without informing us.

On the **technical side**, the exchange between the parties needed to be standardised so that all parties could develop their respective systems independently but still ensure that the systems would communicate with each other. To this end, a RESTful application programming interface (REST-API) was designed and implemented, which here provides a way for any external partner to send documents for computational linguistic processing to us, which they can retrieve after successful processing. The API implements an authentication and authorisation layer that allows us to trace which document originates from whom so that different processing or licence agreements can be accommodated.

The processing of the documents then takes place (optionally) asynchronously in order to also take into account processing that takes a long time. For this purpose, the API assigns an identification token after a job has been successfully submitted, which the remote party can use when returning for querying whether the processing has already been completed. Once the processing has been completed, the bilingual keyword pair candidates suggested by the system for a document can be retrieved. These pairs are then suggested to the authors for further refinement, as explained above.

The keyword pair candidates are compared with those already in the system, and those already known are marked in colour. As explained above, suggestions must be actively accepted, that is, there is no mechanism that automatically assigns suggestions to an article. In addition, the provision of keyword candidates is advantageous but not necessary since, in their absence, authors can also finish editing an article without automatic suggestions by typing their own suggestions manually, which are then also auto-completed with those known to the system.

Keywords can also be curated in a separate interface. Changes in this step are recorded in order to be able to systematically automate traceable changes. For example, a single-number-multiple-number unification can be made, which then receives an entry in a file and is taken into account in future suggestions. This would allow an editor to benefit from both the automated system (the proposal) and the regular curation of the taxonomy in a future article. In order to ensure an exchange of information on the

acceptance or content of the adopted keywords, data reconciliation is carried out at regular intervals. For this purpose, the log file is provided by the news portal developer and transferred to our system.

Finally, questions of **quality control and sustainability** were addressed once the technical environment and workflow was in place and got into production. The assigned keyword pairs could be observed in the context of the portal archive search, for which the generated keywords are used as a search facet. Since the portal search is publicly accessible by all salto.bz costumers, the results need to live up to a minimum quality standard, where quality relates to the adequacy of the assignment of a keyword to an article, the correctness of the keyword pair on both languages, and the coherence of the overall set of keywords (i.e., all keywords referring to the same concept should be merged into one harmonised version[10]). While the assignment can usually be done by the author with high reliability, the correction of keywords can be difficult for some authors, who might not be fully bilingual. In addition, the harmonisation of keywords requires a regular merging activity that is asynchronous to keyword assignments to single articles. To solve the translation and harmonisation issues, we hired a translation professional that carried out the merging and correction task on a weekly basis during the trial phase.

This need for quality control posed a bigger issue for the integration of the news archive, i.e., articles that were written and published before the keywords extraction tool was introduced and for the long term maintenance of the system. To update articles from the archive, we applied the automatic keyword tool and kept only those keywords that occurred more than ten times, which resulted in a list of more than 2000 keywords that were manually corrected and merged by our translation expert. To guarantee the quality of the service in the long term, the news portal has to put in the extra effort of regularly merging and checking newly introduced keywords. The assumption goes that the set of keywords will naturally consolidate over time to some extent, which means that the workload for merging and correction will continuously decrease. Another solution we discussed for reducing an additional workload is to switch to a semi-static model: a closed set of keywords will be fixed at some point and only keywords of this closed set will be assigned to any new articles. This closed set will be updated occasionally to include the most relevant new keywords ('hot topics') that occur over a longer span of time.

5 Conclusions

The reported work on the use case helped better understand what is needed to establish infrastructure cooperations with non-research partners. The most noteworthy challenge we encountered relates to establishing a cooperation and defining common use cases. We observed that besides a lack of awareness about ongoing language infrastructure initiatives, understanding what it can bring in terms of added value is missing on the side of non-research language actors. Resolving this issue required extended interdisciplinary communication efforts. In order to start from tangible scenarios it would be desirable that infrastructure initiatives like CLARIN worked towards a portfolio of use cases for cooperations with non-research stakeholder groups. A second difficulty relates to integrating the technical implementation with established workflows of a partner and fostering the adoption of new procedures. Overall, the communication and decision-making processes required interactions well beyond the technical level and concerned management and editorial participation to a considerable extent. Also, because the realisation of the use case impacted the customer facing newspaper portal search interface, many people outside our direct contacts paid great attention to all the changes. This experience shows that an asynchronous project cooperation with the major workforce on the side of the research partner is not realistically feasible. For future endeavours this suggests that truly interdisciplinary and inter-institutional project cooperations should be targeted, as mentioned above. Targeted project funding for research projects connected to industry are already trying to foster this in other contexts and should be extended to the infrastructure context. Finally, this also brought up the crucial question of quality control and long term maintenance of the service. The fact that through this project, the news portal created a dependency on an external service as part of their daily workflows underlines the importance of sustainability of infrastructure services and strategies for long term maintenance, which both pose unresolved challenges, not only in this kind of interdisciplinary cooperation but in the whole field of technical infrastructure. We conclude that up

[10] e.g., *offener Brief@@lettera aperta* was merged with *Brief@@lettera*

to today, establishing infrastructure cooperations with non-research partners is particularly challenging. In order to further open the paths for collaboration, bigger funded projects should help to pave the way by extending best practices and delivering road maps and business plans for a number of replicable use cases. As of today, on the side of the non-research client of European infrastructures, both awareness for what is doable and support for implementation still need to be promoted.

Acknowledgements

We would like to thank Monica Pretti for the detailed manual analysis and correction of the first comprehensive sample of automatically generated keyword pairs and the harmonisation of newly added keyword pairs during a limited testing period.

References

Martin Bleichner, Eugenie Giesbrecht, Helmar Gust, Eva-Maria Leicht, Petra Ludewig, Sabine Möller, Wiebke Müller, Martin Schmidt, Moritz Stefaner, Egon Stemle, and Katja Wilke. 2005. ASADO: The Analysis and Structuring of Aviation Documents - Final Report. Technical report, Institute of Cognitive Science at the University of Osnabrück and Institute of Applied Linguistics at the University of Hildesheim.

F. M. G. de Jong, Bente Maegaard, Koenraad De Smedt, Darja Fišer, and Dieter Van Uytvanck. 2018. CLARIN: Towards FAIR and Responsible Data Science Using Language Resources. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jennifer Edmond, Frank Fischer, Michael Mertens, and Laurent Romary. 2017. The dariah eric: Redefining research infrastructure for the arts and humanities in the digital age. *ERCIM News*, (111).

Darja Fiser, Jakob Lenardic, and Tomaz Erjavec. 2018. Clarin's key resource families. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Kôiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *LREC*. European Language Resources Association (ELRA).

Twan Goosen, Menzo Windhouwer, Oddrun Ohren, Axel Herold, Thomas Eckart, Matej Ďurčo, and Oliver Schonefeld. 2014. CMDI 1.2: Improvements in the CLARIN Component Metadata Infrastructure. In *Selected Papers from the CLARIN 2014 Conference*, pages 36–53.

Verena Lyding, Alexander König, Elisa Gorgaini, Lionel Nicolas, and Monica Pretti. 2019. DI-ÖSS - Building a digital infrastructure in South Tyrol. In *Selected papers from the CLARIN Annual Conference 2018, Pisa, 8-10 October 2018 / edited by Inguna Skadina, Maria Eskevich*, volume 159, pages 92–102, Linköpings universitet. Linköping Electronic Conference Proceedings.

Massimo Poesio and Bernardo Magnini. 2009. Content extraction meets the social web in the LiveMemories project. In Raffaella Bernardi, Sally Chambers, and Björn Gottfried, editors, *Proceedings of the Workshop on Advanced Technologies for Digital Libraries 2009 (AT4DL 2009)*, pages 42–45. Bozen Bolzano University Press, 09.

Georg Rehm, Stelios Piperidis, Kalina Bontcheva, Jan Hajic, Victoria Arranz, Andrejs Vasiļjevs, Gerhard Backfried, José Manuel Gómez-Pérez, Ulrich Germann, Rémi Calizzano, et al. 2021. European language grid: A joint platform for the european language technology community. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 221–230.

Anna Woldrich, Teja Goli, Iztok Kosem, Ondřej Matuška, and Tanja Wissik. 2021. ELEXIS: Technical and social infrastructure for lexicography, July. Published in K Lexical News (28), pp. 45-52.

ARCHE Suite: A flexible approach to repository metadata management

Mateusz Żółtak
ACDH-CH OEAW
Vienna, Austria
mateusz.zoltak@oeaw.ac.at

Martina Trognitz
martina.trognitz@oeaw.ac.at

Matej Ďurčo
matej.durco@oeaw.ac.at

Abstract

This article presents an innovative approach to metadata handling implemented in the ARCHE Suite repository solution. It first discusses the technical requirements for metadata management and contrasts them with the shortcomings of the existing solutions. Then, it demonstrates how the ARCHE Suite addresses those problems. After one year of productive use, we can assert that the approach implemented in the ARCHE Suite is viable and provides important benefits.

1 Introduction

The ACDH-CH at the Austrian Academy of Sciences operates the repository ARCHE for persistent hosting of humanities research data. ARCHE is certified as a CLARIN B-centre. Between 2017 and 2020, the underlying software technology we used was Fedora Commons version 4 with Blazegraph as a metadata store. Due to many serious shortcomings related to metadata management, the increasing amount of technical issues, and a lack of adequate alternatives, we decided to develop our own repository solution: the ARCHE Suite.

This paper specifies core requirements for metadata management and explains why they are not met by the existing repository solutions Fedora Commons, DSpace, Dataverse or Invenio. We describe how the desired features have been implemented in our solution and how they are used in our metadata management workflows. Finally, we discuss the challenges posed by our solution and summarise our first-year experiences of using it.

2 Technical requirements for metadata handling

Metadata is a vital part of every data repository, indispensable for finding, understanding, and reusing the data. To fully comply with the FAIR Data Principles that emphasise machine-actionability (Wilkinson et al., 2016), data and metadata have to be machine-readable and interoperable, which poses many challenges. The most important ones include ensuring metadata interoperability while preserving its descriptive precision and guarantee metadata consistency. Handling these challenges governs our core technical requirements for the ARCHE Suite.

2.1 Ensuring metadata interoperability

In the humanities and cultural heritage disciplines, the tremendous amount of metadata standards (e.g. (Riley, 2010) stands in the way of metadata interoperability. To overcome this, CLARIN has introduced the Component Metadata Infrastructure (CMDI) (Broeder et al., 2012), a standardised (ISO 24622-1, 24622-2) metadata framework with a built-in interoperability mechanism. Another widely used compromise is to apply the DCMI Metadata Terms (DCMI Usage Board, 2020), with the caveat of losing the potentially richer metadata to one basic common set of metadata descriptors. The repository solutions most popular among CLARIN centres - Fedora Commons 3 and DSpace - force users to use Dublin Core (DC) as a repository-native metadata format in a more or less explicit way.

In the last years, a new concept for (meta-)data interoperability has gained prominence: the Linked Open Data (LOD) principles (Berners-Lee, 2010) with five levels (stars) of compliance. Four-star LOD requires data to be provided in a W3C-compliant standard. This is easily met by using DC because of a well-defined mapping to RDF (W3C, 2014; Nilsson et al., 2008). The real challenge, however, is to additionally meet the requirements of five-star LOD, which includes the use of external links (Holborn, 2014). Using external URLs as DC term values meets the requirements but results in a repository inaccessible to human users, expecting human-readable labels rather than URLs. Using both URLs and human-friendly text labels as values results in problems with DC terms used multiple times (e.g. *dc:creator*) because the corresponding labels and URLs cannot be paired anymore.

Overall, the only viable solution to fully adopt five-star LOD seems to be providing full RDF support. While Fedora Commons 4 provides native RDF metadata support, it suffers from a serious feature drop compared to the previous version (most notably the lack of a search API and dissemination methods). As a result, its adoption has never become widespread. Dataverse presents a mixed approach. On the input side, it requires metadata to follow a bespoke Dataverse schema making it interoperable with other Dataverse repositories only. On the output side, metadata can be serialised into a few schemas (Institute for Quantitative Social Science, 2021), e.g. schema.org's Dataset RDF schema serialised as JSON-LD. Invenio allows any metadata schema which can be defined using the JSON Schema. Such a solution can be considered RDF-compliant to a large extent because RDF metadata can be serialised as JSON-LD, and the resulting JSON-LD structure can be described in the JSON Schema. The limitation here is that there can be many valid RDF to JSON-LD serialisations, and it can be impossible to describe all of them using the JSON Schema.

The interoperability imperative combined with the heterogeneous formats landscape implies that most repositories have to handle more than one metadata format. The enforcement of a metadata schema (often DC) by a repository software is undesired as it either prevents the handling of domain-specific metadata schemas or requires extensive customisation. The typical way of overcoming limitations imposed on the metadata schema by a repository software is to materialise domain-specific formats as separate repository data streams. The main disadvantage of this approach is making the information redundant, which brings the risk of inconsistency. A better solution is to keep a single copy of all metadata values in a schema-agnostic metadata store and to allow for on-demand conversion to the desired metadata format with a templating system. DSpace and Fedora Commons have no embedded support for on-the-fly metadata conversion, Dataverse provides a fixed set of built-in conversions as described above, and Invenio allows to write custom metadata schema conversion plugins in Python.

2.2 Ensuring metadata consistency

Ensuring metadata consistency, preferably at the ingest stage already, involves several aspects to be considered in the context of the repository management software. First, the way in which metadata checks are defined. This can be done either by specifying the allowed schema using configuration files or by plugging in own code which performs the checks. Dataverse only supports the first method, Fedora Commons 4 only the latter, DSpace and Invenio both, and Fedora Commons 3 has no support for custom metadata checks. Executing a pluggable code only after the data were stored in the repository, like in Fedora Commons 4, does not allow for reliable metadata checks because it either allows the metadata to stay in an inconsistent state or rejects it without notifying the client about the ingestion failure.

The second aspect regarding metadata consistency concerns the software layer, in which the metadata restrictions are verified. To ensure that checks can not be bypassed, they have to be enforced by a single software component responsible for handling all data irrespective of the ingestion interface.

The third important factor is the ability to ingest the data using ACID — atomicity, consistency, isolation, durability — (Haerder and Reuter, 1983) transactions. It is especially important from the LOD perspective where consistency of one repository resource metadata may depend on a successful creation (or update) of another repository resource. Unfortunately, ACID transactions are poorly supported by existing repository solutions. Invenio provides only a basic optimistic concurrency control on a single resource modification request level. Dataverse, DSpace and Fedora Commons 3 lack any concurrency

control on the client API level and our experience with the previous ACDH-CH repository based on Fedora Commons 4 proved its transaction support to be intrinsically broken.

2.3 Requirements list

To sum up, the desired repository solution should:

- Provide RDF support as the only viable way of fulfilling the five-star LOD principles.
- Not enforce any particular metadata schema.
- Avoid metadata duplication that come from materialising metadata in different formats.
- Allow for defining upon-ingestion metadata consistency checks in a flexible way.
- Allow for writing extensions in many programming languages.
- Ensure metadata consistency in a way that cannot be easily bypassed.
- Provide fully ACID transactions.

Unfortunately, none of the existing solutions provides support for all the points from this list. For this reason, we have developed a new repository software: the ARCHE Suite.

3 The ARCHE Suite

The ARCHE Suite has been developed from scratch as a bespoke, in-house repository solution within half a year including migration from the old Fedora 4-based repository in 2020. Before going into production it also underwent an external code review. Here, we detail the technical implementation of the metadata-related requirements formulated above. We focus on the developed software solution, ARCHE Suite, as opposed to ARCHE, the specific repository instance provided by the ACDH-CH with the ARCHE Suite as the underlying technology. While ARCHE Suite is schema-agnostic, in ARCHE every resource must be described with metadata respecting a bespoke and elaborate schema (Trognitz and Ďurčo, 2018). In fact, one can argue, the capabilities and specificities of ARCHE as a repository is determined by the combination of ARCHE Suite customised using a bespoke schema.

3.1 RDF support

We decided to avoid dependency on a triplestore and to use a relational database as a metadata storage instead. The database schema is developed in a way it can store any RDF data (does not enforce any particular RDF schema). There were two main reasons. First, using a triplestore makes it difficult to implement ACID transactions because triplestores do not recognize this concept. Second, when running the same queries against the system based on a relational database, we were able to achieve 2.6 times lower CPU usage and a 25 times lower memory footprint compared to our previous triplestore-based implementation. As a result, ARCHE Suite supports RDF as metadata a format both on input and output side but does not natively provide a SPARQL endpoint. A dedicated search API is used instead. However, a triplestore can be paired with ARCHE Suite either using the plugins system described below or by periodic synchronisation. We already successfully tested the periodic synchronisation scenario.[1]

The metadata model assumes a direct connection between the metadata RDF graph and the repository structure. Every node in the RDF metadata graph corresponds to a repository resource. The repository can be configured either to automatically create repository resources when an unknown RDF graph node is spotted in the metadata or to treat it as a metadata inconsistency and raise an error. Notably, such a data model provides a flexible and uniform framework for handling external authoritative data. This is discussed in the full version of the paper.

Metadata fetch and search APIs allow for including metadata of connected resources in the response. There are various ways in which the connection can be defined, e.g. *'all resources that are pointed to by a given resource's metadata'* or *'all resources which can be reached from a given one by following a given RDF predicate'*. This allows the user to fetch all the required metadata in a single call to the repository API making it much easier to develop services that use the repository directly as a persistence layer.

[1] See the arche2sparql Docker image: github.com/csae8092/arche2sparql

3.2 Metadata schema and metadata schema conversion

The ARCHE Suite does not enforce any particular metadata schema. The only requirement is the metadata to be expressed in RDF. The RDF predicates used for storing metadata managed internally by the repository (ifie.g. the date of resource's last modification) can be adjusted in the repository configuration.

The OAI-PMH service shipped with the ARCHE Suite allows to convert metadata into various XML-serialisable formats using a flexible templating system. We have successfully implemented conversions from our internal metadata schema to CMDI profiles as well as to the schema used by Kulturpool (Austrian Europeana aggregator) which allows us to entirely avoid materialising metadata in specific formats (cf. OAI_DC[2], Kuturpool[3] and CMDI profile p_1288172614023[4] serialisations of the same resource).

3.3 Custom metadata consistency checks

The only metadata consistency check performed automatically by ARCHE Suite is the *foreign key* constraint. As described above, all nodes of the RDF metadata graph are represented by repository resources, making it impossible to remove a repository resource that is pointed to by another resource's metadata.

All other checks have to be implemented as plugins by the repository administrator. The plugins can be written in any programming language with the AMQP message queue support[5]. Plugins bind to given events (before/after metadata/binary/transaction creation/modification). When an event occurs, the plugin is provided with resource metadata in the n-triples format and is expected to return metadata in the n-triples format or to raise an error. Plugins can be used both for metadata checks and enrichment as well as for synchronisation with external services (e.g. a triplestore).

The plugins system has turned out to be a very flexible and powerful tool. Dedicated plugins implemented for the ARCHE repository check metadata property cardinalities (applying different rules for resources of different RDF classes), mint PIDs, cast metadata property values to their proper RDF datatypes (including mapping string value labels to SKOS concept URIs for properties with controlled vocabularies) and compute aggregated metadata property values (e.g. summary of license types used by resources within a collection). A detailed discussion on how the plugins system is used in our repository deployment will be provided in the full version of the paper.

3.4 Transactions support

The ARCHE Suite provides full ACID support, although the isolation level is *read uncommitted* only. If consistency enforcement is undesired, it can be turned off by a configuration option. Importantly, all the *before event* plugins are considered part of an ACID transaction and thus, the ACID properties also extend to the plugins' actions. The transactions are backup-safe. In fact, the backup script uses its own transaction with a serialisable isolation level. All in all, the transaction system has proven to be useful in our everyday work. The full version of the paper discusses the incorporation of transactions in our repository ingestion workflows and the implications of transactions on the repository performance.

4 Summary

After a year of using the ARCHE Suite to run the ARCHE repository (currently over 1 TB of data, 86k resources, 2.6M RDF metadata triples), we can confirm it has met our expectations. It allows us to use RDF metadata as input and output format, to perform metadata enrichment and complex consistency checks within the repository software, as well as to avoid duplicating metadata by materialising various metadata formats. Notably, using the ARCHE Suite has significantly reduced server resources consumption compared to the previous solution based on a Fedora Commons 4 coupled with a Blazegraph triplestore.

[2] arche.acdh.oeaw.ac.at/oaipmh/?verb=GetRecord&metadataPrefix=oai_dc&identifier=https://hdl.handle.net/21.11115/0000-000C-29F8-F

[3] arche.acdh.oeaw.ac.at/oaipmh/?verb=GetRecord&metadataPrefix=kulturpool&identifier=https://hdl.handle.net/21.11115/0000-000C-29F8-F

[4] arche.acdh.oeaw.ac.at/oaipmh/?verb=GetRecord&metadataPrefix=cmdi&identifier=https://hdl.handle.net/21.11115/0000-000C-29F8-F

[5] The are more than 20 languages with AMQP Client libraries including Java, C/C++, Python, PHP, Ruby, JavaScript/node.

We are determined to develop the ARCHE Suite further and seek for cooperation with other CLARIN partners.

References

Tim Berners-Lee. 2010. Linked data.

Daan Broeder, Menzo Windhouwer, Dieter Van Uytvanck, Twan Goosen, and Thorsten Trippel. 2012. Cmdi: a component metadata infrastructure. In *Describing LRs with metadata: towards flexibility and interoperability in the documentation of LR workshop programme*, volume 1.

DCMI Usage Board. 2020. DCMI metadata terms.

Theo Haerder and Andreas Reuter. 1983. Principles of transaction-oriented database recovery. *ACM Computing Surveys*, 15(4):287–317.

Timothy Holborn. 2014. What is 5 star linked data?

Institute for Quantitative Social Science. 2021. Dataverse user guide - supported metadata export formats.

Mikael Nilsson, Andy Powell, Pete Johnston, and Ambjörn Naeve. 2008. Expressing dublin core metadata using the resource description framework (RDF).

Jenn Riley. 2010. Seeing standards: A visualization of the metadata universe.

Martina Trognitz and Matej Ďurčo. 2018. One schema to rule them all. the inner workings of the digital archive ARCHE. *Mitteilungen der Vereinigung Österreichischer Bibliothekarinnen und Bibliothekare*, 71(1):217–231, July.

W3C. 2014. Rdf 1.1 concepts and abstract syntax.

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. 2016. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3(1), March.

Legal Issues Related to the use of Twitter Data in Language Research

Pawel Kamocki
IDS Mannheim, Germany
kamocki@ids-mannheim.de

Vanessa Hannesschläger
OeAW, Austria
vanessa.hannesschlaeger@oeaw.ac.at

Esther Hoorn
Rijksuniversiteit Groningen, the Netherlands
e.hoorn@rug.nl

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

Marc Kupietz
IDS Mannheim, Germany
kupietz@ids-mannheim.de

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

Andrius Puksas
Mykolas Romeris University, Lithuania
andrius_puksas@mruni.eu

Abstract

Twitter data is used in a wide variety of research disciplines in Social Sciences and Humanities. Although most Twitter data is publicly available, its re-use and sharing raise many legal questions related to intellectual property and personal data protection. Moreover, the use of Twitter and its content is subject to the Terms of Service, which also regulate re-use and sharing. This extended abstract provides a brief analysis of these issues and introduces the new Academic Research product track, which enables authorized researchers to access Twitter API on a preferential basis.

1 Introduction

Social media data is useful for a wide variety of research disciplines in Social Sciences and Humanities, such as sociology, computer science, media and communication, political science, and engineering, to name a few. Twitter is still one of the most popular platforms for academic research on social media data (see Ahmed 2019). Tweet corpora are also used in linguistics (for example, a tweet sub-corpus is being added to the German Reference Corpus, DeReKo), albeit few tweet corpora are widely known, which may be because due to legal grey areas, such corpora are rarely shared.

Indeed, although most Twitter data is publicly available, its re-use and sharing (especially in a way compatible with Open Science requirements) raise many legal questions related to intellectual property and personal data protection.

2 Intellectual Property perspective on Twitter data

Copyright is an intellectual property right (IPR) that grants the author moral and economic rights over his or her work, including the exclusive right to copy it and make it available to the public.

A work is protected by copyright if it is original, i.e., constitutes the author's own intellectual creation. Although it varies from one jurisdiction to another, very short works such as titles or slogans are often considered unoriginal, as the intellectual creation cannot manifest itself in a very short format. The maximum length of a tweet is currently set at 280 characters (increased from 140 in November 2017), i.e. about 50-60 words in English, which seems more than enough to potentially qualify for copyright

protection. However, in practice, very few tweets reach the maximum threshold, and most of them are considerably shorter: the most common length of a tweet in English has been reported to be only 33 characters long (Perez, 2017), i.e. approximately 6-7 words. Nevertheless, it appears that even works of such short length may still qualify for copyright protection: Kamocki (2020) argues that only n-grams that are no longer than 3 words can safely be regarded as copyright-free.

This does not mean that all tweets are indeed original and protected by copyright. Arguably, in reality, and from the quantitative perspective most tweets (like "Big win!", "LewanGOALski!!!!!!!!!!1111" or "This is crazy LOL") fail to meet the originality criterion. However, a pack of several thousand tweets is likely to contain copyright-protected material (even if it does not include photographs or other media). Therefore, while analysing Twitter data (which necessarily involves at least reproduction, i.e. copying of tweets, an act restricted by copyright law), copyright issues cannot be ignored.

Re-use of copyright-protected content is only legally possible if it is authorised by the author (directly or indirectly) or exempted from authorisation by a statutory exception. The recent EU Directive 2019/790 on Copyright in the Digital Single Market (DSM Directive) introduces a specific framework for text and data mining, including an exception for TDM for research purposes (Article 3). This exception allows research organisations to make copies of the content that they have lawful access to. When it comes to publicly available tweets, the criterion of lawful access is met, as per Recital 14 of the DSM Directive *("Lawful access should also cover access to content that is freely available online"*).

The copies made under the "TDM for scientific research" exception have to be stored with the appropriate level of security, but they can be retained for re-use in other projects or for evaluation purposes. However, the exception does not seem to allow any sharing of the data, although there might be slight variations between implementations in the various EU Member States (for example, the German implementation allows for sharing 'with a specifically limited circle of persons for joint research').

Another intellectual property right that could potentially apply to Twitter is the *sui generis* database right. Under this framework, Twitter could claim an exclusive right in its database of tweets, enabling the company to prevent users from extracting and/or re-using the whole database, or a substantial part thereof, independently from any copyright in the content. Although rarely discussed, this right could considerably limit access to tweets or web content in general. However, it is essential to keep in mind that the *sui generis* database right only applies to companies "formed in accordance with the law of a [EU] Member State and having their registered office, central administration or principal place of business within the [European Union]" (Article 8.2 of the Directive 96/9 on Databases). For companies that, like Twitter, only have registered offices in the European Union (Twitter currently has offices in Dublin, Paris, Berlin, Brussels and Madrid), their operations must be "genuinely linked on an ongoing basis with the economy of a Member State". It is not clear whether this is the case with Twitter. And even if it is, a sample of tweets that can be used in a language research project is quite unlikely to constitute a 'substantial part' of all tweets. In light of the above, the impact of the *sui generis* database right on the re-use of tweets for language research is probably minimal and can be ignored.

3 Contract law perspective on Twitter data

To post tweets, one needs to create a Twitter account and accept (among other documents) Twitter's Terms of Service (ToS)[1]. As per Paragraph 3 of the ToS, although the user retains copyright in his or her tweets, he or she grants Twitter a very broad license to re-use them for free on a non-exclusive basis. This means that someone who would like to copy and share tweets can receive the necessary authorization either directly from the user (which in most cases is unworkable in practice, given the sheer number of Twitter users) or from Twitter. Theoretically, nothing prevents users from re-publishing their tweets outside of Twitter, including, e.g. in .xml format and under an open license.

Twitter ToS also grant every user access to the tweets, although certain actions are expressly forbidden. These include accessing or searching (or attempting to access or search) Twitter content by any means other than interfaces provided by Twitter and scraping tweets without prior consent from the company.

[1] Available at: https://twitter.com/en/tos#intlTerms (access: 27.4.2021).

It seems, therefore, that mining of tweets without specific permission (e.g. without being granted access to the mining interface provided by Twitter), even if done for research purposes, would violate Twitter ToS, which may lead to suspension or termination of the user account(s) that is (are) at the origin of these actions. This might be the reason why those researchers who have indeed scraped data from Twitter may not be transparent about it.

Interestingly, as regards copyright, text and data mining for research purposes by research organisations is covered by the abovementioned exception of Article 3 of the DSM Directive. Article 7 of the same Directive expressly states that any contractual provision contrary to this exception should be unenforceable. In specific contexts, national contract laws (e.g. regulating unfair contractual terms) could also have an impact.

It is far from clear, however, how this will work in practice. In our opinion, if a user (affiliated with a research organisation) scrapes or attempts to scrape tweets for scientific research purposes without specific permission from Twitter, he or she would still violate Twitter ToS (and likely see his account closed or at least suspended), even though he or she would not be liable for copyright infringement. He or she would also be able to retain the copies, according to Article 3 of the DSM Directive. However, Twitter could potentially sue the person for damages for breach of contract. Still, in our opinion, this is quite unlikely to happen taking into account the above-mentioned copyright exception and the limited amount of damages that could possibly be obtained, as well as the hypothetical reputational loss for Twitter for suing a researcher or a research institution. Even if a researcher were sued by Twitter, and found guilty of copyright infringement, the penalty (given the nature of academic research) will likely be moderate (probably not exceeding 10000 EUR); however, the consequences in the relations with funding agencies, and within the research institutions, can potentially be more dire.

4 Technological Protection Measures

Although the authors have not tested it, it can be assumed that scraping tweets is not only in principle forbidden by Twitter ToS but also made impossible (or at least very difficult) by technological protection measures (TPM). In addition to being in principle forbidden by law (Article 6 of the Directive 2001/29 on Information Society), circumvention of TPM is also expressly prohibited by Twitter ToS (and will likely lead to a lifetime ban).

Can the above-mentioned exception of Article 3 of the DSM Directive be interpreted as allowing circumvention of TPM in the context of text and data mining for scientific research purposes? This seems to be the most significant grey area of the new exception, as Article 3 allows rightholders to apply TPM only to the extent necessary to ensure the security and integrity of their networks and databases. In our opinion, Twitter would have an excellent chance to succeed in arguing that TPMs implemented to prevent unauthorised scraping are, in fact, necessary to achieve these goals. However, it remains to be seen how this issue will be worked out in practice.

5 Data protection perspective

In addition to being potentially copyright-protected, tweets should also be regarded as personal data (Gold, 2020), as they contain identifying information (at the very least the user ID, but possibly also location metadata or other identifying content). Therefore, their processing needs to follow the GDPR (even though Twitter is an American company -- as per its Article 3.2, the GDPR applies to foreign companies which offer services to EU citizens), as well as applicable ethical rules.

Twitter provides its users with the possibility to fine-tune their privacy settings, including public availability of their tweets and profile information, which potentially may be interpreted as granting/withdrawing consent. Today, it seems that an average Twitter user should be aware that public tweets can be used for research purposes (Twitter expressly informs their users, in its Rules and Policies, that it conducts research with user data). Therefore, it can be argued that data pertaining to the author of a tweet can be processed for research purposes on several bases: consent, legitimate interest, and (in countries where such legal basis is available for research) public interest. Even sensitive data may be lawfully processed in this context on the ground that such data have been made manifestly public by the data subject (Article 9.2(e) of the GDPR). However, it is recommended to stop processing tweets access to which have been restricted by the user; such action should be interpreted as withdrawal of consent or objection to the processing. Moreover, the processing of tweets still needs to meet all the requirements

of the GDPR (such as security and accountability), especially transparency (unless the applicable national law provides for an exception). As regards the latter, making the information publicly available is a reasonable solution in cases where contacting every user individually would require disproportionate effort (see Article 14.5(b) of the GDPR).

6 Use of Twitter API for research purposes

Albeit it generally seems permissible under Article 3 the DSM Directive, the use of Twitter data for language research purposes is still associated with considerable organizational effort and lack of legal certainty. In this context, simply obtaining specific permission from Twitter may be a reasonable alternative to relying on statutory exceptions. This could clear any copyright-related issues and diminish the burden related to the GDPR -- when the processing is carried out solely through an API provided by Twitter, it can be argued that Twitter is at least a joint controller for the processing.

Twitter has been offering access to APIs for mining tweets for a long time. Recently, in July 2020, Twitter launched a new version (v2) of its API. Reportedly, academic researchers were one of the largest groups of the API users; for this reason, in January 2021, Twitter launched a new Academic Research product track, allowing researchers to get preferential access to the API.

In theory, this Track allows for a 10 000 000 monthly tweet volume cap (compared to 500 000 in the general track). However, this also depends on the streaming endpoint limits, which reportedly are not entirely up to this standard yet (although they are expected to be raised soon). Moreover, it is also possible to use more detailed queries and rules (1024 characters per query/rule in the Academic Research product track, as opposed to 512 in the Standard track). Finally, Twitter Development Agreement allows academic researchers to distribute an unlimited number of Tweet IDs and/or User IDs if they are doing so on behalf of an academic institution and for the sole purpose of non-commercial research (otherwise, 'only' 1 500 000 Tweet IDs per 30-day period can be shared).

The Academic Research product track is available to graduate students, PhD students, post-docs, faculty or research-focused employees at an academic institution or university with a precisely defined research objective and pursuing non-commercial purposes. To apply for the track, a researcher has to answer a very detailed questionnaire including questions about the project, its funding, methodology, the planned use of Twitter data and ways of sharing the outcomes. Arguably, some may see this questionnaire as intrusive and unacceptable from the point of view of academic freedom.

Access to the Track is free of charge. There is no information available as to how many requests are granted. Like anyone with access to the API, successful candidates are bound by the Twitter Development Agreement and Policy. These documents strictly prohibit any attempt to exceed or circumvent access limitations (rate limits). Moreover, Twitter retains the right to immediately terminate or suspend access to the API at any time and for any reason. It can be expected that any attempt to exceed the permissions granted by Twitter, also based on the above-mentioned statutory exception for text and data mining, will be met with the termination of access to the API.

References

Wasim Ahmed. Using Twitter as a data source: an overview of social media research tools (2019). Available at https://blogs.lse.ac.uk/impactofsocialsciences/2019/06/18/using-twitter-as-a-data-source-an-overview-of-social-media-research-tools-2019/ (26.4.2021).

Nicolas Gold. Using Twitter Data in Research. Guidance for Researchers and Ethics Reviewers. University College London (2020). Available at https://www.ucl.ac.uk/data-protection/sites/data-protection/files/using-twitter-research-v1.0.pdf (27.4.2021).

Pawel Kamocki. 2020. When Size Matters. Legal Perspective(s) on N-grams. Proceedings of CLARIN Annual Conference 2020. 05 – 07 October 2020. Virtual Edition. Ed. Costanza Navarretta, Maria Eskevich. CLARIN, 166-169.

Sarah Perez (2017). Twitter officially expands its character count to 280 starting today. Available at https://techcrunch.com/2017/11/07/twitter-officially-expands-its-character-count-to-280-starting-today/?guccounter=1 (26.4.2021).

The Interplay of Legal Regimes of Personal Data, Intellectual Property and Freedom of Expression in Language Research

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

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

Pawel Kamocki
IDS Mannheim, Germany
kamocki@ids-mannheim.de

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

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

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

Vadim Mantrov
University of Latvia, Latvia
vadims.mantrovs@lu.lv

Vanessa Hannesschläger
OeAW, Austria
vanessa.hannesschlaeger gmail.com

Riccardo Del Gratta
ILC, Italy
riccardo.delgratta ilc.cnr.it

Age Värv
University of Tartu, Estonia
age.varv@ut.ee

Gaabriel Tavits
University of Tartu, Estonia
gaabriel.tavits@ut.ee

Andres Vutt
University of Tartu, Estonia
andres.vutt@ut.ee

Abstract

Sometimes legal scholars get relevant but baffling questions from laypersons like: "*The reference to a work is personal data, so does the GDPR actually require me to anonymise it? Or, as my voice data is personal data, does the GDPR automatically give me access to a speech recognizer using my voice sample? Or, can I say anything about myself without the GDPR requiring the web host to anonymise or remove the post? What can I say about others like politicians? And, what can researchers say about patients in a research report?*" Based on these questions, the authors address the interaction of intellectual property and data protection law in the context of data minimisation and attribution rights, access rights, trade secret protection, and freedom of expression.

1 Introduction

There is an awareness that intellectual property (IP) and personal data (PD) protection are relevant in language research. These two regimes are often applicable simultaneously, and their requirements might

seem contradictory. Therefore, the authors have chosen three specific cases[1] to outline the interaction of IP and PD protection and provide preliminary guidance.

Firstly, the authors explore the interplay between the data minimisation principle and the right to be acknowledged as the author (the attribution right). On the one hand, the data minimisation principle enshrined in the General Data Protection Regulation (GDPR) requires processing[2] as little PD as possible (Art. 5 (1) c)). On the other hand, the Berne Convention Art. 6^{bis} gives the authors the attribution right. The relevant question here is whether a researcher who has collected language data (LD) containing copyrighted content should attribute the author of the content or follow the data minimisation principle and remove all PD (e.g., the author's name) that is not necessary for processing.

The second case concerns IP protection and the data subject's access right. In practical terms, a researcher might need to decide what data the access right covers. Is it only raw PD or data derived from PD (e.g., a language model which could contain trade secrets)?

Thirdly, the authors discuss the impact of PD protection on freedom of expression since publications constitute research outcomes. The authors rely on the regulation of their countries.

2 The data minimisation and the right of attribution

According to the data minimisation principle, PD must be "*adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed*" (GDPR Art. 5 (1) clause c).

The European Data Protection Board (EDPB) explains it further: "*minimising can also refer to the degree of identification. If the purpose of the processing does not require the final set of data to refer to an identified or identifiable individual (such as in statistics), but the initial processing does (e.g., before data aggregation), then the controller shall anonymize personal data as soon as identification is no longer needed. Or, if continued identification is needed for other processing activities, personal data should be pseudonymized to mitigate risks for the data subjects' rights*" (2019: 19).

According to Art. 6^{bis} (1) of the Berne Convention, "*the author shall have the right to claim authorship of the work*". The InfoSoc Directive also contains the obligation to identify the source (incl. the author's name) (Art. 5 (3)). The EU case law reiterates the obligation (e.g., C-145/10). In other words, there is a legal obligation to acknowledge the author of the content. It is compatible with the GDPR since it names compliance with a legal obligation as a legal basis for PD processing (Art. 6 (1) c)).

An overarching theme for this and the following section concerns legal obligations relating to derived data (e.g., data derived through text and data mining (TDM). For further discussion, see Kelli et al. (2020). Interestingly, the TDM exception contained in the DSM Directive does not require attribution. However, it should be borne in mind that the TDM exception only limits the reproduction right. In case the results of TDM are disseminated, the attribution right has to be honoured.[3]

3 The scope of access right and trade secret protection

In combination with the right to be informed and the principle of transparency, the access right forms a foundation for exercising the data subjects' rights. The access right requires the controller to provide information on the processing of PD (GDPR Art. 15). The first question for research organisations and researchers (data controllers) is the scope of PD subject to the access right. The data subjects should have access to their raw data. The question is whether the access right applies to data derived from PD as well. This question asks what PD covers.[4] The Court of Justice of the European Union (CJEU) has not been remarkably consistent. For instance, it has explained that "*There is no doubt that the data relating to the applicant for a residence permit and contained in a minute, such as the applicant's name, date of birth, nationality, gender, ethnicity, religion and language, are […] 'personal data' […] As*

[1] It should be mentioned that there are a myriad of IP and PD protection interaction points whose systematic mapping is outside the scope of this article. Therefore, the authors chose cases which could potentially be relevant for language researchers.

[2] The GDPR defines processing so widely that it covers all possible activities with PD (Art. 4 (2)).

[3] The attribution right exists only in case of the existence of copyrighted content. In the EU case law, it is pointed out that 11 consecutive words could be copyright protected (C-5/08). However, it does not say that less than 11 words are not copyrighted. For further discussion, see Kamocki 2020.

[4] For the concept of PD, see WP29 2007.

regards, on the other hand, the legal analysis in a minute, it must be stated that, although it may contain personal data, it does not in itself constitute such data" (C-141/12 paragraphs 38, 39). In another case, the CJEU held that "*the written answers submitted by a candidate at a professional examination and any comments made by an examiner with respect to those answers constitute personal data*" (C-434/16).

Understandably, the concept of PD should be interpreted extensively. However, there is no legal clarity on whether data derived from PD should be made available as well. WP29 (2016: 9) suggests in the context of the right of portability (see GDPR Art. 20) that "*user categorisation or profiling are data which are derived or inferred from the personal data provided by the data subject, and are not covered by the right to data portability*".

Within the context of language research, the question is whether the data subject could require access to a language model which was trained using his PD. If not, there should be a legal basis limiting the scope of access right. One argument here is that model is protected by intellectual property (copyright, trade secret). The GDPR accepts this line of argument in its Recital 63, explaining the nature of the access right: "*That right should not adversely affect the rights or freedoms of others, including trade secrets or intellectual property and in particular the copyright protecting the software. However, the result of those considerations should not be a refusal to provide all information to the data subject*".

To sum up, the access right does not cover the access to language models, especially when their creators consider trade secrets. In this context, IP rights prevail over data protection.

4 Data subject's rights and freedom of expression

The data subject has the right to object to the processing and obtain the erasure, restriction or rectification of PD concerning him (GDPR Art. 16, 17, 18, 21). These rights may conflict with the author's right to make his work available. This question can be framed as an interaction of PD protection and freedom of expression (FoE). Protection of PD is not an absolute right (GDPR Rec. 4). Therefore, the GDPR allows Member States to limit the data subject's rights to reconcile PD protection with the freedom of expression.

The EU countries have followed different implementation routes. For instance, the German Federal Data Protection Act (BDSG) does not contain any rules specifically implementing Article 85 of the GDPR. The existing broad derogation for research and archiving purposes (Article 27 of the BDSG) is based on Article 89, not 85 of the GDPR. It seems to be deemed sufficient by the legislator (Deutscher Bundestag, 2018). Specific state acts regarding media and journalistic expression exist in many federal states (Länder), e.g., Hessisches Pressegesetz or Landesmediengesetz Baden-Württemberg.

Most of the studied countries have adopted a general provision limiting the applicability of the GDPR to ensure freedom of expression (see Austrian, Estonian, Finnish and Lithuanian PDPA). France, Greece, Italy, and Latvia have more detailed regulations, which are explored below.

Article 80 of the French Data Protection Act attempts to reconcile data protection and freedom of expression in France. It derogates from two general principles: the storage limitation and the prohibition of processing sensitive data (including data about criminal convictions and offences). It also limits information rights, access, rectification and restriction, and derogates from the rules on data transfers. This framework applies only when necessary to safeguard freedom of expression and information, and only when the data are processed: 1) for academic (*'universitaire'*), artistic or literary expression, or 2) for journalistic purposes by professional journalists, in a way that respects ethical rules (deontology) of the profession. The Article clearly states that other laws and codes regarding violations of privacy and reputational damage continue to apply.

One can be surprised by the adjective '*universitaire'* in Article 80 of the French Copyright Act (expression *universitaire, artistique ou littéraire*) rather than '*académique'* (as in '*academic, artistic or literary expression*'). However, the same wording is used by the French version of Article 85 of the GDPR. This is because '*académique'* has a very restricted meaning in French (related to the Académie Française) and should not be interpreted as limiting the derogatory framework to processing made by scholars with a university affiliation.

Article 28 of the Greek Personal Data Protection Act, corresponding directly to the GDPR (Art. 85), aims to reconcile the right to personal data protection with the right to freedom of expression and information, *"including the processing for journalistic purposes and for purposes of academic, literary or artistic expression".* More specifically, in the framework of these objectives, Paragraph 1 of this Article

explicitly enumerates cases where the processing of PD is allowed: *"(a) when the subject of the data has given his explicit consent, (b) for PD that have been publicised by the subject, (c) when the right to the freedom of expression and the right to information overweighs the right to PD protection, especially for topics of general interest or when the PD relates to public persons, and (d) when it is restricted to the necessary measure to ensure the right of expression and the right of information, especially with regard to sensitive categories of PD, and criminal cases, and security related measures, taking into account the right of the subject to his private and family life."* We can deduce that the Article looks more into the '*journalistic purposes*' rather than '*academic purposes*'. Paragraph 2 of the same Article provides the exceptions and derogations for processing for such purposes, which are mentioned in Article 85 of the GDPR.

Article 136 of the Italian Personal Data Protection Code (PDPC) implements Art. 85 of the GDPR. It regulates journalistic as well as academic works. Article 137 defines the categories of PD that can be processed without the data subject's consent. Namely, such categories are special categories of PD and PD data related to criminal convictions and offences (GDPR Art. 9, 10). Other sections further restrict these categories to *"Safeguards applying to the processing of genetic data, biometric data, and data relating to health"* (section 2-f) and *"Processing entailing a high risk for the performance of a task carried out in the public interest"* (section 2-p). Article 137 (3) provides: "*It shall be allowed to process the data concerning circumstances or events that have been made known (communicated/disseminated) either directly by the data subject or on account of the data subject's public conduct*".

Article 32(3) of the Latvian PDPA states that when processing data for academic, artistic or literary expression, provisions of the GDPR (except for Article 5) shall not be applied if all of the following conditions are present: 1) data processing is conducted by respecting the right of a person to private life, and it does not affect interests of a data subject which require protection and override the public interest; 2) compliance with the provisions of the GDPR is incompatible with or prevents the exercise of the rights to freedom of expression and information.

There are two intriguing questions concerning the interaction of PD protection and freedom of expression:

1) how to strike a fair balance between PD protection and FoE in research settings? FoE is usually framed in the context of newspapers publishing facts about public figures. The freedom of academic expression is somewhat unclear. Still, it has been interpreted to apply to, e.g., x-ray pictures of medical case studies as standard practice. Such accompanying material is publicly disclosed in a scientific journal as an illustration of the published case. There is no need to obtain the consent of the x-rayed person for this purpose. Usually, a person cannot be directly identified from such an x-ray. However, if the medical condition is rare, the individual may still be identifiable with the help of additional information. As the GDPR defines data concerning health as special categories of PD (Art. 9), this is an especially delicate example.

2) how and where to draw a line between processing for academic expression and research purposes. Research publication requires prior research. The question is whether this research is covered with the freedom of academic expression. The authors admit that when data is present in the research publication, then the processing could be covered by the FoE exception. It is important to emphasise that the principles of data minimisation, purpose limitation, accuracy, fairness (GDPR Art. 5), and other requirements need to be followed. At the same time, research ethics and funders' requirements require the publication of research data to ensure research reproducibility and verifiability. Therefore, there is a tension between open data and personal data protection requirements. For further discussion, see Kelli et al. (2018).

PD protection and FoE are both human rights. This means that one is not prioritized over another. The key issue is to strike a fair balance between them. PD protection should not affect academic freedom of expression.

5 Conclusion

The authors' reached the following preliminary conclusions. Firstly, the data minimisation and the attribution right are not contradictory concepts. The acknowledgement of the author is compatible with the GDPR as the compliance with a legal obligation. The attribution does not concern all PD but only data that is copyrighted. Secondly, the access right primarily applies to raw PD. There is no legal clarity regarding the access to data derived from PD. The access right does not presumably cover language

models containing trade. Thirdly, PD protection usually does not take precedence over the freedom of expression and cannot hinder the academic FoE and the author's right to disseminate his work. Conducting research could be covered FoE.

References

Austrian PDPA. *Data Protection Amendment Act.* Entry into force 2018. Available at https://www.ris.bka.gv.at/eli/bgbl/I/2018/31 (3.4.2021).

Berne Convention. *Berne Convention for the Protection of Literary and Artistic Works of September 9, 1886*. Available at https://wipolex.wipo.int/en/text/283698 (3.4.2021).

C-434/16. *Case C-434/16.* Peter Nowak v Data Protection Commissioner (20 December 2017). Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A62016CJ0434&qid=1617649690306 (5.4.2021).

C-141/12. *Joined Cases C-141/12 and C-372/12.* YS (C-141/12) vs Minister voor Immigratie, Integratie en Asiel, and Minister voor Immigratie, Integratie en Asiel (C-372/12) v M, S (17 July 2014). Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A62012CJ0141&qid=1617633666683 (5.4.2021).

C-145/10. *Case C-145/10.* Eva-Maria Painer vs Standard VerlagsGmbH and Others (1 December 2011). Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A62010CJ0145&qid=1618044850444 (10.4.2021).

C-5/08. *Case C-5/08.* Infopaq International A/S vs Danske Dagblades Forening (16 July 2009). Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1555243488182&uri=CELEX:62008CJ0005 (10.4.2021).

Deutscher Bundestag, 2018. *Deutscher Bundestag,* Ausarbeitung: Die Öffnungsklausel des Art. 85 der Datenschutz-Grundverordnung, WD 3 - 3000 - 123/18. Available at https://www.bundestag.de/resource/blob/560944/956f5930221c807984d40c1df2af5abf/WD-3-123-18-pdf-data.pdf (26.04.2021).

DSM Directive. Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC. *OJ L 130, 17.5.2019*, pp. 92-125. Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1572352552633&uri=CELEX:32019L0790 (10.4.2021).

EDPB 2019. *European Data Protection Board.* Guidelines 4/2019 on Article 25 Data Protection by Design and by Default. Adopted on 13 November 2019. Available at https://edpb.europa.eu/sites/edpb/files/consultation/edpb_guidelines_201904_dataprotection_by_design_and_by_default.pdf (4.4.2021).

Estonian Copyright Act. *Copyright Act.* Entry into force 12.12.1992. Available at https://www.riigiteataja.ee/en/eli/504032021006/consolide (3.4.2021).

Estonian PDPA. *Personal Data Protection Act.* In force from: 15.01.2019. Available at https://www.riigiteataja.ee/en/eli/523012019001/consolide (6.4.2021).

EU Charter of Fundamental Rights. *Charter of Fundamental Rights of the European Union.* 2012/C 326/02. OJ C 326, 26.10.2012, p. 391-407. Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:12012P/TXT (1.4.2021).

Finnish PDPA. *Data Protection Act (1050/2018).* Available at https://www.finlex.fi/en/laki/kaannokset/2018/en20181050.pdf (8.4.2021).

French Data Protection Act. *Loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés*. Available at https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000886460/ (27.4.2021).

French Intellectual Property Code. *Code de la propriété intellectuelle.* Available at https://www.legifrance.gouv.fr/codes/texte_lc/LEGITEXT000006069414/ (27.4.2021).

German Federal Data Protection Act. *Bundesdatenschutzgesetz vom 30. Juni 2017 (BGBl. I S. 2097)*, das durch Artikel 12 des Gesetzes vom 20. November 2019 (BGBl. I S. 1626) geändert worden ist. Available at: https://www.gesetze-im-internet.de/bdsg_2018/BJNR209710017.html (27.4.2021).

GDPR. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *OJ L 119*, 4.5.2016, p. 1-88. Available

at https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1555312258399&uri=CELEX:32016R0679 (1.4.2021).

Greek Data Protection Act. *Personal Data Protection Law (4624/2019)*. Available at: https://www.e-nomothesia.gr/kat-dedomena-prosopikou-kharaktera/nomos-4624-2019-phek-137a-29-8-2019.html (26.4.2021).

InfoSoc Directive. Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society. *Official Journal L 167*, 22/06/2001 P. 0010 – 0019. Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1555254956114&uri=CELEX:32001L0029 (4.4.2021).

Italian PDPA. *Personal Data Protection Code containing provisions to adapt the national legislation to Regulation (EU) 2016/679* of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC. Available at https://www.garanteprivacy.it/documents/10160/0/Data+Protection+Code.pdf/7f4dc718-98e4-1af5-fb44-16a313f4e70f?version=1.3 (8.4.2021).

Kamocki, Pawel. 2020. When Size Matters. Legal Perspective(s) on N-grams. *Proceedings of CLARIN Annual Conference 2020. 05 – 07 October 2020*. Virtual Edition. Ed. Costanza Navarretta, Maria Eskevich. CLARIN, 166-169. Available at https://office.clarin.eu/v/CE-2020-1738-CLARIN2020_ConferenceProceedings.pdf (10.4.2021).

Kelli, Aleksei, Arvi Tavast, Krister Lindén, Kadri Vider, Ramunas Birštonas, Penny Labropoulou, Irene Kull, Gaabriel Tavits, Age Värv, Pavel Stranák, Jan Hajic. 2020. The Impact of Copyright and Personal Data Laws on the Creation and Use of Models for Language Technologies. In: Kiril Simov, Maria Eskevich (Ed.). *Selected Papers from the CLARIN Annual Conference 2019 (53−65)*. Linköping University Electronic Press. Available at https://ep.liu.se/ecp/172/008/ecp20172008.pdf (10.4.2021).

Kelli, Aleksei, Tõnis Mets, Lars Jonsson, Krister Lindén, Kadri Vider, Ramūnas Birštonas, Age Värv. 2018. Challenges of Transformation of Research Data into Open Data: the Perspective of Social Sciences and Humanities. *International Journal of Technology Management and Sustainable Development*, 17 (3), 227−251.

Latvian Copyright Act. *Latvian Copyright Act.* Available at https://vvc.gov.lv/image/catalog/dokumenti/Copyright%20Law.doc (26.4.2021).

Latvian PDPA. *Latvian Personal Data Processing Law.* Available at https://vvc.gov.lv/image/catalog/dokumenti/Personal%20Data%20Processing%20Law.doc (26.4.2021).

Lithuania PDPA. *Republic of Lithuania Law on Legal Protection of Personal Data.* Available at https://www.e-tar.lt/portal/legalAct.html?documentId=43cddd8084cc11e8ae2bfd1913d66d57 (26.4.2021).

WP29 2016. *Article 29 Working Party (WP29)*. Guidelines on the right to data portability. Adopted on 13 December 2016. Available at https://ec.europa.eu/information_society/newsroom/image/document/2016-51/wp242_en_40852.pdf (5.4.2021).

WP29 2007. *Article 29 Working Party (WP29)*. Opinion 4/2007 on the concept of personal data. Adopted on 20th June. Available at https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2007/wp136_en.pdf (5.4.2021).

Ethnomusicological Archives and Copyright Issues: an Italian Case Study

Prospero Marra
University of Siena, Italy
prosperomarra@gmail.com

Duccio Piccardi
University of Siena, Italy
duccio.piccardi@unisi.it

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

Abstract

This paper adds a piece to the puzzle of the complex balance between diffusion and legal restraints in the management of oral archives. We focus on the Caterina Bueno Italian ethnomusicological archive, which is being processed by the *Archivio Vi.Vo.* project and represents a challenging case study in terms of protection of the original informants, the author of the arrangements and the other performers. In particular, the paper expounds problems and partial solutions related to authorship, the fixation of the musical performance, its reproduction, diffusion and the compensation for subsequent uses. Overall, the paper aims to promote awareness on legal protection while defusing the apprehension of potential obstacles and dampening excessive risk aversion in the diffusion of oral materials.

1 Moving targets, moving shields: oral archives and legal issues

In his preface to the *Guide to Oral History and the Law*, Neuenschwander (2014: XIII) bluntly stated through the words of attorney H. M. Welch that "[oral history], like any other [occupation], harbors a possibility of inflicting real or imagined injury and wrongs upon others, and those things usually. . . result in litigation". Indeed, when dealing with copyright and privacy infringement, defamation and access management (Rubel, 2007: 171), legal risks lurk behind any discipline concerned with the recording, archiving and diffusion of human voice excerpts. However, several factors hinder the raising of legal awareness in oral discipline practitioners and researchers. For example, the constant evolution of the legal systems calls for cyclic revisions in the research practices. Interestingly enough, the American Oral History Association promoted the presentation of papers on legal issues (Dixon and Zachert, 1968) at its second National Colloquium in 1967. Nonetheless, a few years later, Eustis (1976: 6) warned the Association that the legal context had vastly changed, invalidating previous knowledge on the related issues. Ultimately, it is only through the work of periodic updates offered by Neuenschwander to the Association in the mid eighties that oral historians have welcomed the beginning of a new awareness on legal issues (Swain, 2003: 149)[1].

The effects of obsolescence in knowledge are amplified by the advent of new technologies which, in turn, complexify case typology. In particular, the internet as a storage resource has critically undermined the researcher's sense of guardianship towards oral materials, engendering uncertainty in management choices (Perks, 2009). Surveys conducted in the US (Brewster, 2000) and in the UK (Perks, 2009) stressed the absence of a consistent level of legal knowledge and uniform practices in the web management of oral archives. It comes as no surprise that this general feeling of uncertainty is making web archivists around the world lean toward risk aversion which, despite some recent developments, excessively limits the diffusion and access to the materials (Stobo, 2019: 51-58).

One way to rationalize risk aversion is to enrich our reference bibliography by presenting several case studies which may provide partial guidance to the assessments of those who are dealing with similar issues (Neuenschwander, 2014). Following attempts in Italy (Calamai et al., 2018) and elsewhere

[1] Elsewhere in Europe a renewed awareness towards ethical and legal affairs in oral archives led to the collective writing of guides such as Ginouvès and Gras (editors, 2018).

(e.g., Chaudhuri, 2009), this paper will synthetize the legal problems which are pertinent to the management of a specific type of web resource, that is, in this case, an Italian ethnomusicological archive. In the following chapter (§2), the Caterina Bueno archive will be contextualized in the framework of the ongoing *Archivio Vi.Vo.* project. Then, a selection of legal issues related to authorship, the fixation of the musical performance, its reproduction, diffusion and the compensation for subsequent uses will be explored through the eyes of the management approach of the project (§3)[2]. At the end of each section, a diagram of the pertinent legal topics will be provided (Figures 1-4). Lastly (§4), conclusions will be drawn in the light of the constant struggle between "economicization" and free diffusion of culturally relevant materials.

2 The Caterina Bueno archive

The *Archivio Vi.Vo.* project started in 2019 through the approval and support of the Region of Tuscany. The project is coordinated by the University of Siena in collaboration with CNR-ILC & CLARIN-IT, *Soprintendenza Archivistica e Bibliografica della Toscana* and *Unione dei Comuni del Casentino*[3]. *Archivio Vi.Vo.* aims to build a web infrastructure which will be available to host several regional archives, with particular attention paid to the proposal of reliable procedures concerning the digital philology of oral materials, metadata and audio restoration. Moreover, accessibility and diffusion are at the core of the project: *Archivio Vi.Vo.* strives to make data findable, accessible, interoperable and reusable (FAIR). At the present time, the Caterina Bueno ethnomusicological archive is serving as a first case study and stress test for the constructs and utilities developed in the context of *Archivio Vi.Vo.* The eponym folk singer and researcher, Caterina Bueno (1943-2007), travelled around Tuscany recording the voices of common men and women singing their folk song repertoires. Through musical rearrangements, Bueno tried to preserve this intangible tradition, which became the object of her intense career as a musician. The archive consists of 476 analogue carriers (audio open-reel tapes and compact cassettes), for a total of more than 700 hours of recordings, which were previously digitised (see Calamai et al., 2021 for a brief history of the archive).

Even though, from the legal standpoint, the physical archive is unitarily governed by Art. 816 Civil Code (*universalità delle cose mobili*) as a collection of things belonging to a single person and having a single purpose, this only helps in solving issues related to the actual property of the physical carriers. On the other hand, as Law 633/1941 clarifies that copyrights apply to the contents of the archival materials and are independent of their physical carriers, the noticeable heterogeneity of the archive contents calls for multiple precautions. For this reason, the reader will not find any further mention to the issues related to the physical archive which, in any case, was returned to the original owners after digitisation. With respect to its musical contents, the archive includes not only some of the original field recordings, with the performances by Bueno's informants, but also live events revolving around published pieces and rehearsals of Bueno and the members of her band. That is to say, *Archivio Vi.Vo.* is involved with the assessment of the risks concerning both authorship and performer rights. In addition to that, the transmission of the recordings through the web interface of the project may suggest keeping a strict eye on defining concepts such as diffusion (see below).

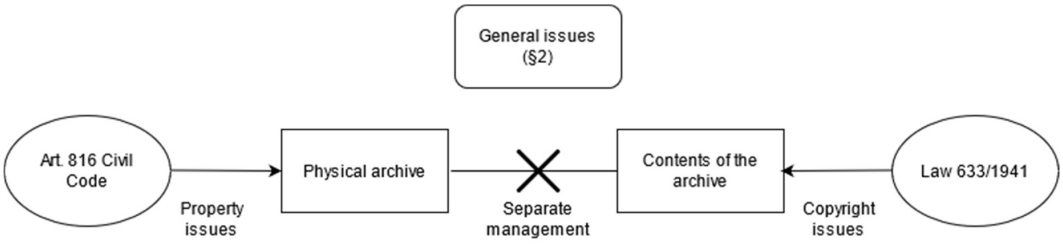


Figure 1. General management of an oral archive

[2] All of the above has been written with the actual Italian legislative context in mind; new rules following EU directive 2019/790 are expected shortly and could change the scenario in a significant way.

[3] The research group is composed as follows: Silvia Calamai, Giovanni Candeo, Monica Monachini, Duccio Piccardi, Niccolò Pretto, Maria Francesca Stamuli. Dissemination activities are coordinated by Pierangelo Bonazzoli.

3 A selection of ethnomusicological copyright hurdles

3.1 Authorship

Two different facets of authorship rights pertain to the management of an ethnomusicological archive: the authorship of the original compositions and that of subsequent rearrangements and executions. As regards the former, it should be noted that the attribution of an orally transmitted composition to a specific author is often unfeasible. Even in the unlikely event that an accurate investigation manages to unequivocally back a claim of composition authorship, Law 633/1941 states that copyright protection ends 70 years after the death of the original author. Thus, composition authorship can hardly be considered a problem in similar ethnomusicological contexts, as the songs are presumably in the public domain.

Rearrangement or execution authorship can be claimed if the original composition incurred in a substantial creative revision. It is unlikely that the improvised elicitations of the source materials can be taken as proof of such endeavour. Therefore, a claim of execution authorship can be more plausibly attempted by the heirs of the folk singer (i.e., Caterina Bueno) or the other members of her band. However, onerous musicological reports are required to back this type of claim as well. In addition to that, the law does not specify the actual level of creative revision required for an attribution of execution authorship, thereby promoting disputability. Lastly, in the case of the Caterina Bueno archive, the donors are also the potential disputers, i.e., the heirs of the musician. Since the donors willingly cooperated with the project's staff and agreed with its general aim, behavioural coherence should be considered a fundamental aspect in the assessment of this kind of legal risk.

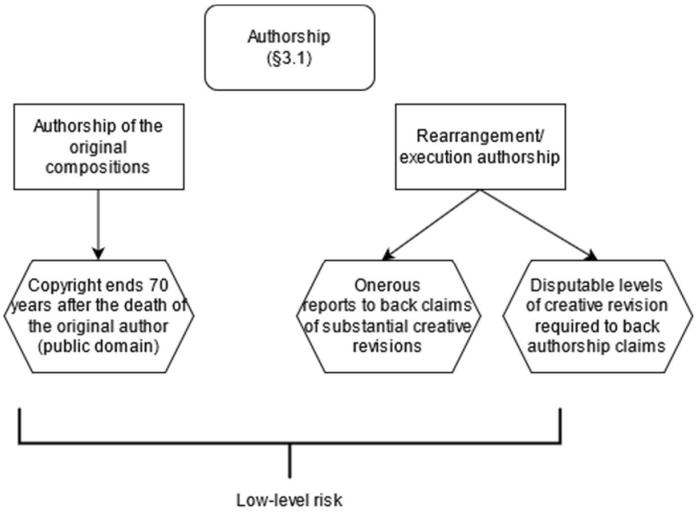


Figure 2. The two facets of authorship rights

3.2 Performer rights

Even though authorship seems to be a minor issue in our ethnomusicological instance, music performers are protected independently of the authorship of the played pieces. Law 633/1941, Art. 80, Par. 2 elucidates the performer rights. Because of their relevance for the management of our archive, we will focus here on the rights concerning the fixation, reproduction and diffusion of music performances.

Fixation rights can be claimed when the recording is not authorized by the performer. With respect to the source materials, it can be safely assumed that Caterina Bueno's informants were aware of being recorded; Bueno always made clear that her intent was to preserve (through fixation) the oral tradition. Since the live events and the rehearsals were recorded by either Bueno or one of the members of her band, the lack of authorization for these contents can be hardly claimed. In the remote chance that specific recordings were taken without the consent of Caterina Bueno and her musicians, the fact that they were welcomed at some point in the Bueno private archive can be interpreted as a manifestation of approval.

Reproduction rights concern the digital copying of the contents of the physical carriers. This legal requirement seems to be per se clear-cut (but see below), since the only exception of the temporary reproduction for technological processes (i.e., caching; Art. 68-bis) is not pertinent to the core management of the archive. Therefore, each of the recorded performers (or their heirs) should ideally authorize the digitisation of the materials.

Lastly, diffusion rights are not always easy to interpret. The original copyright Law 633/1941 and its early reinterpretations were written with respect to synchronous, non-interactive and point-to-mass diffusion (live events, radio broadcasts, television). With the advent of the internet, an effort was made to translate the rules into new communicative contexts. In particular, Art. 80, Par. 2 c and d originally made reference to two different types of diffusion, i.e., live events and diffusion of fixed performances: the judgement by the Milan Tribunal (21/04/2010 n. 4549) equated these types to streaming and downloading services, respectively. Since the assessment of eventual damages depends on the actual propagation of the diffused materials, streaming services are highly preferable for an online archive. Nonetheless, each of the musical performers (or their heirs) should give consent to its streaming diffusion. In addition to that, in the specific case of published materials, producers should give their authorization as well (Art. 72).

The onerousness of providing declarations of consent for each of the music performers represented in the archive (or their heirs) can be mitigated in reference to the legal definition of "orphan work" and "performer". In 3.1 we underlined the difficulties in the identification of an exact author of a folk song. A work is defined as "orphan" when scrupulous investigations fail to pinpoint its copyright holders (Art. 69-quater, Par. 1). This does not apply to authorship only, but also to the identification of performer right holders (Art. 71-decies). Note that Caterina Bueno's informants are often anonymous; in this specific instance, non-profit reproduction and diffusion is allowed (Art. 69-bis). However, this permission is awarded only through a specific certification process, which involves the writing of an extensive report and the approval by the Ministry of Culture.

On a final note, it should be observed that Law 633/1941, Art. 82 leaves room for interpretation in the definition of which kind of performer is entitled to claim copyright. In addition to choir and orchestra leaders, the law protects the performers playing a "noticeable artistic role" in the disputed work. Of course, this phrasing does not unequivocally exclude specific categories of performers from copyright protection. Nonetheless, some Italian music associations of performers are independently trying to circumscribe the features of such "noticeable" individuals. Overall, it seems that terminological uncertainty may add fuel to the disputability of any claim of performer right violation.

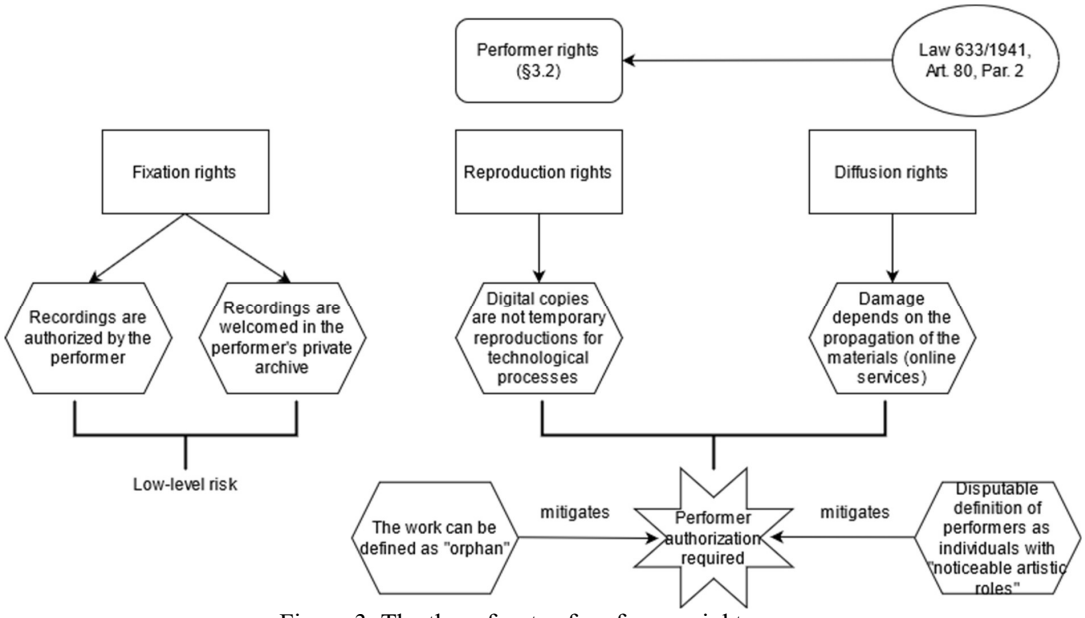


Figure 3. The three facets of performer rights

3.3 Compensation for subsequent uses

Authors, performers and, for published materials, producers are entitled to monetary compensation when their work is repurposed by third parties (Art. 73, 73-bis). This compensation is due in any circumstance, including non-profit repurposing (e.g., our archive) and is considered inalienable. Thus, written authorizations should never include an explicit surrender to compensation. Nonetheless, this should be regarded as a minor concern in the management of the archive, for three main reasons. Firstly, the quantification of monetary compensation for the non-profit repurposing of (mainly) unpublished materials is quite problematic, since estimated profitability usually plays a major role in such estimations. Secondly, the streaming offered by an online archive seems to be excluded from the acceptation of "public" repurposing which triggers the need for compensation (Art. 73). Lastly, the Court of Justice of the European Union (Third Chamber, 15/03/2012, n. 135) has specified that the disputed repurposing should benefit "an indeterminate number of potential listeners, and, in addition, implies a fairly large number of persons". Therefore, the implementation of an authentication system for authorized access to the online archive should be enough to rule out most of the claims for compensation.

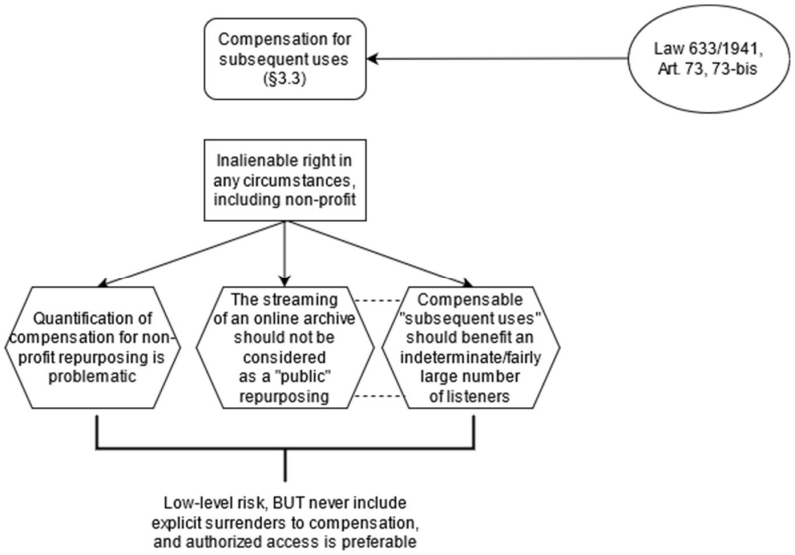


Figure 4. Compensation for subsequent uses

4 Conclusions

This paper presented a set of possible legal obstacles and their foundations when an oral archive is being established and/or managed. Although many such archives are being compiled and offered to different types of users, the absence of a consistent level of legal knowledge and uniform practices in the management of oral archives is still present. Specifically, we assessed the risks of copyright infringement in the online management of the Caterina Bueno Italian ethnomusicological archive. This survey of potential legal issues minimized authorship, performance fixation rights and the compensation for subsequent uses, while underlining the importance of effectively dealing with performance reproduction and diffusion rights. Terminological fuzziness on who is actually entitled to claim those rights is probably not enough to overrule the need to ideally gather consent forms from all the identifiable performers (or their heirs).

Two clashing forces dramatically condition any copyright regulation. One strives to protect the creative efforts of the artists and performers in terms of profit; the other promotes the free diffusion of culturally relevant works. In actual fact, European law systems do not equate these two intents: economic protection is assumed as default, while non-profit cultural promotion has to be contextualized in one of the many "exceptions" to the default. At the present time, those exceptions cannot be generalized, i.e., they cannot constitute alternative categories for the application of copyright regulations (see

the so-called "three-step test"). Therefore, the exceptions have to be defined on a case-by-case basis, and analogical reasoning cannot guarantee the exemption from the default rules. Nonetheless, as risk-aversion runs rampant in the management of online oral archives, presenting legal case-studies can help rationalize common fears and promote procedures to optimize the diffusion of the materials.

References

Brewster, K. 2000. Internet Access to Oral Recordings: Finding the Issues. Oral History Program, University of Alaska Fairbanks.

Calamai, S., Kolletzek, C., Kelli, A. and Biliotti, F. 2018. Authorship and copyright ownership in the digital oral archives domain: The Gra.fo digital archive in the CLARIN-IT repository. *Selected papers from the CLARIN Annual Conference 2017*: 112–127.

Calamai, S., Pretto, N., Stamuli, M. F., Piccardi, D., Candeo, G., Bianchi, S. and Monachini, M. 2021. Community-Based Survey and Oral Archive Infrastructure in the Archivio Vi.Vo. Project. *Selected papers from the CLARIN Annual Conference 2020*: 55–64.

Chaudhuri, S. 2009. *Intellectual Property Management in an Ethnomusicology Archive. An Empirical View from India*. World Intellectual Property Organisation (WIPO), Genève.

Dixon, E. I. and Zachert, M. J. K. 1968. The Second Oral History Colloquium. *The Journal of Library History*, 3(2): 173–178.

Eustis, T. W., III 1976. Get It in Writing: Oral History and the Law. *The Oral History Review*, 4: 6–14, 16–18.

Ginouves, V. and Gras, I., editors 2018. *La diffusion numérique des données SHS: guide des bonnes pratiques éthiques et juridiques*. Presses de l'Université de Provence, Aix-en-Provence.

Neuenschwander, J. A. 2014. *A Guide to Oral History and the Law*, second edition. Oxford University Press, Oxford.

Perks, R. 2009. The challenges of web access to archival oral history in Britain. *IASA Journal*, 32, 74–81.

Rubel, D. T. 2007. Assessing their voice from anywhere: analysis of the legal issues surrounding the online use of oral histories. *Archival Issues*, 31(2): 171–187.

Stobo, V. 2019. Archives, digitisation and copyright: do archivists in the UK avoid risk through strict compliance with copyright law when they digitise their collections? Ph.D. thesis, University of Glasgow.

Swain, E. D. 2003. Oral History in the Archives: Its Documentary Role in the Twenty-First Century. *The American Archivist*, 66(1): 139–158.

Less is more when FAIR. The minimum level of description in pathological oral and written data

Rosalba Nodari
University of Siena, Italy
rosalba.nodari@unisi.it

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

Henk van den Heuvel
Radboud University, Netherlands
H.vandenHeuvel@let.ru.nl

Abstract

This paper presents a case study under the DELAD initiative, on the basis of two different types of data originating in a former neuropsychiatric hospital in Italy: a collection of oral interviews recorded in 1977 by Anna Maria Bruzzone inside the hospital, and a long diary written by a schizophrenic patient in the Seventies. Given the vulnerability of the subjects involved, and the distance in time from the data collection, not all the audio and written material may be accessible). The aim of this work is to address some of the challenges in archiving and storing legacy data referring to vulnerable people in European infrastructures, and to present a minimum set of metadata that can be accessed for further research, according to the FAIR principles.

1 Introduction

Data collections with written or spoken accounts of people with mental disorders are not easy to obtain for research purposes. Considerable effort (and serendipity) is required to find them, but that is only the first hurdle. Permission must then be obtained to use the records for study. Technical challenges may also arise in order to convert the material into digital format to make the data interoperable for analysis with modern technological means. More complications can emerge if the material needs to be shared with other researchers. At all these stages ethical, technical and General Data Protection Regulation (GDPR) issues need to be dealt with. To help researchers do this, the DELAD initiative (see http://delad.net) was established. DELAD stands for Database Enterprise for Language And speech Disorders, (notably: "delad" is Swedish for "shared"). DELAD is an initiative to share corpora among researchers of the speech of individuals with communication disorders (CSD). This is done in a GDPR-compliant way and at secure repositories in the CLARIN infrastructure. DELAD organises workshops focusing on how such corpora can be made shareable with other researchers (Lee et al, 2021). To this end DELAD cooperates with CLARIN data centres such as The Language Archive at the Max Planck Institute (https://archive.mpi.nl/tla/), and Talkbank at CMU (https://talkbank.org/).

In order to offer a corpus of disordered speech shareable with other researchers, the University of Siena research group decided to join the DELAD initiative. The Arezzo Neuropsychiatric Archive preserves different linguistic materials of people with mental disorders. The archive is of notable interest because it offers written and spoken documents of unmonitored speech, unlike other CSD corpora, in which speakers usually perform particular linguistic tasks. It is, furthermore, rich in documentation, offering multiple types of data, and in its population sample. Nevertheless, the corpus in question poses several problems which must be taken into account. We are facing, in fact, not only ethical issues, but also legal issues, given the historical nature of the corpus. The need to find a balance between the safeguarding of vulnerable subjects and the necessity of offering the scientific community a suitable corpus for future research was the reason for starting a feasibility study for storing and archiving the linguistic material present in the Arezzo neuropsychiatric hospital archive into the Language Archive at the Max

Plank Institute for Psycholinguistics in Nijmegen. In the following sections we first offer a description of the Arezzo neuropsychiatric hospital archive. We then focus on the creation of metadata for the written specimens preserved in the Archive according to the best practices for depositing material in the repository bundle of the Language Archive. Finally, we draw conclusions.

2 The historical archive of the neuropsychiatric hospital in Arezzo: its composition, and why it can be important for speech research

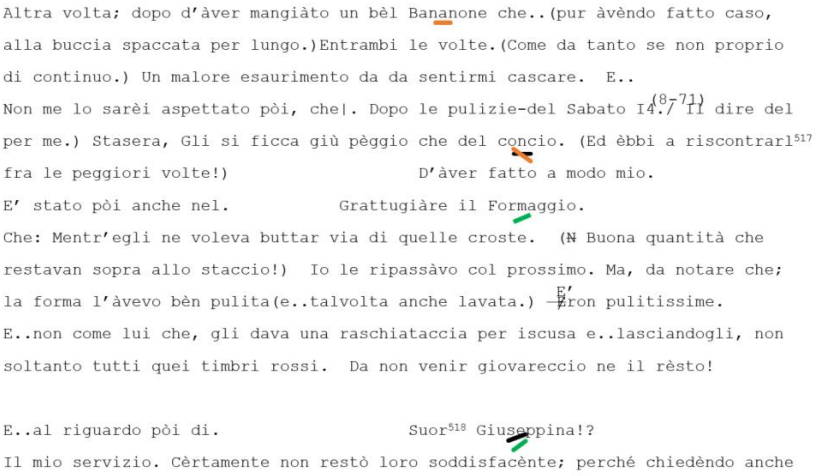
The historical archive of the former Arezzo neuropsychiatric Hospital (Italian acronym ONP) is kept at *Palazzina dell'Orologio* in the Department of Education Sciences, Human Sciences and Intercultural Communication at Siena University, in the Arezzo campus, where the Arezzo sanatorium/mental hospital was originally located. After years of negligence and abandonment following the closure of the Hospital in 1990, the hospital documentation, starting in 1999, was retrieved and catalogued by the University of Siena, in agreement and collaboration with the regional health authority, the owner of the documentation, and the Archival and Bibliographic Superintendence of Tuscany (SabTo). This reorganization made it possible to reconstruct the various original archival series, grouped in two sections and corresponding to the functional sections of the hospital administration: the Directorate for administrative and health affairs, and the Bursar's office for economic management. In 2004, the full inventory of the archive was published (within the series Progetto Archivi of the Province of Arezzo) and can now be consulted online within the "*Carte da legare*" project of the General Directorate for archives (Gherardi, Montani 2004).

The Archive is now open and accessible to university students, researchers and citizens seeking information on their family history. It is composed of about 1500 elements, including files, registers, envelopes, notebooks, and filing cabinets. According to an Agreement renewed in 2019, the University of Siena manages the safekeeping and access in the consultation of the stored material. In addition, a scientific Committee is in charge of enhancing, promoting the study of collections, organizing scientific events and coordinating projects. One of the aims of the Committee is to recover missing archives regarding the neuropsychiatric hospital, and to involve citizens in documentary research and memory conservation, in order to build and preserve hybrid archives (audio, audio-video, paper-based).

Since 2016 several private archives of public figures, who in various ways had a close connection to the Institute of Arezzo, have been traced and collected. Among these is the Bruzzone archive, created by the former teacher and independent researcher Anna Maria Bruzzone. Paola Chiama, Bruzzone's granddaughter and custodian and depositary of Bruzzone's work, decided to progressively donate the researcher's whole archive. Anna Maria Bruzzone conducted various interviews with the patients of the hospital of Arezzo in the summer of 1977. The transcriptions of these interviews were later collated by Bruzzone into the book entitled *'They called us mad. Voices from a mental hospital*' (Bruzzone 1979), republished in 2021. Thanks to her donation, the Arezzo Archive now preserves 19 audiocassettes containing the recordings of the interviews from 34 patients (16 men and 18 women), and 17 other cassettes on different topics. The tapes are associated with the original transcriptions, in different developmental stages – from the Bruzzone handwritten transcription to the published version. The "Anna Maria Bruzzone oral Archive" –which was declared to be of considerable historical and cultural interest in 2018 – is kept at the historical Archive of the former ONP of Arezzo, to which it is ideologically and strongly connected.

Additionally, the archive's nucleus consists in the medical records of patients admitted to the mental and neurological wards. These medical records represent a permanent series in which the files are organized alphabetically and are now preserved in the Directorate section. They consist of two subseries, one devoted to the "mental ward", and the other one to the "neurological ward". Totally, the archive preserves 11,935 medical records of patients hospitalized in the mental unit (see below) and 19,129 medical records regarding the patients in the neurological unit. The medical records are of particular interest in that they contain patients' personal files, such as vital records, administrative and health documents, decrees of partnership and interdiction, medical certificates, family histories, clinical charts, etc., and, until the 1970s, at least one photograph of the patient. In some cases, the charts preserve sections that were directly written by the patients, such as private correspondence, poems, drawings and diaries or autobiographies.

The diary of a former patient is of particular interest. It was meticulously preserved by Fabio Marzi, a psychiatrist who worked at ONP and was loaned to the ONP archive by Gaetano Marzi, the psychiatrist's son. The diary is typewritten, paginated and bound by the author himself, and consists of 323 pages. At the end, it contains a detailed subject index organized according to the places he visited during his life, and to the events he described, in addition to the meetings with people whose stories are told in the diary. The index is organized according to different underlining colours which occur in the text (i.e., the different colours refer to groups of people associated with the Neuropsychiatric Hospital, to family and personal interest, and to different places: see Fig. 1).


Altra volta; dopo d'àver mangiàto un bèl Bananone che..(pur àvèndo fatto caso,
alla buccia spaccata per lungo.)Entrambi le volte.(Come da tanto se non proprio
di continuo.) Un malore esaurimento da da sentirmi cascare. E..
Non me lo sarèi aspettato pòi, che|. Dopo le pulizie-del Sabato 14.7 (8-71) Il dire del
per me.) Stasera, Gli si ficca giù pèggio che del concio. (Ed èbbi a riscontrarl[517]
fra le peggiori volte!) D'àver fatto a modo mio.
E' stato pòi anche nel. Grattugiàre il Formaggio.
Che: Mentr'egli ne voleva buttar via di quelle croste. (N Buona quantità che
restavan sopra allo staccio!) Io le ripassàvo col prossimo. Ma, da notare che;
la forma l'àvevo bèn pulita(e..talvolta anche lavata.) E' ron pulitissime.
E..non come lui che, gli dava una raschiataccia per iscusa e..lasciandogli, non
soltanto tutti quei timbri rossi. Da non venir giovareccio ne il rèsto!

E..al riguardo pòi di. Suor[518] Giuseppina!?
Il mio servizio. Cèrtamente non restò loro soddisfacènte; perché chiedèndo anche

Figure 1. Specimen of the diary (page 59)

Thanks to the medical records, it is now possible to reconstruct personal information about the author. P.A. was a male born in 1916, single, with basic literacy skills and who wanted to join the Church. He was hospitalized two times, the first time from July 1952 to March 1955, the second one in July 1963. Later he returned on his own volition to the hospital, when it was under the direction of Agostino Pirella. On this occasion, he left the hospital only once, in October 1977 for four days.

All this heterogeneous material, by virtue of its nature, needs particular attention in the management, metadata creation and conservation. According to data's peculiar importance (i.e., personal data of vulnerable people) it is indeed crucial to find the right balance between research and the protection of privacy, in order to permit the transmission of knowledge and freedom of research, while maintaining the protection of personal data. In the next section we will propose a minimum set of metadata, in order to make the archive suitable and accessible for researchers interested in disordered speech.

3 The metadata

The Language archive uses the CMDI (CLARIN Metadata Infrastructure) framework as a standard for its descriptive metadata (Broedet et al. 2008; de Vriend et al. 2013). According to TLA deposit manual (https://archive.mpi.nl/tla/deposit-manual-tla), for the Arezzo ONP archive the lat-corpus metadata profile is used as a baseline. The web-based deposit system of TLA includes a webform where the existing metadata profile can be edited for all the relevant collections, sub-corpora and bundles. The CMDI is of profitable use because the CLARIN infrastructure offers researchers the possibility of using ready-made standard component and profile metadata that can be easily adapted to specific linguistic collections. Additionally, the possibility of inserting metadata using the Language Archive web-based interface guarantees a user-friendly tool that does not require any competence in XML syntax. The web interface thus allows to split the work among different collaborators who can then compile the metadata profile online, after having obtained a registered account.

Nevertheless, given the peculiar nature of the Arezzo ONP archive, it is necessary to conduct a preliminary analysis of the architecture of the reference corpus with the aim of selecting the appropriate metadata components and profile from the existing set of metadata. In this respect, at least three crucial

issues have to be taken into account. The first issue is the heterogeneity of the archive. As was mentioned above, the Arezzo Neuropsychiatric Hospital archive contains not only the oral interviews collected by Bruzzone, but also written material (i.e., a private diary). In this regard we have created at the TLA a general collection, called Arezzo Neuropsychiatric Hospital, that, in the future, will contain bundles and other subcollections (such as the Anna Maria Bruzzone archive). In this case, it is then more advisable to consider both granularity – that is, combining components in order to cover just one aspect at a time – and modularity, in order to create a set of metadata that can be suitable for different resources at the same time. For this reason, the first level of metadata profile will thus apply to the collection (i. e. Arezzo ONP archive) and it will contain generic metadata profile, such as Location and Language. Some of these basic components will ~~then~~ be reused for other sublevels of the general collection. Next, different appropriate levels of description will apply to bundles and subcollections.

The second issue to take into account is the peculiar nature of the archive. According to the CMDI best practice guide (https://www.clarin.eu/content/cmdi-best-practices-guide), good component metadata should be "as generic as possible and as specific as needed". This holds particularly true when creating a profile for a corpus that, by virtue of its peculiar nature (i.e. speech and written specimens of vulnerable people), is suitable only for restricted access. Corpora of disordered speech, in fact, usually enclose special categories of personal data (GDPR) such as health information about the patients (see for example van den Heuvel et al. 2020 for a similar case). Thus, metadata creation should not offer sensitive data, such as medical diagnoses, even if these might be useful for other researchers. Nevertheless, the hybrid nature of the archive permits to offer different levels of description depending on specific Bundles and subcollections. For example, the sub-collection Bruzzone archive will contain different Bundles with the complete transcriptions and the different edited versions made by A.M. Bruzzone so that the interviews could be prepared for publishing. However, because the metadata for the Bundle applies to all files within the Bundle, a decision must be made in order to offer a metadata profile that can be applied to all the different versions of the transcription. In fact, the original verbatim transcription contains real names of the patients, that cannot be made publicly available. In this case, the metadata element Actor can be kept anonymous or can contain only the pseudonyms used in the books.

Lastly, the historical nature of the archive makes it necessary to deal with uncertainty. For example, for same patients a medical diagnosis is not available, or the demographic information is incomplete. For this reason, we aim at providing a set of metadata in which it will be possible to infer if some information is not present, rather than omitting possible additional information that could be helpful for other researchers. Following these considerations, we suggest a possible solution for the metadata that will be used for the written material of the archive, that is, at the moment, the schizophrenic patient's diary:

- Name: Diary of a schizophrenic patient
- Title: Diary of a schizophrenic patient / Diario di un paziente schizofrenico
    - Description: Scan and transcription of the personal diary of a schizophrenic patient hospitalised in the Arezzo ONP from July 1952 to March 1955, in July 1963 and, later, under the direction of Agostino Pirella, in the Seventies. The patient wrote the diary his hospitalizations. The diary is typewritten, paginated and bound by the author himself, and consists of 323 pages. At the end, it contains a detailed subject index organized according to the places he visited during his life, and to the events he described, in addition to the meetings with people whose stories are told in the diary. The index is organized according to different underlining colours which occur in the text (i.e., the different colours refer to groups of people associated with the Neuropsychiatric Hospital, to family and personal scope see comment above, and to different places).

- Location
    - Continent: Europe
    - Country: Italy
    - Region: Tuscany
- Project
    - Name: Arezzo ONP archive
    - Contact: […]

    - Description: The corpus contains some of the material preserved in the Arezzo ONP archive. The archive is composed of about 1500 elements, including files, registers, envelopes, notebooks, and filing cabinets, and it documents the history of the Arezzo mental health institution. Along with these materials, the archive comprises different collections of linguistic interest. One of these collections is the Bruzzone archive, created by the former teacher and independent researcher Anna Maria Bruzzone, who conducted various interviews with the patient at the Arezzo ONP. Other collections preserve different specimens that were directly written by the patients, such as private correspondence, poems, drawings and diaries or autobiographies.
- Content
    - Genre: diary
    - SubGenre: spontaneously written
- Content_Languages
    - Content_Language:
    - Id: ISO639-3:ita
    - Name: Italian
- Actors
    - Actor:
    - Role: Patient
    - Name: PA
    - FamilySocialRole: Unspecified
    - EthnicGroup: Unspecified
    - BirthDate: 1916
    - Sex: Male
    - Education: Basic literacy skills
    - Age:
- Resources
    - WrittenResource
        - Date: Unspecified
        - Type: Scan
        - Format: application/zip
        - Size:
    - WrittenResource (Written resource)
        - Date: Trascription
        - Type: Scan
        - Format: application/zip
        - Size:

Following the Deposit Manual of The Language Archive, the metadata will be provided through the online form of the and will be uploaded with the corresponding file, thus creating a bundle with all the resources of the different subcollections.

4 Conclusion

Archiving, managing and sharing corpora of disordered speech is a challenging task, and when managing with legacy data can add even further complications. The option of offering full information when providing metadata will have to, in fact, deal with GDPR issues regarding the protection of special categories of personal data. This feasibility study for storing the linguistic material from the Arezzo ONP archive into the Language Archive aims precisely at striking a good balance between acting in a GDPR-compliant way, while still offering an essential, but still useful, set of metadata suitable for other researchers. However, an effort should be made so as to offer the research community more sensitive information, such as medical diagnoses in the form of documents with restricted access. It is hoped that, at the conclusion of the present work, further studies will be undertaken to assess the feasibility of accessing currently restricted data within the Arezzo ONP archive. DELAD will assist in ensuring compliance to the GDPR guidelines for data protection, transparency, and accessibility.

References

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). OJ L 119, 4.5.2016: 1-88. Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1515793631105&uri=CELEX:32016R0679 (Accessed 14-04-2021).

Broeder, D., Declerck, T., Hinrichs, E., Piperidis, S., Romary, L., Calzolari, N. and Wittenburg, P. 2008. Foundation of a component-based flexible registry for language resources and technology. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC).*

Bruzzone, A. B. 1979. *Ci chiamavano matti. Voci da un ospedale psichiatrico.* Einaudi, Torino.

Bruzzone, A. B. 2021. *Ci* chiamavano matti. Voci da un ospedale psichiatrico. Nuova edizione a cura di Setaro, M. and Calamai, S. Il Saggiatore, Milano.

de Vriend, F., Broeder, D., Depoorter, G., van Eerten, L. and Van Uytvanck, D. 2013. Creating & Testing CLARIN Metadata Components. In *Language Resources & Evaluation (LREC)*, 47: 1315–1326.

Gherardi, S. and Montani, P. 2004. Inventario dell'archivio storico dell'ospedale neuropsichiatrico di Arezzo. Le Balze, Arezzo. http://www.cartedalegare.san.beniculturali.it/fileadmin/redazione/inventari/Arezzo_Ospedale-Neuropscichiatrico.pdf

Kelli, A., Lindén, K., Vider, K., Kamocki, P., Birštonas, R., Calamai, S., Labropoulou, P., Gavriilidou, M. and Straňák, P. (2019). Processing personal data without the consent of the data subject for the development and use of language resources. In *Selected papers from the CLARIN annual conference 2018*, Pisa, 8-10 October 2018. Linköping University Electronic Press: 72-82.

Lee, A., Bessell, N., Van den Heuvel, H., Saalasti, S., Klessa, K., Müller, N., and Ball, M. J. 2021. The latest development of the DELAD project for sharing corpora of disordered speech. Accepted for: *Clinical Linguistics & Phonetics.*

van den Heuvel, H., Kelli, A., Klessa, K., and Salaasti, S. 2020. Corpora of Disordered Speech in the light of the GDPR: two use cases from the DELAD Initiative. In *Proceedings of the 12th Language Resources and Evaluation Conference*: 3317-3321.