# CLARIN Annual Conference 2015
# Book of Abstracts

Wrocław, Poland

October 14–16, 2015

# Preface

This volume contains the abstracts of the presentations at CLARIN2015: CLARIN Annual Conference 2015, held on October 14-16, 2015 in Wrocław, Poland.

The aim of this international conference is to exchange ideas and experiences on the CLARIN infrastructure. This includes the infrastructure's design, construction and operation, the data and services that it contains or should contain, its actual use by researchers, its relation to other infrastructures and projects, and the CLARIN Knowledge Sharing Infrastructure.

There were 33 submissions. Each submission was reviewed by three program committee members. The committee decided to accept 22 abstracts. The abstracts and their titles, author names and affiliations, and keywords were provided by their respective authors. Ten presentations were accepted to be presented orally and twelve as posters. Demos were encouraged. The program also includes an invited talk presented by Andreas Blätte.

I thank all reviewers for their evaluation of the submissions.

October 4, 2015
Bergen

Koenraad De Smedt
Program Committee Chair

# Program Committee

Lars Borin
António Branco
Koenraad De Smedt
Tomaz Erjavec
Eva Hajičová
Erhard Hinrichs
Bente Maegaard
Karlheinz Mörth
Jan Odijk
Rūta Petrauskaitė
Maciej Piasecki
Stelios Piperidis
Kiril Simov
Kadri Vider
Martin Wynne

## Additional reviewers

Daniël de Kok
Maria Gavrilidou

# Table of Contents

# Interaction and dialogue with large-scale textual data: Parliamentary speeches about migration and speeches by migrants as a use case

**Prof. Dr. Andreas Blätte**
Universität Duisburg Essen
andreas.blaette@uni-due.de

## Abstract

My starting point is a discussion on the requirements of an interaction with textual data, on the balance between quantification, visualisation and qualitative analysis, and on the relation between induction and deduction. The resulting views have consequences for the infrastructure you need. Then I will introduce the basic ideas of the implementation of the R environment I work with in the context of the PolMine-project. Finally, I will present work on parliamentary debates on migration and speeches by parliamentarians with a migration background.

# Turkish NLP web services in the WebLicht environment

**Çağrı Çöltekin**
University of Tübingen
Wilhelmstr. 19, 72074 Tübingen
ccoltekin@sfs.uni-tuebingen.de

**Abstract**

This document introduces a number of Turkish natural language processing tools that are being integrated into the CLARIN infrastructure. We describe the tools and resources used in this effort, present their evaluation results, and discuss particular challenges met during this effort both due to some properties of the language and the available resources.

## 1 Introduction

Turkish is a language spoken mainly in Turkey by about 70 million people. It is also one of the major immigrant languages in Europe, e.g., by over 2 million speakers in Germany. Turkish is typologically different from the languages that are well-represented by a large number of tools and resources in the CLARIN infrastructure, and it poses some unusual challenges for the theory and the practice of linguistic research. Hence, besides their practical value because of the large number of speakers, the natural language processing services for Turkish may also be of interest for theoretical (linguistic) research. The services introduced in this document are useful for linguistics research, and potentially for other areas of humanities research.

The main difficulties or differences in computational processing of Turkish are related to its morphological complexity. Morphological complexity, in case of Turkish, does not only mean that the words in the language have a rich set of morphological features. Some linguistic functions that are typically realized in syntax in other languages, e.g., subordinate clause constructions, are realized in morphology in Turkish. As a result, the computational linguistic methods and tools that assume whole words as the units in syntax have a number of difficulties processing Turkish. Example (1) demonstrates one of these problems.

(1) *Sorun* *tarafların* *konuşmamasıydı*
Problem side-PL-GEN talk-NEG-INF-POSS3P-PAST-PERS3P

‘The problem was (the fact that) the parties did not talk.’

Besides the diverse list of morphological features assigned to the last word in (1) that would not fit into a single morphological paradigm, it is clear from the English translation that there are two different predicates in this example (‘be’ and ‘talk’), both having subjects of their own. However, the Turkish sentence does not have two separate words for each predicate. Both predicates are confined into a single word, *konuşmamasıydı*. Furthermore, the negation clearly belongs to the verb *konuş-* ‘talk’, not to the copula, and the past tense marker belongs to the copula (‘was’). As a result, the proper analysis of such sentences requires syntactic relationships between parts of the words, and hence, presenting challenges to typical computational linguistic methods which assume the word is the minimal syntactic unit. In the remainder of this document we describe the way we fit a set of Turkish NLP tools to allow morphological analysis/disambiguation and dependency parsing within the WebLicht environment.

## 2 The Turkish NLP tools in the WebLicht environment

WebLicht (E. Hinrichs et al. 2010; M. Hinrichs et al. 2010) is a natural language processing environment that enables researchers to use NLP web services offered by a large number of institutions. WebLicht allows chaining these services in custom ways to obtain the desired linguistic annotations, and visualize the results through a user-friendly web-based interface. A number of different NLP tasks, e.g., tokenization, POS tagging, dependency or constituency analysis, for a number of languages (including German, English, Dutch, French, Italian) are readily available in the WebLicht environment. WebLicht enables researchers without substantial computational experience to make use of these automated tools. WebLicht is developed within the CLARIN project, and it is fully integrated to the rest of the CLARIN infrastructure. In this section, we describe the new Turkish NLP web services that are being integrated to the WebLicht environment. Some of the tools described here are based on existing tools and resources, and some of them are developed from scratch or improved substantially during the process of integrating them to WebLicht. Although similar efforts exist (most notably Eryiğit 2014), our approach differs in the choice of tools in the pipeline, and integration into the WebLicht provides an easy-to-use and familiar system for the users of the CLARIN infrastructure.

### 2.1 Sentence and word tokenization

Sentence and word tokenization is typically the first task in an NLP pipeline. Since Turkish is written with a Latin-based alphabet, this task is similar to tokenization of most European languages. For both tokenization tasks, we modify existing tokenization services based on Apache OpenNLP, and add statistical models for Turkish. The sentence splitter model is trained using 1 million sentences from the Turkish news section of the Leipzig corpora collection (Quasthoff et al. 2014). The $F_1$-score of the resulting sentence splitter on the same corpus is $95.8\,\%$ (average of 10-fold-cross validation, sd=$0.000\,5$). A qualitative analysis of the results indicates that a sizable part of the mismatch between the model's output and the gold standard is not due to errors made by the model, but errors in the original automatic sentence tokenization. The F-score goes up to $98.7\,\%$ if the model is trained on full Leipzig corpus and tested on about five thousand sentences from the METU-Sabancı treebank (Say et al. 2002).

As noted in Section 1, words are not the tokens that are used for some of the NLP tasks, especially for parsing. The tokens that are input to the parsing can only be determined after the morphological analysis. However, we still need a first tokenization pass for other NLP tasks, including morphological analysis. We train the OpenNLP tokenizer on the METU-Sabancı treebank, which results in a model with $0.44\,\%$ error rate (average of 10-fold cross validation).

### 2.2 Morphological analysis and disambiguation

The morphologically complex nature of the language puts morphological analysis on a central place in Turkish NLP. For morphological analysis, we use the open-source finite-state morphological analyzer TRmorph (Çöltekin 2010; Çöltekin 2014). Besides the morphological analyzer, TRmorph distribution contains a guesser which is useful if the root of the word is unknown. The output of the morphological analyzer for the verb in Example 1 is given in (2).

(2) `konuş⟨V⟩⟨neg⟩⟨vn:inf⟩⟨N⟩⟨p3s⟩⟨0⟩⟨V⟩⟨cpl:past⟩⟨3s⟩`

It is important to note, for the purposes of this work, that the analysis contains multiple part of speech tags. That is, the root *konuş* 'talk' is a verb which is inflected for negation, then the verb becomes a noun (nominal) with the meaning 'the state of not talking', and it again becomes a verb (a nominal predicate). This process is rather different than usual morphological derivation. The crucial difference is that each step in this derivation may participate in syntactic relations outside the word (see in Section 2.3 for an example). The conventional solution for analyzing such words in Turkish NLP involves assuming subword syntactic units called *inflectional groups* (IG). An IG contains a root or a derivational morpheme and a set of inflectional features, or inflections.[1] Each IG may potentially participate in syntactic relations with other IGs outside the word. As well as determining the inflections, or morphological features, of each IG, identifying these IGs is also part of the morphological analysis. Hence, morphological analysis also functions as a tokenization step for the syntactic analysis. For use in WebLicht, we have implemented

[1] This definition is slightly different than earlier use in the literature (e.g., Hakkani-Tür et al. 2002), where derivational morphemes were considered as part of the 'inflectional features'.

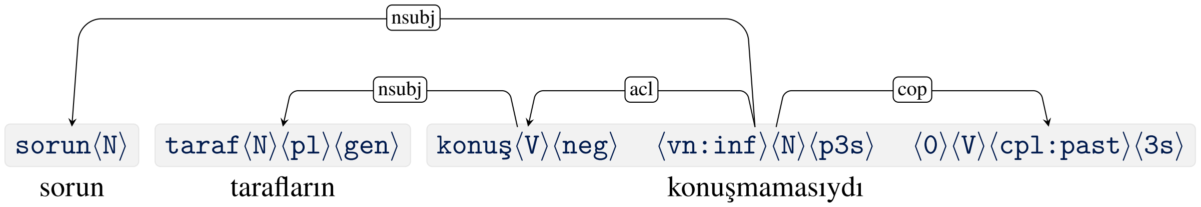


Figure 1: A dependency analysis of Example 1. The dependency labels (roughly) follow Universal Dependencies (`http://universaldependencies.github.io/docs/`). Dependency annotation produced by our parsing service currently follows the METU-Sabancı treebank format which is slightly less intelligible. The morphological tags/labels are from TRmorph.

a simple finite-state decoder that uses the finite-state transducer produced by TRmorph description. For each word, the transducer output is converted to a set of IGs with a POS tag, and a set of morphological features relevant for that POS tag.

Although finite-state tools are efficient at producing analyses strings like in (2), the analyses are often ambiguous. For example, the first word *sorun* in (1), among a few others, can also be analyzed as sor⟨V⟩⟨imp⟩⟨2p⟩ '(please) ask!' or soru⟨N⟩⟨p2s⟩ 'your question'. Disambiguation of these analyses is important, and it has attracted considerable interest in Turkish NLP literature (Hakkani-Tür et al. 2002; Yüret and Türe 2006; Sak et al. 2007, to name a only few). Unfortunately, none of the earlier methods or tools could easily be adjusted to work with TRmorph output. For the present work, we have implemented a new morphological disambiguation system, that we briefly describe below.

Turkish morphological disambiguation is often viewed as a POS tagging with a few additional difficulties. These difficulties include (1) the data sparseness problems due to large tagset, (2) the difficulties of applying standard POS tagging algorithms because of variable number or inflectional groups in alternative analyses of a word and (3) the limited utility of features extracted from the local context of words in history-based POS tagging algorithms due to free-word-order nature of the language.

Like earlier work, we alleviate the data sparseness problem by making use of IGs. However, we do not view the morphological disambiguation in the usual setting of sequence labeling with hidden (POS) labels. We exploit the fact that the analyzer limits the choices for possible morphological analysis of a word, and the analyzer output is available in both training and testing time for the complete sentence. We extract a set of features, $\Phi$, from all analyses offered by the morphological analyzer for the input sentence. Some features depend on the position of the word containing the IG, e.g., 'the last POS tag of the previous word', some features are word-internal, e.g., 'the word is capitalized', and others are general features extracted from the sentence or the available analyses of it, e.g., 'the analyses contain a finite verb'.

Recalling that an IG contains a root (or derivational morpheme) $r$, a POS tag $c$, and a set of inflections $f$, we assume that given $\Phi$, analysis of a word is independent from the other words in the sentence, and similarly, an IG is independent of the other IGs in the same word given $\Phi$. This allows us to define analysis of a word with $m$ inflectional groups as,

$$\prod_{i=1}^{m} P(f|r,c,\Phi)P(r|c,\Phi)P(c|\Phi) \qquad (1)$$

We estimate components of Equation 1 using discriminative models (logistic regression models for the results reported here). The resulting disambiguator has an accuracy of 91.2 % on the METU-Sabancı treebank with 10-fold cross validation. Although the results may not be directly comparable due to use of different morphological analyzers, this is similar to earlier results obtained on the same data set using Sak et al.'s (2007) disambiguator by Çetinoğlu (2014) (90.2 % with a similar setting).

### 2.3 Dependency parsing

Since the syntactic relations in Turkish are between inflectional groups, rather than words, the dependency links relate IGs. A dependency parse of Example 1 is given in Figure 1.

For dependency parsing, we currently include an additional model to an already existing web service based on MaltParser (Nivre et al. 2006). The model is trained on the METU-Sabancı treebank. We use the version used in CoNNL-X shared task (Buchholz and Marsi 2006) with minor corrections. The resulting model has a labeled attachment score of 66.8 and unlabeled attachment score of 77.2 (10-fold-cross validation on the METU-Sabancı treebank). The results are obtained with coarse POS tags, with default learning method and without additional features or optimization. Initial experiments with additional features did not yield substantially better results. The (rather low) numbers we obtain are similar to earlier parsing results reported in the literature. Parsing Turkish was found to be difficult in earlier studies (Buchholz and Marsi 2006). Part of this difficulty seems to stem from the properties of the language, some of which are discussed above. However, our initial impression is that difficulty also stems from the small and not-very-high-quality resources available for the language. The only treebank available for Turkish (METU-Sabancı treebank) contains only 5 635 sentences and 56 424 tokens, and includes many annotation errrors and some unusual annotation choices.

## 3 Concluding remarks

We summarized the effort of integrating a Turkish NLP pipeline into the WebLicht infrastructure. The pipeline contains web services for sentence and word tokenization, morphological analysis and disambiguation, and dependency parsing. For some of the services, we used existing tools with some improvements and customization, and for others we developed some in-house tools. The tools and services described in this document are fully implemented and ready for use, and we are still improving some of the services. The services only make use of freely available tools, and the tools developed during this work will also be made available with a free/open-source license.

## References

[Buchholz and Marsi 2006] Buchholz, Sabine and Erwin Marsi (2006). *CoNLL-X shared task on multilingual dependency parsing*. In: *Proceedings of the Tenth Conference on Computational Natural Language Learning*, pp. 149–164.

[Çetinoğlu 2014] Çetinoğlu, Özlem (2014). *Turkish Treebank as a Gold Standard for Morphological Disambiguation and Its Influence on Parsing*. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Ed. by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Reykjavik, Iceland: European Language Resources Association (ELRA). ISBN: 978-2-9517408-8-4.

[Çöltekin 2010] Çöltekin, Çağrı (2010). *A freely available morphological analyzer for Turkish*. In: *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*. Valetta, Malta, pp. 820–827.

[Çöltekin 2014] Çöltekin, Çağrı (2014). *A set of open source tools for Turkish natural language processing*. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. Reykjavik, Iceland: European Language Resources Association (ELRA).

[Eryiğit 2014] Eryiğit, Gülşen (2014). *ITU Turkish NLP Web Service*. In: *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Gothenburg, Sweden: Association for Computational Linguistics, pp. 1–4.

[Hakkani-Tür et al. 2002] Hakkani-Tür, Dilek Z., Kemal Oflazer, and Gökhan Tür (2002). *Statistical Morphological Disambiguation for Agglutinative Languages*. In: *Computers and the Humanities* 36.4, pp. 381–410.

[E. Hinrichs et al. 2010] Hinrichs, Erhard, Marie Hinrichs, and Thomas Zastrow (2010). *WebLicht: Web-Based LRT Services for German*. In: *Proceedings of the ACL 2010 System Demonstrations*. Uppsala, Sweden: Association for Computational Linguistics, pp. 25–29.

[M. Hinrichs et al. 2010] Hinrichs, Marie, Thomas Zastrow, and Erhard Hinrichs (2010). *WebLicht: Web-based LRT Services in a Distributed eScience Infrastructure*. In: *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. Ed. by N. Calzolari, K. Choukri, B. Maegaard,

J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias. Valletta, Malta: European Language Resources Association (ELRA). ISBN: 2-9517408-6-7.

[Nivre et al. 2006] Nivre, Joakim, Johan Hall, and Jens Nilsson (2006). *Maltparser: A data-driven parser-generator for dependency parsing*. In: *Proceedings of LREC*, pp. 2216–2219.

[Quasthoff et al. 2014] Quasthoff, Uwe, Dirk Goldhahn, and Thomas Eckart (2014). "Building Large Resources for Text Mining: The Leipzig Corpora Collection". In: *Text Mining*. Ed. by Chris Biemann and Alexander Mehler. Theory and Applications of Natural Language Processing. Springer, pp. 3–24. ISBN: 978-3-319-12654-8. DOI: 10.1007/978-3-319-12655-5_1.

[Sak et al. 2007] Sak, Haşim, Tunga Güngör, and Murat Saraçlar (2007). *Morphological Disambiguation of Turkish Text with Perceptron Algorithm*. In: *CICLing 2007*. Vol. LNCS 4394, pp. 107–118.

[Say et al. 2002] Say, Bilge, Deniz Zeyrek, Kemal Oflazer, and Umut Özge (2002). *Development of a Corpus and a TreeBank for Present-day Written Turkish*. In: *Proceedings of the Eleventh International Conference of Turkish Linguistics*. Eastern Mediterranean University, Cyprus.

[Yüret and Türe 2006] Yüret, Deniz and Ferhan Türe (2006). *Learning morphological disambiguation rules for Turkish*. In: *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*. HLT-NAACL '06. New York, pp. 328–334. DOI: 10.3115/1220835.1220877.

# Bootstrapping a neural net dependency parser for German using CLARIN resources

**Daniël de Kok**
University of Tübingen
Wilhelmstraße 19
72074 Tübingen
daniel.de-kok@uni-tuebingen.de

## Abstract

Statistical dependency parsers have quickly gained popularity in the last decade by providing a good trade-off between parsing accuracy and parsing speed. Such parsers usually rely on hand-crafted symbolic features and linear discriminative classifiers to make attachment choices. Recent work replaces these with dense word embeddings and neural nets with great success for parsing English and Chinese. In the present work, we report on our experiences with neural net dependency parsing for German using CLARIN data.

## 1 Introduction

A dependency parser transforms a sentence $w_0^n$ with the artificial root token $w_0$ in a dependency graph $G =< V, A >, V \subseteq w_0^n$ and $A \subseteq V \times R \times V$, wherein R is a set of dependency labels. A transitional dependency parser (Nivre, 2003; Kübler et al., 2009) performs this transduction using a transition system, where the states are configurations consisting of a buffer $\beta$ of unprocessed tokens, a stack $\sigma$ of partially processed tokens, and $A$ the set of dependency arcs that were found so far. The transitions transform the configuration, typically by introducing a leftward arc, a rightward arc, or moving a token from $\beta$ to $\sigma$. Transformations are applied until the system reaches a final configuration wherein $\beta$ is empty and $\sigma$ only contains the artificial root token.

In a particular configuration $c$, there are often multiple possible transitions. A classifier $\lambda(t|c)$ is used to find the most probable transition $t$ in a configuration $c$. If the parser only considers the most probable transition, parsing time is linear in the commonly-used transition systems. The classification function $\lambda$ is typically a linear classifier that is parametrized using features that decompose the configuration that is under consideration. $\lambda$ is trained by running a parser over the training data using an oracle that produces a sequence of transitions that results in $A = A_{gold}$.

## 2 Neural net parsers

Chen and Manning (2014) propose a transition-based parser that uses a neural net for $\lambda$. Moreover, the input of the parser does not consist of symbolic features, but the concatenation of (dense) word, tag, and relation embeddings of tokens that are in interesting positions of the parser configuration.

Their approach has three potential benefits over dependency parsing with a linear model: (1) the use of a non-linear activation function in the hidden layer(s) avoids the need to carefully craft effective feature combinations; (2) the use of word embeddings can counter data sparseness, e.g. *dog* and *puppy* will behave similarly with regards to syntax and selectional preferences; and (3) transition-based dependency parsers that use symbolic features are known to spend most of their time on feature extraction (Bohnet, 2010), while the use of embeddings only requires a small number of lookups and fast dense matrix-vector multiplications.

Chen and Manning (2014) report large improvements in accuracy over existing transition-based parsers such as Malt and the graph-based MSTParser for English and Chinese. Of course, their work leaves some interesting questions: how well does this approach work for other languages and how stable are the network topology and chosen hyperparameters? Our goal is to reproduce this approach for German using CLARIN resources, with the twofold goal of validating their results and bootstrapping a neural net dependency parser for German.

## 3 Training of embeddings

Since the input of a neural net parser consists solely of embeddings, it is of prime importance that the embeddings are of a high quality. It is well-known that such embeddings can only be obtained using huge training sets (Mikolov et al., 2013).

Although a vast amount of textual data for German is available through the CLARIN-D project,[1] we also require part-of-speech and morphological annotations without any divergence in annotation schemes. To this end, we use two large corpora for German: TüBa-D/W (de Kok, 2014) and TüPP-D/Z (Müller, 2004). The TüBa-D/W is a 615 million token treebank of German Wikipedia text that provides part-of-speech and morphological annotations. The TüPP-D/Z is a 204 million token corpus of text from the German newspaper Taz. To ensure that the annotation schemes for part-of-speech tags and morphology are the same as in TüBa-D/W, we apply the annotation pipeline described in De Kok (2014), except that we retain the token and sentence boundaries of TüPP-D/Z. We also remove sentences that are part of TüBa-D/Z (Telljohann et al., 2004). Our training data for the embeddings consist of the concatenation of these two corpora (818 million tokens, 48 million sentences). We prepend a special root token to each sentence that is used internally by the parser.

We train the embeddings using an adaptation of Word2Vec (Mikolov et al., 2013) that uses a structured variant of the skip-gram model (Ling et al., 2015). This adaptation uses a dedicated output matrix for each position in token's context window. As a result, the embeddings capture order dependence, making them better suited for syntax-based tasks. We use the same training hyperparameters as proposed in Ling et al. (2015) and a minimum token frequency of 30.[2] Experiments with other hyperparameters did not provide a consistent improvement.

The dependency parsers were trained and evaluated using the dependency version (Versley, 2005) of the TüBa-D/Z (Telljohann et al., 2004). We first split the treebank into five (interleaved) parts. Two parts were used for training and validation. The remaining three parts were held out and used for the evaluation. The morphology annotations in TüBa-D/Z are replaced by annotations provided by RFTagger (Schmid and Laws, 2008) to use the same tag set as TüBa-D/W and the reprocessed TüPP-D/Z.

## 4 Parser configurations

### 4.1 SVM parser

We use the dpar[3] parser as the 'traditional' parser in our comparison. dpar is a linear SVM-based parser that is architecturally similar to MaltParser (Nivre et al., 2006), adding hash kernels (Shi et al., 2009; Bohnet, 2010), and providing more fine-grained control over morphology features. We use this parser and the neural net parser with the stack-projective transition system.

The initial set of feature templates is obtained by applying MaltOptimizer (Ballesteros and Nivre, 2012) to the training and validation data. We then add extra hand-crafted morphology-based feature templates,[4] such as: take the *number* morphological feature of $\sigma_1$ and $\sigma_2$ (the two tokens on top of the stack), and the *case* feature of token $\sigma_1$.

### 4.2 Neural net parser

**Parser input** The dparnn parser is used as the neural net parser in our comparison. Since a neural net that uses a non-linear activation in its hidden layer can infer complex features, the parser does not need feature templates. Instead, we specify the positions of the stack ($\sigma_0^{n-1}$) and buffer ($\beta_0^{n-1}$) that the parser is allowed to use, together with the information layers. The possible layers are: TOKEN, TAG, MORPH, and DEPREL (the

[1] `http://clarin-d.de/de/auffinden/referenzressourcen`

[2] The embeddings are available at: `http://hdl.handle.net/11022/0000-0000-8507-2`

[3] `http://github.com/danieldk/dpar`

[4] The feature template specification is included in the *dpar* repository.

relation to a token's head). In addition, the LDEP and RDEP indirections can be used, which give the leftmost and rightmost dependent of a token. For feature formation, we provide the parser with the following: TOKEN($\sigma_0^3$), TOKEN($\beta_0^2$), TAG($\sigma_0^3$), TAG($\beta_0^2$), TOKEN(LDEP($\sigma_0^1$)), TOKEN(RDEP($\sigma_0^1$)), TAG(LDEP($\sigma_0^1$)), TAG(RDEP($\sigma_0^1$)), DEPREL($\sigma_0$), DEPREL(LDEP($\sigma_0^1$)), and DEPREL(RDEP($\sigma_0^1$)). For experiments with morphology, we also give the parser access to MORPH($\sigma_0^1$) and MORPH($\beta_0$).

**Embeddings** For word embeddings and morphology embeddings, we always use the embeddings obtained through unsupervised training as described in Section 3. The embeddings for dependency relations are trained during the training of the parser, by learning weights through backpropagation to the relation vectors. For part-of-speech tags, we explore two possibilities: learning embeddings unsupervised from a large corpus following Section 3 and learning embeddings through backpropagation while training the parser. Chen and Manning (2014) did not explore unsupervised training of tag embeddings nor unsupervised training of morphology embeddings.

**Training** The training data for the neural network is obtained by running the parser over the training data using an oracle which extracts each transition and the corresponding parser input. We normalize each embedding using its $\ell_2$ norm and the input vector to the neural net per index by its mean and variance.

We use Caffe (Jia et al., 2014) to train the network. Since Caffe was developed for image processing, we represent each input as an $1 \times 1$ image with the number of channels that is equal to the input size. To train tag and relation embeddings, we take the following approach: tags and dependency relations are encoded using one-hot encoding in the input. We create a hidden linear layer for each input position of the form TAG($\cdot$) and DEPREL($\cdot$). The weights of layers of a particular type (TAG and DEPREL) are shared. The output of these hidden layers, along with the remainder of the input, is the input of a hidden layer with a non-linear activation function. After training, the aforementioned hidden linear layers contain the tag and relation embeddings – the $n$-th weight of the $m$-th neuron stores the $m$-th element of the embedding vector of the tag/relation corresponding to the $n$-th one-hot bit. In our parser, we use 50 neurons in the embedding layers, to obtain embeddings of size 50. The network topology that is used during training is shown schematically in Figure 1. After training, we extract the embeddings, to simplify the network topology during parsing so that it does not need the hidden linear layers.

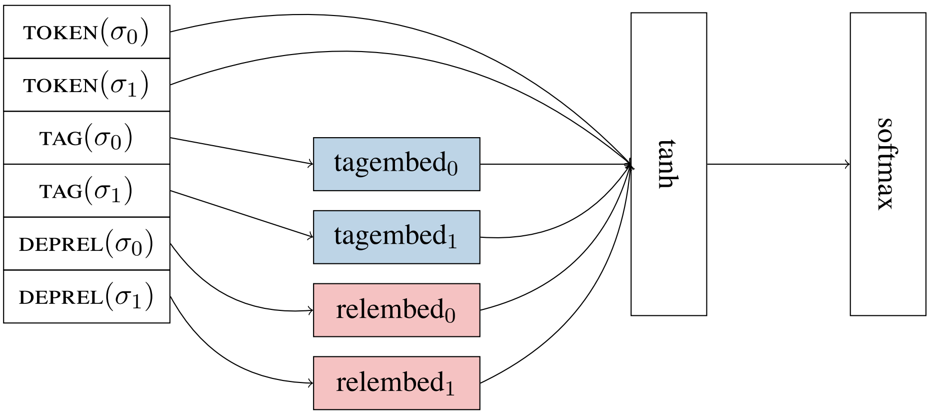


Figure 1: Network topology used during training, with supervised training of tag and relation embeddings. Embeddings with the same color have shared weights. The number of inputs is reduced for illustrative purposes.

As Chen and Manning (2014), we use a non-linear hidden layer with 200 neurons, an output layer that uses the softmax activation function, and the Adagrad solver (Duchi et al., 2011). However, their other hyperparameters, a cube activation function for the hidden layer and 50% dropout (Srivastava et al., 2014), did not work well in our experiments. Instead, opted for more conservative hyperparameters: a hyperbolic tangent activation function for the hidden layer, and mild input layer and hidden layer dropout of 10% and 5% respectively.

## 5 Results

In our discussion of the results, we will start with the most basic versions of both types of parsers: dpar (*dpar*) and dparnn trained with unsupervised word embeddings and supervised tag and relation embeddings (*dparnn STE*). Then, we will discuss dparnn with tag embeddings learned through unsupervised training (*dparnn UTE*). Finally, we will look at the effect of giving access to morphological information to both dpar (*dpar morph*) and dparnn (*dparnn UTE morph*).

As can be seen in Table 1, the *dparnn STE* parser clearly outperforms *dpar*. We expect that this is the case because the neural net is able to infer features that are difficult to get by hand-crafting feature templates or doing a greedy search such as that of MaltOptimizer. Moreover, since the word embeddings are trained on a very large corpus, the *dparnn STE* has much better lexical coverage (97.13% of tokens, 75.73% of types) of the evaluation data than *dpar* (89.32% of tokens, 30.67% of types).

| Model | LAS |
|---|---|
| dpar | 88.28 |
| dpar (morphology) | 88.85 |
| dparnn STE | 88.83 |
| dparnn UTE | 89.08 |
| dparnn UTE + Morphology | **89.40** |

Table 1: Results of parsing using SVM classifiers using symbolic features and neureal nets using word embeddings.

Using unsupervised tag embeddings improves the result by 0.25%. This is perhaps surprising, because the unsupervised tag embeddings were trained on non-gold standard data. However, it seems that the sheer amount of examples puts each tag more accurately in vector space.

Finally, adding morphological information to the mix improves the SVM parser profoundly (0.57%), while also providing an improvement for the neural net parser (0.32%). In the final parser configurations, the neural net parser outperforms the SVM parser by 0.55%.

## 6 Conclusion

In this work we have shown that neural net dependency parsers that rely on word embeddings outperform 'traditional' SVM dependency parsers on German. Since neural net parsers rely on large corpora of automatically annotated text, efforts to create and distribute such corpora are paramount to the improvement of the state of the art in parsing. In future work, we hope to address our poor man's treatment of morphology, by making morphological analysis an integral part of parsing.

## References

[Ballesteros and Nivre2012] Miguel Ballesteros and Joakim Nivre. 2012. MaltOptimizer: an optimization tool for MaltParser. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 58–62. Association for Computational Linguistics.

[Bohnet2010] Bernd Bohnet. 2010. Very high accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 89–97. Association for Computational Linguistics.

[Chen and Manning2014] Danqi Chen and Christopher D Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, volume 1, pages 740–750.

[de Kok2014] Daniël de Kok. 2014. TüBa-D/W: a large dependency treebank for german. In *Proceedings of the Thirteenth International Workshop on Treebanks and Linguistic Theories (TLT13)*, page 271.

[Duchi et al.2011] John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, 12:2121–2159.

[Jia et al.2014] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional Architecture for Fast Feature Embedding. *arXiv preprint arXiv:1408.5093*.

[Kübler et al.2009] Sandra Kübler, Ryan McDonald, and Joakim Nivre. 2009. Dependency parsing. *Synthesis Lectures on Human Language Technologies*, 1(1):1–127.

[Ling et al.2015] Wang Ling, Chris Dyer, Alan Black, and Isabel Trancoso. 2015. Two/too simple adaptations of word2vec for syntax problems. *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL), Denver, CO*.

[Mikolov et al.2013] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

[Müller2004] Frank Henrik Müller. 2004. Stylebook for the tübingen partially parsed corpus of written German (TüPP-D/Z). In *Sonderforschungsbereich 441, Seminar für Sprachwissenschaft, Universität Tübingen*, volume 28, page 2006.

[Nivre et al.2006] Joakim Nivre, Johan Hall, and Jens Nilsson. 2006. Maltparser: A data-driven parser-generator for dependency parsing. In *Proceedings of LREC*, volume 6, pages 2216–2219.

[Nivre2003] Joakim Nivre. 2003. An efficient algorithm for projective dependency parsing. In *Proceedings of the 8th International Workshop on Parsing Technologies (IWPT)*, pages 149–160.

[Schmid and Laws2008] Helmut Schmid and Florian Laws. 2008. Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 777–784. Association for Computational Linguistics.

[Shi et al.2009] Qinfeng Shi, James Petterson, Gideon Dror, John Langford, Alex Smola, and SVN Vishwanathan. 2009. Hash kernels for structured data. *The Journal of Machine Learning Research*, 10:2615–2637.

[Srivastava et al.2014] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

[Telljohann et al.2004] Heike Telljohann, Erhard Hinrichs, and Sandra Kübler. 2004. The TüBa-D/Z treebank: Annotating German with a context-free backbone. In *In Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004*. Citeseer.

[Versley2005] Yannick Versley. 2005. Parser evaluation across text types. In *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories (TLT 2005)*.

# INESS meets CLARINO

**Koenraad De Smedt**
University of Bergen
Bergen, Norway
desmedt@uib.no

**Gunn Inger Lyse**
University of Bergen
Bergen, Norway
gunn.lyse@uib.no

**Paul Meurer**
Uni Research Computing
Bergen, Norway
paul.meurer@uni.no

**Victoria Rosén**
University of Bergen
Bergen, Norway
victoria@uib.no

### Abstract

The INfrastructure for the Exploration of Syntax and Semantics (INESS) is an extensive online system which supports research on treebanks. Its services address CLARIN goals and have become closely integrated in the CLARINO Bergen center (Norway).

## 1 Introduction

The rich annotation in treebanks makes them good sources for empirical research on syntax and the lexicon, for work on parsers, and to some extent also for exploring semantics, for information extraction, and for other 'deep' processing. The accessibility of treebanks is therefore important for several target researcher groups in CLARIN.

Work on INESS[1] began in 2010, two years before the start of the CLARINO project, which implements and runs the Norwegian part of CLARIN. The INESS project has two main goals: (1) the construction of NorGramBank, a large Norwegian parsebank (i.e. treebank obtained by parsing), and (2) making treebanks more accessible. While the former will not be discussed here, the latter goal is relevant for a wide CLARIN audience. One of the objectives of the CLARINO project has therfore been to integrate INESS into the world of CLARIN.

Although hundreds of treebanks exist which are potentially useful for research, their effective exploitation has until recently often been impeded by practical issues regarding distribution, access, metadata, licensing and search tools. Treebanks have typically been distributed over many different websites and have not had compatible metadata. Search within treebanks has often required downloading the data to one's own computer as well as downloading and installing standalone tools. Furthermore, query languages and tools were often specific to certain annotations and formats. Such limitations are typical of the kinds of problems that CLARIN in general wants to see solved.

In recent years, some treebanking efforts linked to CLARIN projects have started to address these issues. For instance, whereas the standalone tool Dact[2] already provides a user-friendly alternative to the earlier Alpino treebank tools (van Noord et al., 2013), online search alternatives for these tools have also become available, such as the example-based Gretel (Vandeghinste and Augustinus, 2014) and PaQu[3] which is especially handy for relations between word pairs.

Access to the Czech treebanking resources and services has also considerably evolved. The distribution of treebanks in the LINDAT repository (based on DSpace) has become well integrated in the overall CLARIN architecture by the consistent use of CMDI metadata (Broeder et al., 2010), persistent identifiers (PIDs), federated authentication and license handling. The current LINDAT infrastructure offers a wide selection of dependency and constituency treebanks for different languages which can be individually searched and visualized through its online service PML Tree Query.[4]

Taking another approach at CLARIN integration, the TüNDRA[5] web tool for treebank research is

[1] http://clarino.uib.no/iness/

[2] http://rug-compling.github.io/dact/

[3] http://zardoz.service.rug.nl:8067/info.html

[4] http://lindat.mff.cuni.cz/services/pmltq/

[5] http://weblicht.sfs.uni-tuebingen.de/Tundra/

accessible online in WebLicht (Martens, 2013). It provides federated authentication, browsing, search and visualization for TüBA treebanks and some other treebanks with constituency or dependency annotation. WebLicht also offers a service for building one's own parsebank.[6]

The INESS infrastructure is similar to these efforts in its goal of making treebanks more accessible, but handles a larger range of treebank types. INESS hosts treebanks of any current annotation type and format. This includes structural representations in Lexical Functional Grammar (LFG) and Head-Driven Phrase Structure Grammar (HPSG), besides constituency annotation and three current flavors of dependency annotation. It also handles parallel treebanks, even those combining different annotation types. INESS handles any Unicode-compatible language, alphabet and writing system and currently hosts treebanks for 43 languages.

In order to offer more uniform exploration, the online search tool INESS-Search has a readable, compact and expressive query language (Meurer, 2012) which shares important syntax features across annotation frameworks. Thus, notations for categories and words, operators for dominance and precedence, etc. are the same, regardless of the grammatical approach or type of annotation, to the largest extent possible. It also allows simultaneous search in several treebanks selected by the user, in other words, virtual treebank collections can be defined as search domains.

Similarly, INESS offers visualization of common structural representations in any type of treebanks and has user dependent options for visualization preferences (e.g. tree or arc diagrams for dependencies). INESS also supports online parsing, interactive parse selection and parsebank construction with LFG grammars and discriminant disambiguation. Briefly, INESS has evolved into a virtual laboratory for treebank management and exploration (Rosén et al., 2012, inter alia).

The remainder of this abstract describes how INESS has become better integrated with respect to CLARIN standards, how it addresses needs of the CLARIN user community, and how INESS together with LINDAT have formed a K-center in the CLARIN Knowledge Sharing Infrastructure (KSI).

## 2 Metadata and PIDs

INESS currently provides access to more than 200 treebanks and therefore offers treebank selection based on user choices, which currently include language, collection, annotation type, and linguality (monolingual or parallel). Initially, INESS accommodated treebanks for META-SHARE (Losnegaard et al., 2013), but since INESS wants to comply to CLARIN requirements, it has started to manage CMDI metadata. The CLARINO working group on metadata found, however, that it was difficult to identify suitable existing profiles and components for treebanks and other resources in the Component Registry.[7] In particular, the CMDI profiles derived from the corresponding META-SHARE schemata did not have sufficient descriptive coverage. Therefore, improved CLARINO profiles and components were developed, including the *corpusProfile* which is being applied to all treebanks in INESS.

As an example, META-SHARE had individual components for the different *actor roles* that persons and organizations may have in relation to a resource, e.g. one component for *creatorPerson*, one for *creatorOrganization* and so on for any role such as IPR holder, funder, validator, annotator, etc. However, the existing variety of *actor role* components did not always cover the descriptive needs seen in CLARINO. For instance, the META-SHARE *author* component did not seem to accommodate the author's year of death, which may be quite relevant for historical texts, which form the source of several treebanks relevant to e.g. philology studies. Although building on the different META-SHARE components for person and organization roles, CLARINO decided to collapse all of them into a generic component *actorInfo*, applicable for any role, independently of whether the *actor* is a person or an organization, and where a sufficient number of elements are available (e.g. year of death). Replacing all role-specific components with the generic component *actorInfo* meant that a considerable number of original META-SHARE components had to be replaced by new CLARINO components, even if each of the changes is a minor one.

Another significant departure from META-SHARE concerns the license component in the metadata. CLARINO has modified this component so as to include the three recommended main usage categories

[6] `http://danieldk.eu/Research/Publications/cac2014.pdf`

[7] `http://catalog.clarin.eu/ds/ComponentRegistry/`

(‘laundry tags’) PUB (public), ACA (academic) or RES (restricted).

The obligatory component *resourceCommonInfo* contains fields for general information relevant for all resource types, and is required in CLARINO to facilitate search across profiles, by ensuring that some basic information is always present, similarly to the minimal metadata schema in META-SHARE. This includes basic information such as resource type, resource name, identifiers (e.g. PIDs), licences, origin, owner, contact information, metadata details, version info, validation info and relation to other resources. The improved CLARINO profiles and components, which are being applied to resources in INESS, are stable and will soon be published in the Component Registry.

When sufficient metadata are provided for a treebank, a CLARIN compatible persistent identifier (*handle*) is created which redirects to a landing page displaying the metadata. These metadata are available in compact and full views. However, in its effort to host as many treebanks as possible, INESS is currently also hosting treebanks for which documentation has not yet been fully supplied.

## 3 Access policies and licenses

CLARIN depositor’s license agreements are as far as possible used for new treebanks, and INESS follows the CLARIN recommendation to make resources as freely and openly available as possible. However, INESS also accommodates treebanks with legacy licenses from different sources which may impose restrictions. In line with CLARIN recommendations, INESS streamlines the description of licenses using the CLARIN license categories.[8]

Treebanks which are not public require that users log in. Like LINDAT, INESS has implemented SAML 2.0-based single sign-on covering the CLARIN IdP and federations participating in eduGAIN.[9] Selection of one’s ID provider is achieved through integration of the DiscoJuice[10] discovery service by Uninett.

Many treebanks have a complex provenance; futhermore, the license conditions may vary according to the type of access (more open license for access through the INESS search interface, more limited access for downloadability). Therefore, INESS is able to handle multiple licenses. Indeed, the *Distribution info* component in the CMDI metadata may contain more than one *License info* component. This is the case, for instance, in the *Distribution info* for the BulTreeBank,[11] which has different licenses for users wishing to search it (*Distribution access medium: accessibleThroughInterface*) and for users wishing to download it (*Distribution access medium: downloadable*).

Authorization to use treebanks is in many cases handled locally in the infrastructure, by asking identified users to agree to certain conditions for use.

## 4 INESS as a K-center

INESS and LINDAT were in 2015 approved as a joint virtual K-center in the CLARIN Knowledge Sharing Infrastructure.[12] This implies that knowledge will be shared with users in a systematic way, so as to assist researchers in managing and using resources efficiently and in line with good practice. To that effect, the INESS website contains a menu link to an overview page for getting started, while a FAQ intends to answer common questions for troubleshooting. There is also a link to extensive documentation about grammars, the query language, the web interface, annotation guidelines, and sharing treebanks. Furthermore, there are links to publications and to internal and external resources. Users can interact through a user forum, while a contact email address is also provided. The K-center also organizes treebanking tutorials. The first such event was organized by INESS in Warsaw on February 2 and 6, 2015.[13]

## 5 Users and use cases

INESS fills the gap between two groups targeted by CLARIN: those who have resources but need a place to deposit them, and those who wish to use resources and who need advanced online tools to explore them.

[8] `http://www.clarin.eu/content/license-categories`

[9] `http://services.geant.net/edugain`

[10] `http://discojuice.org`

[11] `http://hdl.handle.net/11495/D918-1214-6A7E-1`

[12] KSI, `http://www.clarin.eu/content/knowledge-centres`

[13] `http://pargram.b.uib.no/meetings/spring-2015-meeting-in-warsaw/`

Several projects and organizations, including also philology and historical linguistics initiatives such as MeNoTa,[14] ISWOC[15] and PROIEL,[16] have in INESS found both an archive to deposit their resources and a virtual laboratory for exploring resources.

INESS has also proved useful, for instance, in the context of the Parseme COST action, which aims, among other things, at investigating how multiword expressions are annotated in treebanks. Since searches for specific dependency relations can be performed in several treebanks simultanously, INESS-Search is a practical tool for making comparisons and checking consistency.

The treebank building facilities in INESS have been used by researchers at IPI-PAN (Warsaw) who have set up their own instance of the software and have developed a Polish LFG treebank (POLFIE).[17]

## 6 Concluding remarks and outlook

INESS is an example of a specialized, open infrastructure. It is encompassing in its range of online treebanking resources and services open to the CLARIN community.

INESS is also an example of an infrastructure which has been initiated outside of CLARIN, but which has gradually been incorporated in CLARINO and has to a large extent become compliant with good practice in CLARIN.

The main software is written in Common Lisp for expressiveness, extensibility and rapid development and updating. This package is available to others and has so far been installed at two sites: the Bergen CLARINO center and the IPI-PAN center in Warsaw, which is becoming an associated partner in INESS. We hope to see more of this cross-border sharing between countries in CLARIN.

Among possible future extensions, we are considering user-serviced uploading of treebanks, of corpora to be parsed, and of grammars. However, in our experience, there are often unpredictable divergences from standard formats which need to be manually solved.

Access to a treebank based on authorization granted by an external rightsholder remains an interesting challenge in the wider context of CLARIN. This is illustrated by the following cases in INESS. The use of LASSY-Klein, a treebank distributed by TST-Centrale,[18] is conditional upon the user signing a license agreement exclusively with TST-Centrale. Thus, end user licensing for this treebank cannot be handled by INESS. Since licensed users can download and search this treebank as they wish, they can request access to this resource through INESS, but only if they present a piece of paper signed by TST-Centrale — a procedure which is not practical. There is no automated way of verifying if potential users have obtained a license for this resource from the rightsholders. Similarly, our research group has a license for TüBa-D/Z,[19] for which a paper-based signed agreement is required as well, but we are not allowed to give users access to this treebank unless they are explicitly authorized by the University of Tübingen. Again, there is no easy way of verifying if potential users have signed a license for this resource.

The adoption of a common resource entitlement management system such as REMS[20] would make authorization a more streamlined process, not only for treebanks, but for any restricted resources which may be accessible through services at more than one CLARIN center. In such a scheme, any authorization by a rights owner (e.g. TST-Centrale) would be recorded in a secure database, which in turn could be consulted by service providers (such as INESS). The use of such a shared AAI architecture will be an important step to reach a European dimension. It will, however, only be effective if it is centrally promoted by the CLARIN ERIC and and widely adopted by CLARIN resource rightsholders and service providers alike.

[14] http://www.menota.org

[15] http://www.hf.uio.no/ilos/english/research/projects/iswoc

[16] http://www.hf.uio.no/ifikk/english/research/projects/proiel/

[17] This effort is reported in more detail elsewhere at the CLARIN2015 conference.

[18] http://tst-centrale.org/nl/producten/corpora/lassy-klein-corpus/6-66?cf_product_name=Lassy+Klein-corpus

[19] http://www.sfs.uni-tuebingen.de/en/ascl/resources/corpora/tueba-dz.html

[20] https://confluence.csc.fi/display/REMS/Home

## References

[Broeder et al.2010] Daan Broeder, Marc Kemps-Snijders, Dieter Van Uytvanck, Menzo Windhouwer, Peter Withers, Peter Wittenburg, and Claus Zinn. 2010. A data category registry- and component-based metadata framework. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

[Losnegaard et al.2013] Gyri Smørdal Losnegaard, Gunn Inger Lyse, Anje Müller Gjesdal, Koenraad De Smedt, Paul Meurer, and Victoria Rosén. 2013. Linking Northern European infrastructures for improving the accessibility and documentation of complex resources. In Koenraad De Smedt, Lars Borin, Krister Lindén, Bente Maegaard, Eiríkur Rögnvaldsson, and Kadri Vider, editors, *Proceedings of the workshop on Nordic language research infrastructure at NODALIDA 2013, May 22–24, 2013, Oslo, Norway. NEALT Proceedings Series 20*, number 89 in Linköping Electronic Conference Proceedings, pages 44–59. Linköping University Electronic Press.

[Martens2013] Scott Martens. 2013. TüNDRA: A web application for treebank search and visualization. In Sandra Kübler, Petya Osenova, and Martin Volk, editors, *Proceedings of the Twelfth Workshop on Treebanks and Linguistic Theories (TLT12)*, pages 133–144. Bulgarian Academy of Sciences.

[Meurer2012] Paul Meurer. 2012. INESS-Search: A search system for LFG (and other) treebanks. In Miriam Butt and Tracy Holloway King, editors, *Proceedings of the LFG '12 Conference*, LFG Online Proceedings, pages 404–421, Stanford, CA. CSLI Publications.

[Rosén et al.2012] Victoria Rosén, Koenraad De Smedt, Paul Meurer, and Helge Dyvik. 2012. An open infrastructure for advanced treebanking. In Jan Hajič, Koenraad De Smedt, Marko Tadić, and António Branco, editors, *META-RESEARCH Workshop on Advanced Treebanking at LREC2012*, pages 22–29, Istanbul, Turkey.

[van Noord et al.2013] Gertjan van Noord, Gosse Bouma, Frank Van Eynde, Daniël de Kok, Jelmer van der Linde, Ineke Schuurman, Erik Tjong Kim Sang, and Vincent Vandeghinste. 2013. Large scale syntactic annotation of written Dutch: Lassy. In Peter Spyns and Jan Odijk, editors, *Essential Speech and Language Technology for Dutch*, Theory and Applications of Natural Language Processing, pages 147–164. Springer, Berlin/Heidelberg.

[Vandeghinste and Augustinus2014] Vincent Vandeghinste and Liesbeth Augustinus. 2014. Making a large treebank searchable online. The SoNaR case. In Marc Kupietz, Hanno Biber, Harald Lüngen, Piotr Bański, Evelyn Breiteneder, Karlheinz Mörth, Andreas Witt, and Jani Taksha, editors, *Challenges in the Management of Large Corpora (CMLC-2)*, Reykjavik, Iceland.

# Influence of Interface Design on User Behaviour in the VLO

**Thomas Eckart**
NLP group
University of Leipzig
Leipzig, Germany
teckart@informatik.uni-leipzig.de

**Alexander Hellwig**
NLP group
University of Leipzig
Leipzig, Germany
alexander.hellwig@t-online.de

**Twan Goosen**
CLARIN ERIC,
The Language Archive
Nijmegen, The Netherlands
twan@clarin.eu

## Abstract

The metadata search engine *Virtual Language Observatory* (VLO) is one of the central facilities in CLARIN to search for and identify relevant linguistic resources by end users. Its purpose to provide a generic interface for all kinds of linguistic data and services resulted in user feedback from different communities with partially very specific requirements. As a consequence a major overhaul of the user interface was carried out in 2014. To identify improvements and impairments VLO log files from before and after this revision were analysed and compared. The results show changes in the user behaviour including changes in preferences of selected facets or typical query patterns.

## 1 Introduction

The Virtual Language Observatory (VLO)[1] provides a Web interface for accessing language resources on the basis of CMDI (see Broeder et al. (2012)) metadata files. Therefore it plays a central role for end users searching for specific resources provided in the context of CLARIN or similar projects. Naturally it is object of intensive discussions and various suggestions regarding its functionality or its user interface. These proposals are initiated by single end users as well as a CLARIN task force dedicated especially to improvements in user experience, extraction mechanisms, and documentation (see Haaf et al. (2014)). As a consequence the VLO is subject to regular changes in all its components.

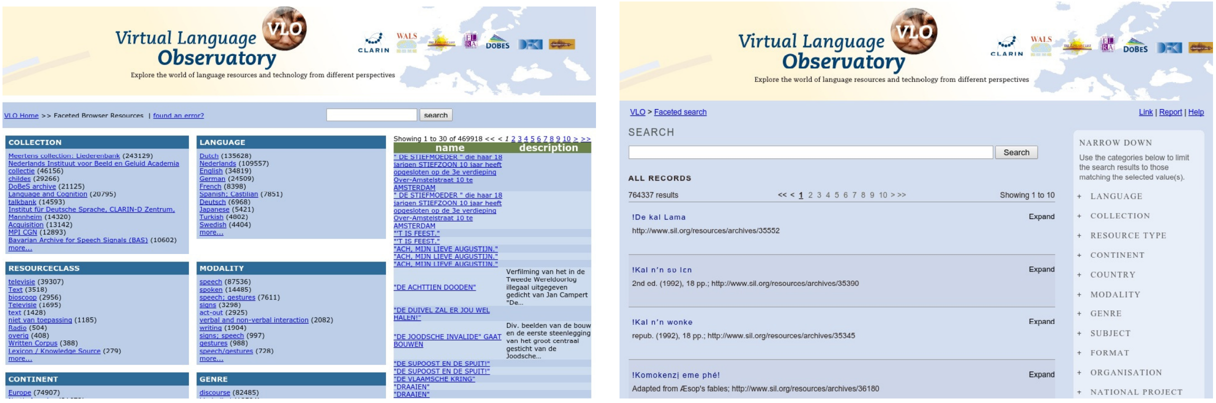
From the beginning one particular hotly contested topic was the design and functionality of the user interface. Hence the VLO was redesigned two times reflecting new wishes by the community and providing enhanced search functionality. Despite mostly positive feedback for the last major overhaul in 2014 (for details see Goosen and Eckart (2014)) it is hard to estimate the concrete consequences on the users' behaviour. For a more objective evaluation a deeper analysis is necessary. Based on anonymized log files such a web log analysis (Grace et al. (2011)) was conducted to answer the following questions regarding the revised interface:

- What is the influence on usage of full text search versus selection via facets?
- What is the influence regarding popularity of facets?
- What combination of facets are popular?
- Are there typical query patterns?

Additionally, this contrastive evaluation log files can also give insight in questions like:

- What concrete values are selected most often?
- Where do the requests come from?
- What is the typical distribution of requests over time?

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



(a) VLO version 2 (facets in the centre, resources on the right)

(b) VLO version 3 (default view on displayed resources, resources in the centre, facets on the right)

Figure 1: Changes in the VLO user interface

## 2 User Interface Redesign

The third version of the VLO was released in May 2014. Aside from some changes in the back end (i.e. database and metadata importer) the most prominent changes affected the user interface. Version 2 contained a full text search field on top but had a strong focus on the selection of resources via facets. Every facet was displayed including the ten most frequent values in a 2x7 grid layout in the centre. The list of selected results was only displayed in a small box on the right of the interface.

Version 3 reversed this layout by showing all facets on the right corner and an extended result list in the centre of the interface (similar to most faceted search systems). The result list supports an expanded view for every resource to display more attributes besides title and descriptive text. Furthermore, the start page of the new VLO only shows an elementary interface containing the full text search field, a preselection of links to useful facets and an option to show all resources with collapsed facets via two different links. Figure 1 compares both interfaces.

For a more detailed description of all changes, see Goosen and Eckart (2014).

## 3 Log File Data

The evaluation was based on log files for two timespans: 47 days (2013-12-02 – 2014-01-17) for the old interface (from now on *timespan 1*) and the first 164 days (2014-05-15 – 2014-10-25) after the new interface was deployed (*timespan 2*). Used log files included the access log of the VLO web server (with anonymized IP addresses) and the Apache Solr[2] log containing all requests on the database. As the current configuration of the VLO does not allow a direct mapping from user requests to database queries, a session detection was developed to map database queries to connected web server requests. As a large share of all requests were issued by automatic crawlers or bots an intensive cleaning of the log files was necessary. As a consequence the majority of all requests had to be removed. In the end 7409 requests for timespan 1 and 20,035 user requests for timespan 2 could be extracted.

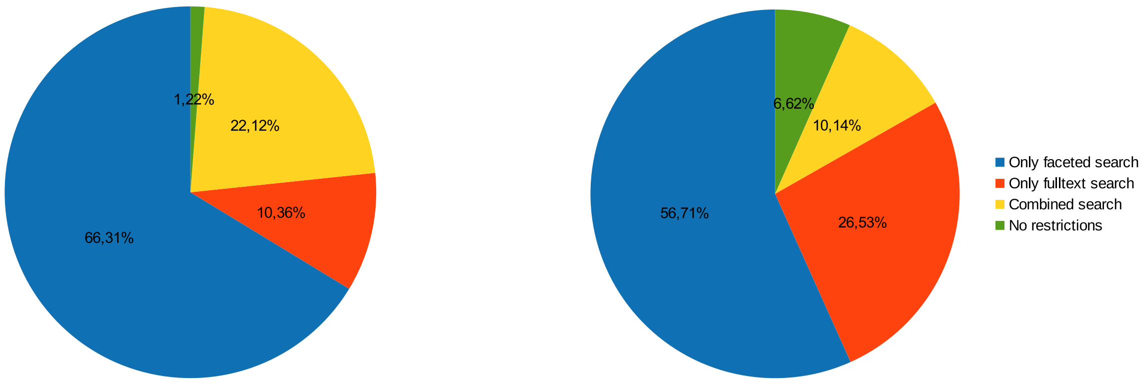
## 4 Influence on User Behaviour

Based on these extracted requests the behaviour and preferences of users were analysed. In the following three of these analyses are presented. Where applicable differences between the old and the new interface are shown.

### 4.1 Distribution of Query Types

The modifications to the user interface especially contained a stronger focus on full text search with less space for the facets and a preselection of facet values only after additional user interaction. Therefore

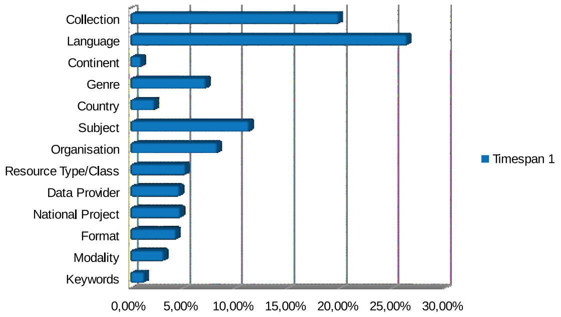
[2] http://lucene.apache.org/solr/



(a) Timespan 1

(b) Timespan 2

Figure 2: Query types in percent



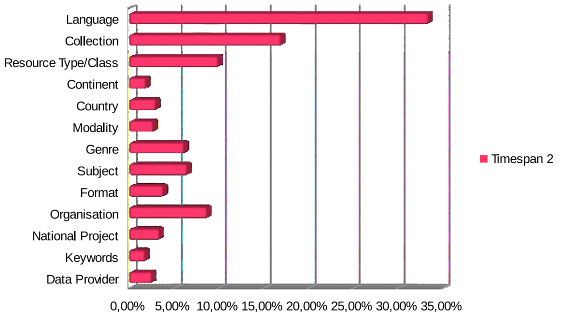


Figure 3: Used facets in timespan 1

Figure 4: Used facets in timespan 2

it could be expected that there would be changes in the usage of query types (divided in queries using only facet selection, queries using only full text, queries using both full text and facet selection, and queries with no selection, i.e. the start page). Figure 2 shows the differences between the old and the new interface. As expected, the number of queries using only full text search has increased. However, this had a negative impact on the number of faceted queries and combined queries.

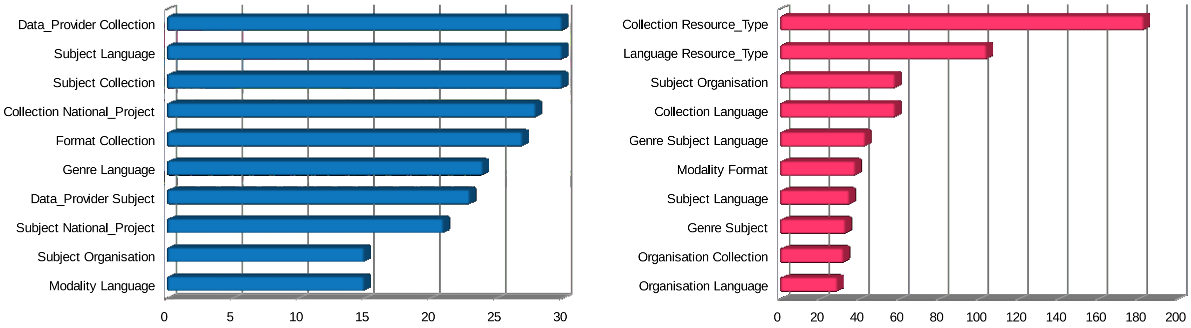
## 4.2 Usage of Facets

### 4.2.1 Absolute Facet Usage

A controversial topic regarding the VLO interface has always been selection and order of supported facets. Figure 3 and 4 show the usage of facets (in percent, based on all queries that use facet selection) for the old and the new interface. The facet order in the diagrams reflects the order in the respective UI. Especially for the old version there seems to be a correlation between the order of facets and their frequency of use, although it is unclear if this can be explained by an appropriate design or because users may only have used the first facets "above the fold" ignoring facets that can only be reached by explicit interaction (scrolling down). In comparison with the new interface it is obvious that the most popular facets are still popular for restricting result sets. These are especially *Language*, *Collection*, and *Organisation*. It is also noteworthy that unpopular facets like *Continent*, *Country*, or *Keywords* stay unpopular. This may be seen as an indication of missing interest by the users but also as a sign of inadequate quality, missing support by most of the resources or an unclear definition.

### 4.2.2 Popular Facet Combinations

The general idea of a faceted search interface is the process of "narrowing down" the result set to an acceptable amount of resources. Therefore, it is expected that users may use more than one facet to



(a) Timespan 1

(b) Timespan 2

Figure 5: Most frequent facet combinations (in absolute numbers)

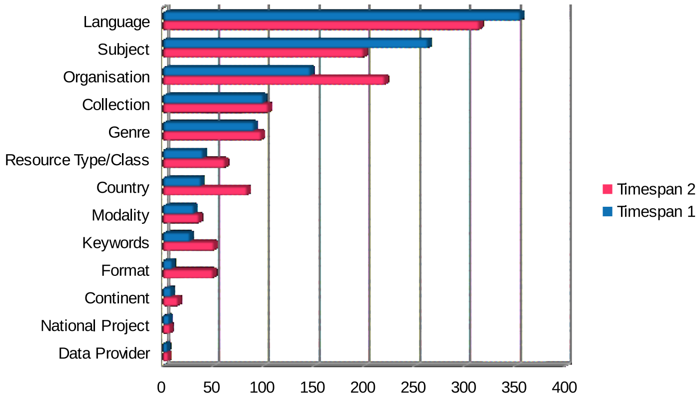


Figure 6: Number of selected facet values

get adequate results. The log files show that most queries using facet restriction only use one facet.[3]. Nevertheless, the combinations of facets may give insights in the appropriateness of facets and their exploitation by the users. Figure 5 shows the ten most frequent facet combinations for both timespans with their absolute frequencies. Some used facets seem to be over-represented compared with their absolute frequency of use (cp. 4.2.1). For example the facet *Subject* is used quite often in combination with others facets.

### 4.3 Selected Facet Values

Aside from the frequency of use for every facet it is also relevant to look at the selected values. Some facets provided by the VLO are based on rather small vocabularies (like *Continent* or *Country*) whereas open facets (like *Subject* or *Keywords*) have no such natural limitation. Figure 6 shows the absolute number of values selected from every facet for both timespans. Despite the fact that timespan 2 contains around three times the number of requests the absolute number of values does not vary as much as one could expect. For a facet like *Language* this may be due to a kind of “saturation” in the selection of relevant values where non-selected values may contain an above-average number of extraction errors. On the other hand, changes in postprocessing and metadata curation may have lead to reduced number of facet values over time. In contrast this does not hold for the facet *Country* with similar characteristics.

## 5 Conclusion

The analysis suggests the impact of interface design on the users’ behaviour and type of user requests. This especially holds for the stronger focus on full text search and a less dominant presentation of facets. In addition, the order of facets seems to guide user selections to some degree.

[3] This resembles results from evaluations in the context of Dublic Core based metadata (cf. Stvilia and Gasser (2008)).

In the future more significant results will be generated by an improved session logging to map associated requests to the same (anonymized) user. As a result more valuable hints may be extracted to improve functionality and interface of the VLO. Furthermore, this data may be useful for evaluating CMDI metadata schemata to enhance visibility and discoverability of resources based on typical user requests.

## References

[Broeder et al.2012] 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*, page 1.

[Goosen and Eckart2014] Twan Goosen and Thomas Eckart. 2014. Virtual Language Observatory 3.0: What's New? In *CLARIN Annual Conference 2014 in Soesterberg, The Netherlands*.

[Grace et al.2011] L.K.Joshila Grace, V. Maheswari, and Dhinaharan Nagamalai. 2011. Web log data analysis and mining. In Natarajan Meghanathan, BrajeshKumar Kaushik, and Dhinaharan Nagamalai, editors, *Advanced Computing*, volume 133 of *Communications in Computer and Information Science*, pages 459–469. Springer Berlin Heidelberg.

[Haaf et al.2014] Susanne Haaf, Peter Fankhauser, Thorsten Trippel, Kerstin Eckart, Thomas Eckart, Hanna Hedeland, Axel Herold, Jrg Knappen, Florian Schiel, Jens Stegmann, and Dieter Van Uytvanck. 2014. CLARINs virtual language observatory (VLO) under scrutiny - the VLO taskforce of the CLARIN-D centres. In *CLARIN annual conference 2014 in Soesterberg, The Netherlands*.

[Stvilia and Gasser2008] Besiki Stvilia and Les Gasser. 2008. Value-based metadata quality assessment. *Library & Information Science Research*, 30(1):67–74.

# Operationalisation of Research Questions of the Humanities within the CLARIN Infrastructure – An Ernst Jünger Use Case

**Dirk Goldhahn**
Natural Language
Processing Group
University of Leipzig
Germany
dgoldhahn@informatik.uni-leipzig.de

**Thomas Eckart**
Natural Language
Processing Group
University of Leipzig
Germany
teckart@informatik.uni-leipzig.de

**Thomas Gloning**
Department of
German Philology
University of Gießen
Germany
thomas.gloning@germanistik.uni-giessen.de

**Kevin Dreßler**
Natural Language
Processing Group
University of Leipzig
Germany
kvndrsslr@gmail.com

**Gerhard Heyer**
Natural Language
Processing Group
University of Leipzig
Germany
heyer@informatik.uni-leipzig.de

## Abstract

CLARIN offers access to digital language data for scholars in the humanities and social sciences. But how can the linguistic resources help to answer real research questions in the respective fields? By adressing research questions concerned with the work of German author Ernst Jünger an example of a successful operationalisation within the CLARIN infrastructure will be introduced. In addition a new versatile Web application for answering a wide range of research questions based on vocabulary use will be presented.

## 1 Introduction

CLARIN envisions itself as a research infrastructure for the humanities and social sciences. This makes the concerns of scholars of the respective fields a central aspect for CLARIN. A very important question which arises in this regard is: How can the linguistic resources offered by CLARIN be used to answer research questions in the humanities or social sciences?

In the following a use case of the humanities for the CLARIN infrastructure will be depicted. Research questions concerned with the work of German author Ernst Jünger will be adressed and operationalised. The CLARIN infrastructure will be used to search for necessary resources, to process the data and to analyse and visualise the results.

Part of this analysis is the Web application Corpus Diff for difference analysis currently developed in the context of CLARIN. It allows to answer a wide range of research questions which can be mapped to differences in vocabulary use. Typical workflows within this Web application will be presented utilizing the Jünger use case.

## 2 Research Questions

Ernst Jünger's political texts ("Politische Publizistik") from the years 1919 to 1933 are available in a philologically reviewed and well annotated edition (Berggötz, 2001), which has been digitalized for the purpose of internal investigation. The explosiveness of these texts lies in a wide range of topics regarding the development of Germany in the 1920s. This contains dealing with front experiences in World War I, consequences of the lost war, issues of national reorientation, and superordinated aspects of time interpretation. Jünger's texts change considerably in their thematic priorities and in their linguistic form in the 15 years of their creation.

Key issues that arise from a linguistic and discourse historical perspective on this corpus, include:

1. How does language use, in particular the use of words, correlate with specific topics and "perspectives", which are expressed in the texts?
2. How can the lexical profile of Jünger's political texts and its development be characterized in the temporal dimension?
3. How can the lexical profile of the political texts be characterized in comparison with contemporary material such as newspaper texts of the 1920s or the literary works of authors from the same period?

## 3 Operationalisation

In order to answer these research questions in a systematic matter, they need to be operationalised. Important aspects of this process are:

- data: collections of texts matching the research question (including reference corpora)
- algorithms: methods to carry out the desired analysis and their combination to more complex applications or workflows
- results and visualisation: structure, size, presentation and how to browse or search the data that lead to the results

Focus of the operationalisation will be on using the CLARIN infrastructure for searching for data and algorithms and performing the analysis by combining them to workflows.

First of all, texts matching the research question are required. As mentioned before, a digitized version of Ernst Jünger's political texts from 1919 to 1933 was already available. This corpus includes publishing dates for all included texts and will be the starting point for further analyses.

Next, a method needs to be chosen to discover differences in the use of vocabulary. One method that allows for such insights is difference analysis (Kilgarriff, 2001). Using this analysis we can investigate differences between different years of Jünger's work or between Ernst Jünger's texts and some reference corpora.

This will then allow to:

- quantify corpus similarity,
- discover differences in vocabulary and
- analyse prominent results (vocabulary) further.

### 3.1 Reference Data – DWDS

Another requirement for a difference analysis is the availability of reference data. A central entry point for scholars searching language resources in CLARIN is the Virtual Language Observatory (VLO). Using the VLO's search features such as facettes, it is easy to navigate and narrow down resources and identify those of interest for the respective research questions of the social sciences or humanities.

As our Jünger texts are in German and from the years 1919 to 1933, the same needs to hold for our reference corpora. When restricting the facets to "corpus" in "German" and adding the search term "20th century" one of the most prominent results is the DWDS Kernkorpus[1].

The DWDS corpus (Digitales Wörterbuch der deutschen Sprache) (Geyken, 2006) was constructed at the Berlin-Brandenburg Academy of Sciences between 2000 and 2003. The DWDS Kernkorpus contains approximately 100 million running words, balanced chronologically and by text genre.

[1] http://catalog.clarin.eu/vlo/search?q=20th+century&fq=languageCode:code:deu&fq=resourceClass:Corpus

The main purpose of the DWDS Kernkorpus is to serve as the empirical basis of a large monolingual dictionary of the 20th century. The Kernkorpus is roughly equally distributed over time and over genres: journalism, literary texts, scientific literature and other nonfiction.
Using the webservices of the DWDS we extracted texts for all genres. We collected corpora for each year and genre separately, allowing for analyses using both dimensions.

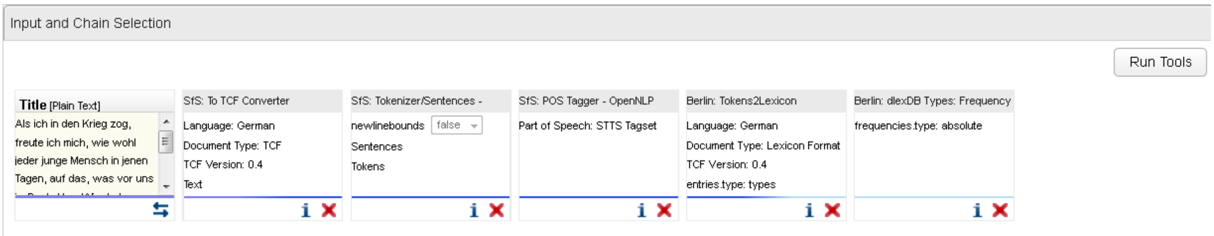
## 4 Combination to workflows

### 4.1 Preprocessing

Preprocessing of the raw material is the basis for conducting a difference analysis as word frequencies for the underlying texts need to be computed. Therefore, especially sentence segmentation and tokenization are relevant preliminary work. In addition, to allow for part of speech specific analyses, POS tagging needs to be performed.

For data processing WebLicht (Hinrichs, 2010), the Web-based linguistic processing and annotation tool, is an obvious choice. Within WebLicht one can easily create and execute tool chains for text processing without downloading or installing any software.

Figure 1 depicts a preprocessing chain created in WebLicht. It includes import of plain text files, conversion to an internal format (TCF), sentence segmentation, tokenization, part of speech tagging and counting of word frequencies. This processing can be run online and the results can then be downloaded. For an even more convenient data transfer an export of results to the personal workspace of CLARIN users will be available in the future.



**Title** [Plain Text]
Als ich in den Krieg zog, freute ich mich, wie wohl jeder junge Mensch in jenen Tagen, auf das, was vor uns

SfS: To TCF Converter
Language: German
Document Type: TCF
TCF Version: 0.4
Text

SfS: Tokenizer/Sentences -
newlinebounds false
Sentences
Tokens

SfS: POS Tagger - OpenNLP
Part of Speech: STTS Tagset

Berlin: Tokens2Lexicon
Language: German
Document Type: Lexicon Format
TCF Version: 0.4
entries.type: types

Berlin: dlexDB Types: Frequency
frequencies.type: absolute

Figure 1: Preprocessing chain in WebLicht.

The tool chain was executed on the Jünger texts for all annual time slices from 1919 to 1933 separately. This resulted in information on word frequencies for 15 corpora, one for each year. Furthermore we used the preprocessing for the annual corpora of each of the four DWDS-genres, resulting in 60 subcorpora, four for each year.

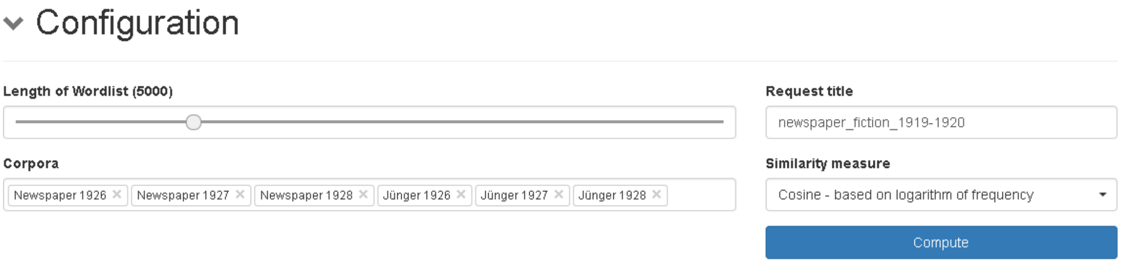
### 4.2 Data Analysis

The actual analysis can be carried out via the dedicated Web application Corpus Diff[2]. An easy to use JavaScript-based interface allows creating several analysis processes in parallel. The generation of corpus similarity is solely based on lists of word frequencies or word rank representing their respective corpus. The user interface provides several similarity measures that are using cosine similarity on these word vectors. Cosine similarity is just one possible option for the proposed comparison but has the advantage of flexibility concerning features and their weighting (e.g. logarithmic scaling). The result is a normalised value for every pairwise comparison between 0 (no resemblance of the word lists) and 1 (identically distributed vocabulary). The application is completely based on a RESTful webservice that delivers all information necessary: an overview of all provided corpus representations and the complete word lists for every corpus.
Using word frequency lists for corpus comparison has several advantages: these lists are dense representations of the content of a corpus, but due to their small size easy to transfer and to process. Furthermore there are hardly any copyright restrictions as no access to full texts or text snippets is necessary. This means that even for textual resources with very restrictive licences an exchange of this data is in most cases feasible.

By using the Web interface a user can select a set of corpora, the used similary measure and how many of the most frequent words of a word list should be taken into account for the analysis (figure 2). As a

[2] http://corpusdiff.informatik.uni-leipzig.de

result the user is provided with a matrix visualising pairwise similarity of corpora using different colour schemes. These colour schemes also emphasize clusters of similar corpora. In addition, a dendogram shows a representation of a single-linkage clustering for all word lists used in the analysis. Both, matrix and dendogram, are a means of identifying interesting corpora with an especially high or low similarity of their vocabulary. This can be used to perform a diachronic comparison to identify changes over time, but also for comparing corpora of different genre or origin with each other.



Configuration

Length of Wordlist (5000)

Request title: newspaper_fiction_1919-1920

Corpora: Newspaper 1926 ×, Newspaper 1927 ×, Newspaper 1928 ×, Jünger 1926 ×, Jünger 1927 ×, Jünger 1928 ×

Similarity measure: Cosine - based on logarithm of frequency

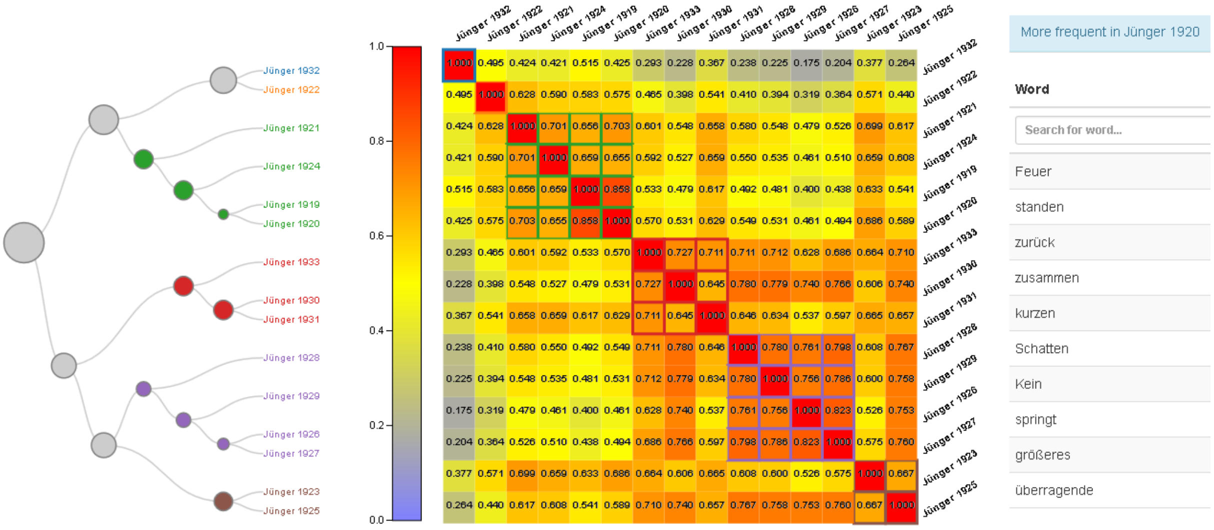
Figure 2: Configuration of corpus comparison tasks.

By selecting two corpora more detailed information about differences in their vocabularies is shown. This especially includes lists of words that are more frequent or that exclusively occur in one of the corpora. Both are valuable tools to identify terms that are specific or more significant for the respective resource. Moreover the results are a starting point for further analyses with hermeneutic approaches by experts of the respective fields.

If the user is interested in a specific word, a frequency timeline is provided via a line chart. This will usually be relevant for important key words of the texts in question or words extracted in the previous analysis steps. Here the diachronic development of the word's usage can be easily analysed and put into relation to other words or comparisons can be made of its frequency in different genres over time.

## 5 Examples

Figure 3 (left) shows the similarity matrix and the dendogram for the Jünger texts between 1919 and 1933. One interesting pair of corpora are, among others, those from 1920 and 1927 since a low similarity holds for them. When looking at salient vocabulary for this comparison (figure 3, right), words like "Feuer" ("fire") are much more prominent in the texts from 1920.



| | Jünger 1932 | Jünger 1922 | Jünger 1921 | Jünger 1924 | Jünger 1919 | Jünger 1920 | Jünger 1933 | Jünger 1930 | Jünger 1931 | Jünger 1928 | Jünger 1929 | Jünger 1926 | Jünger 1927 | Jünger 1923 | Jünger 1925 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Jünger 1932 | 1.000 | 0.495 | 0.424 | 0.421 | 0.515 | 0.425 | 0.293 | 0.228 | 0.367 | 0.238 | 0.225 | 0.175 | 0.204 | 0.377 | 0.264 |
| Jünger 1922 | 0.495 | 1.000 | 0.628 | 0.590 | 0.583 | 0.575 | 0.465 | 0.398 | 0.541 | 0.410 | 0.394 | 0.319 | 0.364 | 0.571 | 0.440 |
| Jünger 1921 | 0.424 | 0.628 | 1.000 | 0.701 | 0.656 | 0.703 | 0.601 | 0.548 | 0.658 | 0.580 | 0.548 | 0.479 | 0.526 | 0.699 | 0.617 |
| Jünger 1924 | 0.421 | 0.590 | 0.701 | 1.000 | 0.659 | 0.655 | 0.592 | 0.527 | 0.659 | 0.550 | 0.535 | 0.461 | 0.510 | 0.659 | 0.608 |
| Jünger 1919 | 0.515 | 0.583 | 0.656 | 0.659 | 1.000 | 0.858 | 0.533 | 0.479 | 0.617 | 0.492 | 0.481 | 0.400 | 0.438 | 0.633 | 0.541 |
| Jünger 1920 | 0.425 | 0.575 | 0.703 | 0.655 | 0.858 | 1.000 | 0.570 | 0.531 | 0.629 | 0.549 | 0.531 | 0.461 | 0.494 | 0.686 | 0.589 |
| Jünger 1933 | 0.293 | 0.465 | 0.601 | 0.592 | 0.533 | 0.570 | 1.000 | 0.727 | 0.711 | 0.711 | 0.712 | 0.628 | 0.686 | 0.664 | 0.710 |
| Jünger 1930 | 0.228 | 0.398 | 0.548 | 0.527 | 0.479 | 0.531 | 0.727 | 1.000 | 0.645 | 0.780 | 0.779 | 0.740 | 0.766 | 0.606 | 0.740 |
| Jünger 1931 | 0.367 | 0.541 | 0.658 | 0.659 | 0.617 | 0.629 | 0.711 | 0.645 | 1.000 | 0.646 | 0.634 | 0.537 | 0.597 | 0.665 | 0.657 |
| Jünger 1928 | 0.238 | 0.410 | 0.580 | 0.550 | 0.492 | 0.549 | 0.711 | 0.780 | 0.646 | 1.000 | 0.780 | 0.761 | 0.798 | 0.608 | 0.767 |
| Jünger 1929 | 0.225 | 0.394 | 0.548 | 0.535 | 0.481 | 0.531 | 0.712 | 0.779 | 0.634 | 0.780 | 1.000 | 0.756 | 0.786 | 0.600 | 0.758 |
| Jünger 1926 | 0.175 | 0.319 | 0.479 | 0.461 | 0.400 | 0.461 | 0.628 | 0.740 | 0.537 | 0.761 | 0.756 | 1.000 | 0.823 | 0.526 | 0.753 |
| Jünger 1927 | 0.204 | 0.364 | 0.526 | 0.510 | 0.438 | 0.494 | 0.686 | 0.766 | 0.597 | 0.798 | 0.786 | 0.823 | 1.000 | 0.575 | 0.760 |
| Jünger 1923 | 0.377 | 0.571 | 0.699 | 0.659 | 0.633 | 0.686 | 0.664 | 0.606 | 0.665 | 0.608 | 0.600 | 0.526 | 0.575 | 1.000 | 0.667 |
| Jünger 1925 | 0.264 | 0.440 | 0.617 | 0.608 | 0.541 | 0.589 | 0.710 | 0.740 | 0.657 | 0.767 | 0.758 | 0.753 | 0.760 | 0.667 | 1.000 |

More frequent in Jünger 1920

| Word |
|---|
| Feuer |
| standen |
| zurück |
| zusammen |
| kurzen |
| Schatten |
| Kein |
| springt |
| größeres |
| überragende |

Figure 3: Similarity matrix and dendogram for the Jünger texts 1919-1933 (left), List of words with higher relative frequency in 1920 when compared to 1927 (right).

The example of "Feuer" (in a military sense) shows the fruitfulness of the tool's visualizations. Both in respect of its usage from 1919 to 1933 in the edition of the "Politische Publizistik" and in comparison with newspaper texts from the same period, differences in word usage can be identified (figure 4), making it an ideal starting point for further analyses by experts of the respective fields. Since the focus of this paper is on the operationalization we will not follow up on these subject specific tasks.

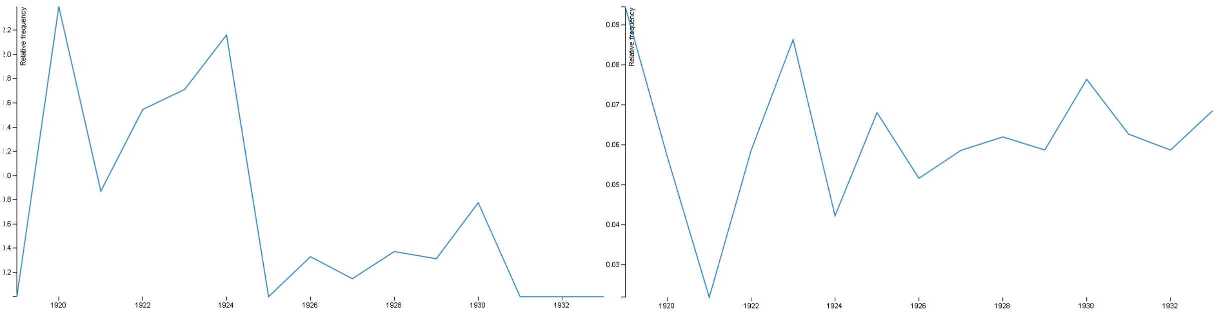


Figure 4: Frequency timeline for "Feuer" in texts of Ernst Jünger (left) and in newspaper texts (right).

A second example of the same kind are the dynamics of the word "Frontsoldat" which shows a comparable dynamical timeline. The word usage mirrors the status of the front soldier experience as a topic both in Jünger's texts and in the newspapers of the time.

## 6 Outlook

The current state is to be seen as a first working implementation to depict the workflow and its potential value for scholars from a broad range of scientific fields. For the near future it is planed to implement a seamless integration in the CLARIN infrastructure. This will especially include the support of TCF-based files and hence the usage of WebLicht as the first choice preprocessing chain.
It also planed to support the CLARIN personal workspaces as data storage to address copyright issues. As a consequence users will be enabled to store their data in a secure and private environment, and still make use of the full potential of a distributed infrastructure.

From the philological standpoint the following tasks are considered to be especially important for the further development:

- Methods to deal with polysemy and thematic specificity of terms (like "Feuer"/"fire")
- Answers to the question what fields of "traditional" research (e.g. Gloning to appear) can be treated by means of Digital Humanities and in what fields this is not feasible yet. Reference point for such a comparison or evaluation of methods could be by contrasting the performance of DH tools with results of traditional approaches. Such a comparison could provide indications for the future development of DH tools for specific research questions. In the context of this usecase this may be especially relevant for the field of word formation (e.g. *Bierreden*, *Bierstimmung*, or *Biertisch* used for degradation*)*.

In addition to the existing analysis and visualisation components it will be important to connect results with the original text material by providing an option to display and check the usage contexts. A Key Word in Context (KWIC) view would be one of these base functions where results generated by automatic analysis can be coordinated with the textual findings. As a typical user experience is that the corpus comparison tool provides surprising results that require further explanation, it would be desirable to provide easy to use methods for pursuing these traces in the original text. Besides a KWIC component (where available) an overview of typical contexts will be generated by using co-occurrences analysis (Büchler, 2006). This graph-based approach allows the visualisation of typical word contexts per time slice and especially following diachronic changes of these contexts.

## 7 Conclusion

In this paper we showed how a research question of the humanities and social sciences can be operationalised and answered using the CLARIN infrastructure. Different resources and services of CLARIN were utilized in this endeavour, such as the VLO or the WebLicht execution environment for automatic annotation of text corpora.

In addition we introduced a generic tool for difference analysis which can be used to answer a wide range of research questions based on vocabulary use. A workflow for utilizing this Web application and its different ways of visualisation to guide the exploration of the research results was presented. Finally, an outlook for a seamless integration into the CLARIN infrastructure was given.

## References

Berggötz, S.O. (2001). Ernst Jünger. Politische Publizistik 1919 bis 1933. Klett-Cotta, 2001.

Büchler, M. (2006). Flexibles Berechnen von Kookkurrenzen auf strukturierten und unstrukturierten Daten. Diploma Thesis, University of Leipzig.

Geyken, A. (2006). A reference corpus for the German language of the 20th century. In: Fellbaum, C. (ed.): Collocations and Idioms: Linguistic, lexicographic, and computational aspects. London: Continuum Press, 23-40.

Gloning, Th. (to appear). Ernst Jünger Publizistik der 1920er Jahre. Befunde zum Wortgebrauchsprofil. To appear in: Benedetti, Andrea/Hagestedt, Lutz (eds.): Totalität als Faszination. Systematisierung des Heterogenen im Werk Ernst Jüngers. Berlin/Boston: de Gruyter (to appear january 2016).

Hinrichs, M., Zastrow, T., & Hinrichs, E. W. (2010). WebLicht: Web-based LRT Services in a Distributed eScience Infrastructure. In Proceedings of LREC 2010, Malta.

Kilgarriff, A. (2001). Comparing corpora. International journal of corpus linguistics, 6(1), 97.

# DiaCollo: On the trail of diachronic collocations

**Bryan Jurish**
Berlin-Brandenburgische Akademie der Wissenschaften
Jägerstrasse 22-23 · 10117 Berlin · Germany
jurish@bbaw.de

## Abstract

This paper presents DiaCollo, a software tool developed in the context of CLARIN for the efficient extraction, comparison, and interactive visualization of collocations from a diachronic text corpus. Unlike other conventional collocation extractors, DiaCollo is suitable for extraction and analysis of diachronic collocation data, i.e. collocate pairs whose association strength depends on the date of their occurrence. By tracking changes in a word's typical collocates over time, DiaCollo can help to provide a clearer picture of diachronic changes in the word's usage, in particular those related to semantic shift. Beyond the domain of linguistics, DiaCollo profiles can be used to provide humanities researchers with an overview of the discourse topics commonly associated with a particular query term and their variation over time or corpus subset, while comparison or "diff" profiles highlight the most prominent differences between two independent target queries. In addition to traditional static tabular display formats, a web-service plugin also offers a number of intuitive interactive online visualizations for diachronic profile data for non-technical users.

## 1 Introduction

In recent years, an increasing number of large diachronic text corpora have become available for linguistic and humanities research, including the *Deutsches Textarchiv*[1] (Geyken et al., 2011) and the Corpus of Historical American English[2] (Davies, 2012). While the broad time spans represented by these corpora offer unique research opportunities, they also present numerous challenges for conventional natural language processing techniques, which in turn often rely on implicit assumptions of corpus homogeneity – in particular with respect to the temporal axis. Indeed, even putatively synchronic newspaper corpora have a nontrivial temporal extension and can reveal date-dependent phenomena if queried appropriately (Scharloth et al., 2013). This paper addresses the problem of automatic *collocation profiling* (Church and Hanks, 1990; Evert, 2005) in such diachronic corpora by introducing a new software tool "DiaCollo" explicitly designed for this purpose which allows users to choose the granularity of the diachronic axis on a per-query basis.

DiaCollo is a modular software package for the efficient extraction, comparison, and interactive visualization of collocations from a diachronic text corpus. Unlike conventional collocation extractors such as DWDS Wortprofil[3] (Didakowski and Geyken, 2013), Sketch Engine[4] (Kilgarriff and Tugwell, 2002; Rychlý, 2008), or the UCS toolkit[5], DiaCollo is suitable for extraction and analysis of diachronic collocation data, i.e. collocate pairs whose association strength depends on the date of their occurrence and/or other document-level properties. By tracking changes in a word's typical collocates over time and applying J. R. Firth's famous principle that "you shall know a word by the company it keeps" (Firth, 1957), DiaCollo can help to provide a clearer picture of diachronic changes in the word's usage.

[1] http://www.deutschestextarchiv.de

[2] http://corpus.byu.edu/coha

[3] http://zwei.dwds.de/wp

[4] http://www.sketchengine.co.uk

[5] http://www.collocations.de/software.html

DiaCollo was developed in the context of CLARIN in order to aid historians participating in the CLARIN Working Groups in their analysis of the changes in discourse topics associated with selected terms as manifested by changes in those terms' context distributions, and has been successfully applied to both mid-sized and large corpus archives, including the *Deutsches Textarchiv* (1600–1900, ca. 2.6K documents, 173M tokens) and a large heterogeneous newspaper corpus (1946–2015, ca. 10M documents, 4G tokens).

## 2 Implementation

DiaCollo is implemented as a Perl library, including efficient re-usable classes for dealing with native index structures such as $(string \leftrightarrow integer)$ mappings, $n$-tuple inventories and component-wise equivalence classes, or hierarchical tuple-pair frequency databases. DiaCollo indices are suitable for use in a high-load environment, since no persistent server process is required and all runtime access to native index data structures occurs either via direct file I/O or (optionally) via the `mmap()` system call for efficient kernel-managed page caching, relying on the underlying filesystem cache to optimize access speed.

In addition to the programmatic API provided by the Perl modules, DiaCollo provides both a command-line interface as well as a modular plugin for the DDC/D* corpus administration framework which includes a publicly accessible RESTful web service (Fielding, 2000) and a form-based user interface for evaluation of runtime database queries and interactive visualization of query results. The remainder of this paper describes the DiaCollo web service, whose architecture closely mirrors the independently documented command-line and Perl APIs. A publicly accessible web front-end for the *Deutsches Textarchiv* corpus can be found at `http://kaskade.dwds.de/dstar/dta/diacollo/`, and the source code is available via CPAN at `http://metacpan.org/release/DiaColloDB`.

**Requests & Parameters** DiaCollo is a request-oriented service: it accepts a user request as a set of *parameter=value* pairs and returns a corresponding *profile* for the term(s) queried. Parameters are passed to the service RESTfully via the URL query string or HTTP POST request as for a standard web form. Each request must contains at least a *query* parameter specifying the target term(s) to be profiled. The date-range to be profiled can be specified with the *date* parameter, while the *slice* parameter can be used to alter the granularity of the returned profile data by specifying the size in years of a single profile epoch. Candidate collocates can be filtered by means of the *groupby* parameter, and result-set pruning is controlled by the *score*, *kbest*, and *global* parameters.

**Profiles & Diffs** The results of a simple DiaCollo user request are returned as a tabular *profile* of the $k$-best collocates for the queried word(s) or phrase(s) in each of the requested date sub-intervals ("epochs" or "slices", e.g. decades) specified by the `date` and `slice` parameters. Alternatively, the user may request a comparison or "diff" profile in order to highlight the most prominent differences between two independent queries, e.g. between two different words or between occurrences of the same word in different date intervals, corpus subsets, or lexical environments.

**Indices, Attributes & Aggregation** For maximum efficiency, DiaCollo uses an internal "native" index structure over the input corpus content words to compute collocation profiles. Each indexed word is treated as an $n$-tuple of linguistically salient token- and/or document-attributes selected at compile-time, in addition to the document date. User `query` and `groupby` request parameters are interpreted as logical conjunctions of restrictions over these attributes, selecting the precise token tuple(s) to be profiled. For finer-grained selection of profiling targets, DiaCollo supports the full range of the DDC[6] query language (Sokirko, 2003; Jurish et al., 2014) via the `ddc` and `diff-ddc` profile types whenever the DiaCollo instance is associated with an underlying DDC server back-end.

**Scoring & Pruning** DiaCollo assigns each collocate $w_2$ in a unary profile for a target term $w_1$ a real-valued score by means of a user-specified *score function*. Supported score functions include absolute raw- and log-frequency (*f, lf*), normalized raw- and log-frequency per million tokens (*fm, lfm*), pointwise

[6] `http://www.ddc-concordance.org`

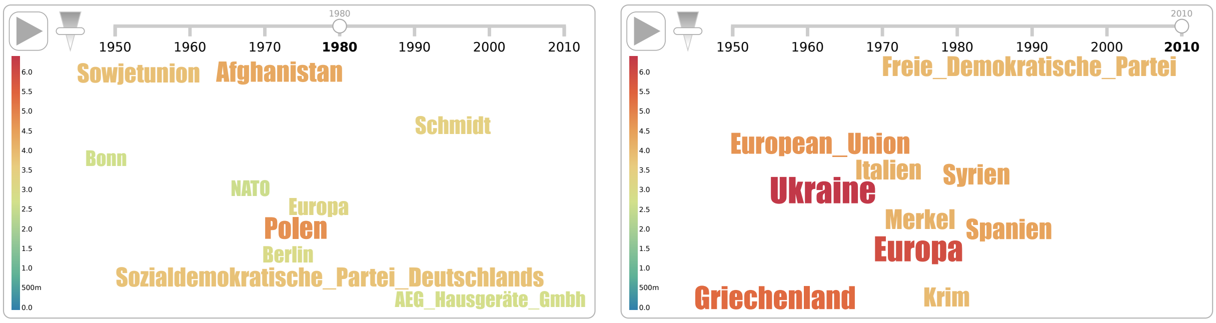


Figure 1: DiaCollo interactive tag-cloud visualization of the ten best proper name collocates of the noun *Krise* ("crisis") in the German daily newspaper *DIE ZEIT* for the epochs 1980–1989 (left) and 2010–2014 (right).

mutual information × log-frequency product (*mi*), and the scaled log-Dice coefficient (*ld*) as proposed by Rychlý (2008). Candidate collocates are ranked in descending order by the associated scores, and the $k$-best candidates in each epoch are selected and returned. "Diff" queries compute independent profiles $p_{\mathrm{a}}$ and $p_{\mathrm{b}}$ for the *query* and *bquery* parameters, respectively. After ranking according to the selected score function, a comparison-profile $p_{\mathrm{a-b}}$ is computed as $p_{\mathrm{a-b}} : w_2 \mapsto p_{\mathrm{a}}(w_2) - p_{\mathrm{b}}(w_2)$ for each of the up to $2k$ collocates $w_2 \in k\text{-best}(p_{\mathrm{a}}) \cup k\text{-best}(p_{\mathrm{b}})$ and the $k$-best of these with the greatest absolute differences $|p_{\mathrm{a-b}}(w_2)|$ are selected and returned.

**Output & Visualization** DiaCollo supports a number of different output formats for returned profile data, including TAB-separated plain text suitable spreadsheet import, native JSON for further automated processing, and a simple tabular HTML format. In addition to the static tabular formats, the DDC/D* web-service plugin also offers several interactive online visualizations for diachronic profile data, including two-dimensional time series plots using the Highcharts JavaScript library, flash-based motion charts using the Google Motion Chart library, and interactive tag-cloud and bubble-chart visualizations using the D3.js library. The HTML and interactive D3-based display formats provide an intuitive color-coded representation of the association score (rsp. score-difference for "diff" profiles) associated with each collocation pair, as well as hyperlinks to underlying corpus hits ("KWIC-links") for each data point displayed.

## 3 Example

Figure 1 contains example tag-cloud visualizations for a unary DiaCollo profile of proper name collocates for the noun *Krise* ("crisis") in 10-year epochs over an archive of the German daily newspaper *DIE ZEIT* spanning the interval 1950–2014. Since the term "crisis" usually refers to a short-lived and inherently unstable situation, its typical collocates can be expected to vary widely over time, reflecting changes in the discourse environment which in the case of a newspaper corpus can themselves be assumed to refer to events in the world at large. Indeed, the data in Figure 1 can easily be traced to prominent political events of the associated epochs. Conspicuous *Krise*-collocates and associated events in the 1980s include *Afghanistan* and *Sowjetunion* ("Soviet Union") for the Soviet-Afghan war (1979–1989); *Polen* ("Poland") due to the *Solidarność* movement and declaration of martial law in 1981; *Schmidt* and *Sozialdemokratische Partei Deutschlands* (SPD) referring to the collapse of the SPD-led German government coalition under Helmut Schmidt in 1982; and *NATO*, *Bonn*, and *Berlin* in the context of NATO's deployment of mid-range missiles in western Europe. The foreshortened final epoch (2010–2014) can be traced to civil wars in the Ukraine and Syria (*Syrien*), the Greek government-debt crisis and its repercussions in the European Union (*Griechenland, Italien, Spanien*), the Russian annexation of Crimea (*Krim*), and the German FDP (*Freie Demokratische Partei*) party's loss in the 2013 federal elections.

## 4 Summary & Outlook

This paper introduced DiaCollo, a new software tool for the efficient extraction, comparison, and interactive online visualization of collocations specially tailored to the unique demands of diachronic text corpora. In its top-level incarnation as a modular web service plugin for the DDC/D* corpus administration framework, DiaCollo provides a simple and intuitive interface for assisting linguists, lexicographers, and humanities researchers to acquire a clearer picture of diachronic variation in a word's usage over time or corpus subset. Future work will focus on implementing new visualization techniques for DiaCollo profile data, as well as extending the profiling back-end to handle other corpus input formats or online search APIs such as CLARIN Federated Content Search.

## References

[Church and Hanks1990] Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

[Davies2012] Mark Davies. 2012. Expanding horizons in historical linguistics with the 400-million word corpus of historical american english. *Corpora*, 7(2):121–157.

[Didakowski and Geyken2013] Jörg Didakowski and Alexander Geyken. 2013. From DWDS corpora to a German word profile – methodological problems and solutions. In Andrea Abel and Lothar Lemnitzer, editors, *Network Strategies, Access Structures and Automatic Extraction of Lexicographical Information*, (OPAL X/2012). IDS, Mannheim.

[Evert2005] Stefan Evert. 2005. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.D. thesis, Institut für maschinelle Sprachverarbeitung, Universität Stuttgart.

[Fielding2000] Roy Thomas Fielding. 2000. *Architectural Styles and the Design of Network-based Software Architectures*. Ph.D. thesis, University of California, Irvine.

[Firth1957] John Rupert Firth. 1957. *Papers in Linguistics 1934–1951*. Oxford University Press, London.

[Geyken et al.2011] Alexander Geyken, Susanne Haaf, Bryan Jurish, Matthias Schulz, Jakob Steinmann, Christian Thomas, and Frank Wiegand. 2011. Das deutsche Textarchiv: Vom historischen Korpus zum aktiven Archiv. In Silke Schomburg, Claus Leggewie, Henning Lobin, and Cornelius Puschmann, editors, *Digitale Wissenschaft. Stand und Entwicklung digital vernetzter Forschung in Deutschland*, pages 157–161.

[Jurish et al.2014] Bryan Jurish, Christian Thomas, and Frank Wiegand. 2014. Querying the deutsches Textarchiv. In Udo Kruschwitz, Frank Hopfgartner, and Cathal Gurrin, editors, *Proceedings of the Workshop "Beyond Single-Shot Text Queries: Bridging the Gap(s) between Research Communities" (MindTheGap 2014)*, pages 25–30, Berlin, Germany, 4th March.

[Kilgarriff and Tugwell2002] Adam Kilgarriff and David Tugwell. 2002. Sketching words. In Marie-Hélène Corréard, editor, *Lexicography and Natural Language Processing: A Festschrift in Honour of B. T. S. Atkins*, EURALEX, pages 125–137.

[Richling2011] Julia Richling. 2011. Referenzkorpus Altdeutsch (Old German reference corpus): Searching in deeply annotated historical corpora. Talk presented at the conference *New Methods in Historical Corpora*, 29–30 April, 2011. Manchester, UK.

[Rychlý2008] Pavel Rychlý. 2008. A lexicographer-friendly association score. In *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN*, pages 6–9.

[Scharloth et al.2013] Joachim Scharloth, David Eugster, and Noah Bubenhofer. 2013. Das Wuchern der Rhizome. linguistische Diskursanalyse und Data-driven Turn. In Dietrich Busse and Wolfgang Teubert, editors, *Linguistische Diskursanalyse. Neue Perspektiven*, pages 345–380. VS Verlag, Wiesbaden.

[Sokirko2003] Alexey Sokirko. 2003. A technical overview of DWDS/Dialing Concordance. Talk delivered at the meeting *Computational linguistics and intellectual technologies*, Protvino, Russia.

# The Regulatory and Contractual Framework as an Integral Part of the CLARIN Infrastructure: an Estonian and Finnish Perspectives

**Aleksei Kelli**
University of Tartu
School of Law
Estonia
aleksei.kelli@ut.ee

**Kadri Vider**
University of Tartu
Center of Estonian Language Resources
Estonia
kadri.vider@ut.ee

**Krister Lindén**
University of Helsinki
Department of Modern Languages
Finland
krister.linden@helsinki.fi

## Abstract

The article focuses on the regulatory and contractual framework in CLARIN. The discussion is based on the process analysis approach, which allows an evaluation of the functionality and shortcomings of the entire legal framework concerning language resources and technologies. The article reflects the personal knowledge and insights of the authors gained through their work with legal aspects of language resources and technologies in Estonia and Finland. The analysis may be helpful to CLARIN partners facing similar problems.

**Keywords**: regulatory and contractual framework, CLARIN agreement templates, contractual and exception model.

## 1 Introduction[1]

The nature of language resources (LR) and language technologies (LT) can be analyzed from several perspectives such as technological, linguistic, ethical and legal. The authors focus on the legal challenges relating to the development and dissemination of language resources and technologies. From this point of view, the regulatory and contractual framework (legal framework) constitutes one of core infrastructures of CLARIN.

The discussion is based on the process analysis approach, which allows the evaluation of the functionality and shortcomings of the entire legal framework concerning LR and LT. The process starts with the development and ends with the dissemination of language resources and technologies. Different process phases are not addressed separately since they affect each other. Legal issues are defined and analyzed within each phase.

The authors use traditional social science methods and draw on previous legal research conducted by the authors on LR and LT. The analysis relies on the Estonian and Finnish experience. The article also reflects the personal knowledge and insights of the authors gained through work with the legal aspects of LR and LT in Estonia and Finland. The authors make suggestions for improving the existing legal framework. The analysis could also be helpful to other CLARIN partners facing similar problems.

The paper is organized into two main sections. The first section focuses on the establishment of institutional control over the developed language resources and technologies. The second addresses the issue of the development of language resources and deals with the dissemination and potential subsequent utilization of LR and LT. We also study the case of providing public access to fragments of resources in a concordance service versus distributing resources in full for research purposes in light of a research exception in the copyright regulation and the CLARIN contractual framework.

[1]

## 2 Establishment of the institutional control over language resources and technologies

The dissemination and utilization of language resources and technologies depends on several conditions. Institutions distributing resources must have sufficient technological capabilities. Additionally, they must be legally entitled to do so. In practical terms, this means that institutions managing resources must have the capacity to enter into valid transactions and have concluded all relevant contracts regarding LR and LT. In order to avoid too abstract an analysis, the authors use Estonia as an example in addressing these issues.

Estonia has set up the Center of Estonian Language Resources (CELR) as a consortium of 3 institutions at the national level on December 2, 2011. The consortium consists of the University of Tartu (UT) (as leading partner in CELR), the Institute of Cybernetics at Tallinn University of Technology, and the Institute of the Estonian Language. The consortium constitutes an organizational framework for the coordination and implementation of the obligations of Estonia as a member in CLARIN ERIC.

The national consortium is expected to perform obligations, which may bind the whole consortium. The problem, however, is that the national consortium is not a legal person in private or public law (legal entity). The consortium is an agreement. Technically speaking the consortium partners could represent each other but this could create legal uncertainty. In the Estonian consortium, each partner works with certain types of resources. Therefore, the partners have agreed that each one concludes agreements for his respective field of activity.

The Estonian consortium agreement regulates issues relating to the partners' background and foreground IP. However, it does not provide a clear framework concerning the LR and LT developed and owned by persons outside the consortium. To acquire these language resources and technologies, the consortium partners have to conclude agreements with these persons. Since the aim of CLARIN ERIC is to standardize and unify its members' activities, CLARIN has developed standard agreement templates (Licenses, Agreements, Legal Terms). CLARIN also has standard deposition agreement templates. CLARIN deposition license agreements are divided into three categories:

1) CLARIN-DELA-PUB-v1.0 (for public use resources);
2) CLARIN-DELA-ACA-v1.0 (for academic use resources);
3) CLARIN-DELA-RES-v1.0 (for restricted use resources).

In principle, the authors support the ideology of CLARIN having three categories of resources and integrating the approach into a contractual framework (deposition license [DELAs] and end-user license agreements [EULAs]). However, when we analyze specific agreements, we can identify ways of improving them. The process of translating the CLARIN contractual framework (DELAs and Terms of Services) into Estonian and making them compatible with the Estonian law provided a good opportunity to scrutinize once again the existing contracts. A very preliminary meeting was held in Helsinki on 21 May 2015 to discuss the possible amendments to the CLARIN agreement templates and develop them further.[2] The results and observations are discussed below.

The first observation concerns the structure of the DELAs. All DELAs have almost identical provisions. The main difference comes from the provisions concerning intellectual property rights and access rights (Section 7).[3] Therefore, it would be practical to divide the DELA into two parts: a general part for all persons depositing resources and a separate part for selecting a specific category (PUB, ACA, RES). It should be easily achievable in an e-environment.

According to the second observation, the provisions on the warranties and indemnity are among the most important clauses of the DELAs.[4] In the current version of the DELAs, Section 10 regulates liability and indemnity. The provision should be revised to make the regulation clearer. In the following table the current and amended Section 10 is presented:

| The current provisions | The amended provisions |
|---|---|
| **10. Legal Obligations** | **10. Warranties and indemnity** |
| 10.1 The Copyright holder shall be responsible for | 10.1 The Depositor warrants and represents that (i) |

[2] The meeting was held in Helsinki on 21 May 2015. Participants: Krister Linden, Kadri Vider and Aleksei Kelli.

[3] There are some differences in annexes too but it should be easy to unify the approach.

[4] The other important part is the license which owners of resources and technologies grant to repositories.

| | |
|---|---|
| holding a copyright or a sufficient license and/or other rights based on intellectual property law to the Resource and that any use of the Resource for the purposes compliant with this Agreement does not in any form violate any third party copyright or any other rights based on intellectual property law or other incorporeal right. | it possesses all proprietary rights, title and interest in the Resource and has full authority to enter into this Agreement. The Depositor shall be responsible for holding copyright, related rights and other rights or a sufficient license and/or other rights to the Resource and that any use of the Resource for the purposes compliant with this Agreement does not in any form violate any third party copyright, related rights or any other rights. |
| 10.2 The Copyright holder is held liable for all damages and costs he causes CLARIN or the Trusted Centres in the CLARIN Service by breaching any of the obligations in 10.1. | 10.2 The Depositor undertakes to indemnify and hold harmless the Repository of any liability, directly or indirectly, resulting from the use and distribution of the Resources, including but not limited to claims from third parties. The Depositor is held liable for all damages and costs he causes CLARIN or the Trusted Centres in the CLARIN Service by breaching any of the obligations in 10.1. |
| 10.3 Should a third party present a justified claim that the Resource violates the obligations in 10.1., the Resource can be removed from the CLARIN Service. | 10.3 Should a third party present a claim that the Resource violates the obligations in 10.1., the Resource can be removed from the CLARIN Service. |

In the current version, Section 10 is called “Legal Obligations”. This is not the best choice of words since all obligations arising from a contract are legal. Therefore Section 10 should be called “Warranties and indemnity”. The amended version also reflects a new terminological approach. The DELA terms identifying the parties to the agreement are replaced as follows: the Copyright curator (CLARIN Centre receiving LR and LT) is replaced with “repository” and the Copyright holder (person licensing LR and LT) with “depositor”. Subsection 10.1 and 10.2 are elaborated further to increase clarity. Subsection 10.3 was amended to provide sufficient grounds for removal of resources if a third party presents a claim that her rights are violated. The repository does not have to prove that the claim was justified. In addition, Subsection 10.3 must be compatible with the CLARIN Notice and Take Down Policy.

An additional issue regarding the deposition of resources concerns the question whether institutions should accept LR and LT on an as-is basis without any representations and warranties. In case a depositor does not have title to the resources, it would be out of the question. If resources are developed based on the copyright exception and/or include personal data, DELAs for the categories ACA or RES are suitable.

## 3 Development and dissemination of language resources

Language resources have two tiers of rights: 1) the rights of the persons who developed the resources and 2) the rights of the persons whose copyright-protected content (sometimes also content with related rights) was used when creating the resources. In the previous section, we addressed the first tier and here we focus on the second tier.

From a legal perspective, language resources constitute copyright protected databases (see Kelli et al. 2012; Tavast et al. 2013). The creation of language resources often requires the use of copyright protected works. The use, however, can be based on two models: 1) the contract model and 2) the exception model. The contract model means that a person developing language resources acquires permission to use copyrighted works (books, journal articles, etc.). The exception model is based on a copyright exception allowing free use of works for research purposes. Both models have their pros and cons.

The contract model allows negotiating terms for commercial use of resources and making them publicly available. The model is expensive even if copyright holders do not ask for remuneration. Administrational costs arise during the negotiations and the management of contracts. There is no guarantee that the right-holders use identical contracts. This could lead to incompatibility between different contracts and restrict the development and dissemination of resources. Another problem is *de* *facto* orphan works (anonymous web posts, blogs etc.) since there is no one identifiable who can give permission for their use.

The advantage of the exception model is that there is no need to ask for permission from the right-holders. It is possible both to use works of identified authors and works of unidentifiable authors (*de facto* orphan works). There is no administrative burden to negotiate licenses. The main disadvantage is that it is not possible to use the developed resources for commercial purposes or make them available in the PUB category. Dissemination is possible only in the categories ACA and RES.

The Estonian Copyright Act has a general research exception allowing development of language resources (Autoriõiguse seadus § 19). The draft Copyright and Related Rights Act introduces a specific exception for data mining and text analysis worded as follows: "reproduction and processing of an object of rights for the purpose of text analysis and data mining, on the condition of attributing the name of the author of the used work, the name of the work and the source of publication, except if such attribution is impossible, and on the condition that such use is not carried out for commercial purposes". It was added for the sake of legal clarity.

Finland relies on the contract model. FIN-CLARIN has refrained from paying for resources but has contributed a minimal sum towards the collective extended license for the Finnish billion word newspaper corpus which has been scanned and OCRed by the National Library of Finland comprising newspapers from 1792 to the present. FIN-CLARIN provides access to the full corpus for non-commercial research purposes and access to anyone for small excerpts based on search results.

Similarly the billion word blog Suomi24 maintained and distributed by the commercial company AllerMedia is available in full for non-commercial research purposes via FIN-CLARIN but excerpts can be used by anyone. The motivation for this by AllerMedia is that it welcomes ideas provided by the research community by facilitating access and hopes to provide access to the same data to commercial companies against a fee.

Language resources and technologies are made available within CLARIN through a specific contractual framework. Firstly, a person interested in using LR and/or LT has to accept the Terms of Services (TOS) (Licenses, Agreements, Legal Terms). The DELAs and TOS are two sides of the same coin. When DELA shifts all liability regarding language resources and technologies to the depositors, the TOS disclaims and limits CLARIN's liability regarding resources to the maximum extent allowed by law. Drawing on public licenses such as EUPL, GPL and Creative Commons, we suggest amending Section 5 of TOS so that it is absolutely clear that the resources are provided on an as-is and as-available basis and no liability is assumed.

In addition to the TOS, the prospective user also has to accept the EULA attached to the language resources and technologies.

When it comes to the dissemination of language resources, it is useful to remember the maxim of Roman law saying "*Nemo plus iuris ad alium transferre potest, quam ipse haberet*" (Dig. 50.17.54). This means you cannot give others more rights to something than you have yourself (see Zimmermann, 1996). In other words, resources developed based on a research exception cannot be licensed in the PUB category. In view of this, we study how to provide public access to fragments of resources *versus* distributing resources in full for research purposes using the CLARIN contractual framework. A set of excerpts (*i.e.* search results) may be considered derived works, which are subject to the same conditions as the original work unless otherwise agreed. We may therefore still need a DELA to acquire the right to distribute the search results publicly.

In most cases, the right-holders are willing to make excerpts publicly available while the full corpus is only distributed for academic or restricted purposes. In case there is no research exception, this can still be agreed using one DELA (as the PUB/ACA/RES templates are quite similar). In both cases, the resource needs two metadata records: one for pointing to the PUB excerpts and one for pointing to the original ACA/RES resource, *i.e.* we have data with two different uses provided by two licenses in one agreement.

## 4 Conclusion

The regulatory and contractual framework constitutes an integral component of the CLARIN infrastructure. Similar to technical standards CLARIN also needs unified legal standards. It is in the interest of CLARIN to lobby for an EU-wide mandatory text and data mining exception.

CLARIN agreement templates have to be integrated and evaluated as one functional system. There are two ways for language resources and technologies to enter CLARIN. Firstly, the employees' rights regarding LR and LT are transferred to the employer (i.e. CLARIN national consortium member). Secondly, authors who are not employed by a CLARIN national consortium member have to conclude a deposition agreement. The aim of the DELA is to shift the liability for the resources to its depositors. The wording of the relevant provisions needs to be amended accordingly so that it is very explicit that a depositor is responsible for the resource.

Users can access CLARIN resources after they have agreed to the Terms of Services. The TOS objective is *inter alia* to limit CLARIN's liability towards users. Resources and technologies are offered on an as-is and as-available basis. The wording of the provisions regulating liability in TOS should be amended to limit CLARIN liability to the maximum extent.

## Reference

[Autoriõiguse seadus § 19] Autoriõiguse seadus [Copyright Act] (as entering into force on 12.12.1992). RT I 1992, 49, 615; RT I, 29.10.2014, 2 (in Estonian). Unofficial translation available via https://www.riigiteataja.ee/en/eli/531102014005/consolide (accessed on 12 July 2015);

[Kelli et al. 2012] Aleksei Kelli, Arvi Tavast, Heiki Pisuke (2012). Copyright and Constitutional Aspects of Digital Language Resources: The Estonian Approach. – Juridica International (19), 40-48;

[Dig. 50.17.54]. Available at http://www.thelatinlibrary.com/justinian/digest50.shtml (13.7.2015);

[Licenses, Agreements, Legal Terms]. Available at http://clarin.eu/content/licenses-agreements-legal-terms (13.7.2015);

[Tavast et al. 2013] Arvi Tavast, Heiki Pisuke, Aleksei Kelli (2013). Õiguslikud väljakutsed ja võimalikud lahendused keeleressursside arendamisel (Legal challenges and possible solutions in developing language resources). – Eesti Rakenduslingvistika Ühingu Aastaraamat (9), 317-332;

[The draft Copyright and Related Rights Act] Autoriõiguse ja autoriõigusega kaasnevate õiguste seaduse eelnõu. Versioon: 21.7.2014 [The Estonian draft Copyright and Related Rights Act. Version: 19.7.2014]. (in Estonian), https://ajaveeb.just.ee/intellektuaalneomand/wp-content/uploads/2014/08/Aut%C3%95S-EN-19-7-2014.pdf, (accessed on 5 May 2015);

[Zimmermann, 1996] Reinhard Zimmermann. The Law of Obligations Roman Foundations of the Civilian Tradition. – Oxford University Press, 1996.

# Variability of the Facet Values in the VLO – a Case for Metadata Curation

**Margaret King**
ACDH-OEAW
Vienna, Austria
margaret.king
@oeaw.ac.at

**Davor Ostojic**
ACDH-OEAW
Vienna, Austria
davor.ostojic
@oeaw.ac.at

**Matej Ďurčo**
ACDH-OEAW
Vienna, Austria
matej.durco
@oeaw.ac.at

## Abstract

In this paper we propose a strategy for metadata curation especially with respect to the variability of the values encountered in the metadata records and hence in the facets of the main CLARIN metadata catalogue, the VLO. The approach concentrates on measures on the side of the infrastructure and the interaction between human curators and the automatic processes.

## 1 Introduction

CLARIN runs a mature well-established metadata infrastructure, harvesting metadata from more than 60 providers on a weekly basis using the standardized OAI-PMH protocol. Some seven hundred thousand records are collected and provided via the main metadata catalog, the Virtual Language Observatory or VLO (Van Uytvanck et al, 2010). It aims to provide access to a broad range of linguistic resources from many disciplines and countries based on the flexible metadata framework CMDI (Broeder et al., 2010, 2012). After a few years of intensive use by the community and continuous growth of the body of data made available via this service a number of issues have been identified (Broeder et al., 2014) concerning the functionality of the catalog, but mainly the quality of the metadata provided by the data providers such as the variation in metadata values. These irregularities seriously hamper the discoverability of resources.

After reviewing the work done within the CLARIN community until now, this paper concentrates on the issue of variant values within the facets in the VLO, exemplifying primarily by the Resource Type facet, and proposes a strategy for the implementation of a metadata curation workflow that could rectify (some of) the described problems.

## 2 State of Research

The CLARIN community is acutely aware of the issue at hand, and has discussed the question of how to curate metadata and especially normalize the VLO's facet values on multiple occasions. A Metadata Curation Taskforce was established in 2013 by the Centre's Committee (SCCTC) with delegates from member countries, however this taskforce until now could only collect ideas, describe the situation and tried to remedy some of the encountered problems. It wasn't able to sustain a concerted level of activity to systematically approach this problem.

CLARIN-D established a separate VLO Taskforce in October 2013 (Haaf et al., 2014) which worked out recommendations for the VLO facets in an attempt to provide more guidance and clarity regarding the usage and meaning of the facets to the data providers. The VLO Taskforce meetings throughout 2014 and 2015 provided small steps towards a solution. However the Taskforce has concentrated on recommendations and sound definitions, the actual implementation is not seen as one of its tasks.[1] A sound definition of the facets and recommended values for the facets is certainly a necessary condition and a good starting point towards answering the problem under consideration. However it is of little use if it is not integrated in the infrastructure nor taken up by resource providers.

In 2014, Odijk conducted an in depth survey of the VLO from the point of view of discoverability of linguistic resources (Odijk, 2014). The comprehensive report identified a number of concrete issues and

[1] as indicated in informal talks with members of the taskforce

proposed possible solutions. These identified problems pertain both to the schema level (e.g. crucial elements not obligatory), to the instance level of the data (fields not filled, variation of the values), and also to the functionality provided by the VLO (missing facets, multi-selection). He also underscored the aspect of granularity, a related point currently much discussed throughout CLARIN but one which falls outside the scope of this paper.

In an unpublished follow-up internal CLARIN report in 2015, Odijk lays out a strategy for metadata curation, concentrating on the main goal to achieve clean facets. Based on the assumption that "the providers in general case cannot improve their metadata" the main actor in the curation process is the curation task force operating on the harvested metadata (Odijk, 2015). The main reason why the metadata cannot be improved on the side of the data providers is the lack of resources to invest in improving legacy data. CMDI in its complexity may pose a steep challenge to data provider with limited resources, it seems not trivial for data providers to select the right CMDI profile without guidance. Finally, in provider's own realm the metadata may be perfectly consistent and homogeneous, it is just through aggregation that inconsistencies arise.

## 3 VLO Metadata: a closer look

Thus the mission of the CLARIN metadata curation task force in (in normalizing the variant facets) is twofold. In the first place it must analyze the different problems of variation and its effect on discoverability. The second practical aim is that of creating and implementing a strategy for curation within the framework of CLARIN's social structures.

### 3.1 Variation of Values

We can identify different types of variation. From trivial ones like case or whitespaces ("WrittenCorpus" vs. "Written Corpus"), to combination of multiple values in one field with arbitrary (or even no) delimiters (e.g. "AddressesAnthologiesLinguistic corporaCorpus"), synonyms ("spoken" vs. "audio", "text" vs "written") and, most problematically, complex (confusing) values that carry too much information and need to be decomposed to multiple values possibly in multiple facets.

Odijk points to the data provider isolation as a main cause for the variation of values (Odijk, 2014). Indeed, it is clear that different people describe things in different ways. Some providers assigned the value "text" to Tacitus' Annals while someone else chose to create a new value called "Annals". This assumption is also supported by the fact that once the data is restricted to a single collection or organization the values in facets mostly "clear up" and appear as a consistent set.

The obvious answer from the infrastructure point of view is to reach better coordination between the data providers, basically applying shared controlled vocabularies (Durco and Moerth, 2014). Presently the only guidance regarding recommended vocabularies for individual facets was provided in the Recommendations by the VLO-Taskforce. Even these vocabularies are rarely used. In the Resource Type facet only 15,000 records use one of the 25 recommended values. All in all round 250 different values are used in the Resource Type facet, the most common reason for variation is the inclusion of extra information (unrelated to Resource Type but to some other facet). For example Shakespeare's King Lear is described by the Resource Type "poem" which belongs in the Genre facet with the Resource Type "text". A controlled vocabulary could help data providers to assign the details to the correct facet.

### 3.2 Missing values

Even worse than the variation of the values is the fact, that many records do not provide any value for some of the facets. Odijk attributes this mainly to the lack of obligatory metadata elements in CMDI and the fact that the metadata authors are often 'blind' to the 'obvious' aspects of their resources, like language or type. For the special case of the Resource Type the main reason may be the fact that the type is implicit in the underlying CMDI-Profile (e.g. TextCorpusProfile, LexicalResourceProfile).

Whatever the reasons, the extent of the problem is alarming. Most of the facets cover only about ⅓ of the records, so from the some 700 thousand records around 500 thousand are not visible and findable in each facet (except for the automatic/obligatory ones: Collection, Data Provider). Table 1 lists the number of null values for each facet.

A minimal remedy to deal with facets without specified values, would be to collect all records without appropriate value facets should have default value (e.g. “unspecified” or “unknown”)[2]. More advanced solution would be to evaluate values for certain facets from other facets or metadata fields, like “continent” from “country.” We aim for complete coverage, i.e. every record needs to be represented at once.

| Facet | null count | Facet | null count |
|---|---:|---|---:|
| Language Code | 240 183 | Subject | 503 233 |
| Collection | 0 | Format | 62 381 |
| Resource Type | 482 935 | Organisation | 520 560 |
| Continent | 472 048 | Availability | 580 907 |
| Country | 474 637 | National Project | 104 316 |
| Modality | 490 195 | Keywords | 567 347 |
| Genre | 329 114 | Data Provider | 0 |

Table 1 Number of records not covered within given facet in the VLO (on a sample of 631 000 records)

### 3.3 Missing facets

One source of the problem with confusing values may be the lack of appropriate facets. When trying to normalize the values of the Resource Type facet it was sometimes unclear in dealing with an overloaded value exactly where the information should go. For example, mediums of information such as radio, internet, mobile phone as well as more technical entries did not have a clear value among the recommendations for this facet. This lack of facets was also identified by Odijk (2014), who suggests adding a dedicated facet for Linguistic Annotation, as well as by the VLO task force, proposing new facets for Lifecycle Status, Rights Holder and License. However adding more facets also raises the complexity of the user interface, and the mapping, so the impact of such additions would need to be carefully examined.

### 3.4 Need for an Efficient Curation Workflow

As mentioned in the State of Research section a great deal of inquiry has been spent on establishing exactly what types of problems exist in the area of facet value normalization most notably in Odijk (2014). While some of the trivial problems in value variation can be solved programmatically (case folding, whitespace normalization), all the more complex issues like synonyms and complex values require human input - a mapping of variant values to recommended ones. There exist some first, tentative mappings available as a result of the analysis done by Odijk or the team of authors. Besides the question of the reliability of such mappings, the next challenge is how to integrate such a mapping into the established harvesting and ingestion workflow, especially how to ensure a sustainable and consistent process over time.

At the moment any automatic curation steps happen during the ingestion of the metadata into the indexer (the so-called “post-processing”). However this is currently limited to simple programmatic corrections of values, a mapping between actual and normalized values is only applied for the “Organization” facet. What is especially missing is a procedure to ensure that the mappings are kept up to date (new previously unseen values are added and mapped) and the curation process has access to the most current version of the mappings.

We will concentrate in the following section on the general framework that needs to be in place to ensure collaborative maintenance of vocabularies and mappings, their integration in the automatic curation process, and their wider adoption by the data providers. It is crucial to ensure that all changes are transparent to the data provider and to the user of the VLO. Another requirement is to make the workflow more modular, especially allow for the curation module to be encapsulated enough so as to be reusable in other contexts.

## 4 Proposed solution for normalization of facet values

The first step in coming to a solution will be to ensure that the recommended values are sufficient for the current state of the VLO. Once there is a (relatively) stable version of these recommendations, a manual, case by case mapping must be completed for the individual facets. Very importantly, these

[2] As we actually did on our test instance of VLO when evaluating data for this work.

mappings must be constantly modified (new values/mappings added) as new metadata is added to the VLO) In practice, the work on the recommendations and the mappings will go hand in hand. Also given the sheer size of the value lists of the individual facets, we need to set up a priority list, and process the facets one by one. Only the combination of human and automatic curations can lead to an efficient and workable solution. Only a human can unpack the kinds of variations that exist but only an automatic procedure can implement the corrections consistently

This strategy was applied by the team of authors to the Resource Type facet in a testing instance of VLO, and has proven workable and lead to a substantial consolidation of the facet values. The wider adoption of the process, especially inclusion of colleagues from other national consortia still needs to be implemented.

### 4.1 Integration into Workflow

There are multiple strategies where changes to the processed data can be introduced in the workflow:

a) The most radical approach is to define a separate profile guided solely by the goal of better discoverability, with elements reflecting one to one the facets from VLO. While this would make the life of the VLO developers very easy, it would move the burden of mapping the existing profiles to this one profile either to the data providers or to the curation task force.

b) The other extreme is to normalize values only at the moment of indexing the records, which is the approach currently already adopted in some facets within the VLO (“post-processing”).

c) Next option is to create amended copies of the metadata records by the curation module while staying within the confines of the original schema/profile.

d) A variant of the previous option is to keep the original records and only indicate proposed changes by means of annotations.

### 4.2 Management of Vocabularies and Mapping

A relatively simple (and partly already implemented) approach to the management of the mappings is to maintain a vocabulary in the vocabulary repository CLAVAS, where, based on the SKOS data model, every entity or concept is maintained as a separate record/item (skos:Concept), with a skos:prefLabel as the normalized name/label for given concepts and all variants encountered in the actual metadata stored as skos:altLabel (or skos:hiddenLabel). This information can be easily retrieved from CLAVAS and injected in the harvesting/curation workflow of the VLO. This is currently being done for the Organization names. The change introduced in CMDI 1.2 (Goosen et al., 2014) allowing to indicate a controlled vocabulary for given element in the CMDI-profile should in the mid-term also help with the handling of vocabularies with relation to the metadata elements.

What is still missing is an automatic procedure to add new previously unseen values to CLAVAS. The application underlying CLAVAS, OpenSKOS exposes a rich RESTful API that allows us not only to query but also to manipulate the data. So technically it would be possible for the curation module to add new candidate concepts. Human interaction is crucial here. These candidate concepts need to be clearly marked and kept in “quarantine” until they are checked and approved by the curators.

However, even if this whole process is set up it does not offer a solution to the more complex problem, when a value in one facet needs to be decomposed to multiple values in multiple facets. The ad-hoc experiments until now showed that a natural data structure would be a simple table with the encountered values in first column, a separate column for the other facets, allowing the curators to decompose the facet value intuitively/ergonomically into the appropriate facets.

If this file is stored as text/csv file and maintained under version control in the CLARIN’s code repository, it can be easily edited by a team of curators, seeing who has done what when and can equally easily be retrieved and processed by any application, most notably the curation module.

A final technical issue is the testing phase. In order to prove that metadata quality and VLO discoverability are improved by curation module, test cases have to be designed by experts. Each class of identified problems should be covered and generated reports should be used by metadata curators and software developers for further improvements.

### 4.3 Prevention – fighting the problem at the source

While we pessimistically stated at the beginning that we cannot expect the providers to change their metadata, we cannot give up on them, as it is clearly better to combat the problem at the source. There are indeed a number of measures that can (and need to) be undertaken on the side of the data provider:

a) best practices guides and recommendations (like the CLARIN-D VLO-Taskforce recommendations on the VLO facets), especially a list of recommended profiles (one or two per resource type) needs to be provided, with profiles that have good coverage of the facets and use controlled vocabularies wherever possible

b) provision of detailed curation reports to the providers as separate output of the curation step

c) provision of curated/amended metadata records directly back to the data providers (automatically available with option c) and d))

d) availability of the controlled vocabularies via a simple API (as is provided by the OpenSKOS-API) to be integrated with metadata authoring tools. This functionality has been already planned to be added for at least two metadata editors used by the CLARIN community: Arbil (Withers, 2012) and COMEDI (Lyse et al., 2014)

A crucial ingredient to the proposed strategy is the question of governance, i.e. who is going to steer the process and if not force than still persistently keep reminding data providers of the problems encountered and proposing solutions. CLARIN has well-defined organizational structures and a number of bodies with delegates from all member countries, where decisions can be agreed upon on different levels. In the described case, the primary operative unit is definitely the metadata curation task force with representatives from national consortia, in a tight collaboration with the CMDI task force, both reporting to the SCCTC, which in turn reports to the Board of Directors. Thus both the horizontal coverage over the member countries is ensured, so that national metadata task forces can report shortcomings they have identified, as well as the vertical integration of the decision-making bodies, allowing to integrate small practical technical solutions as well as to propose substantial structural changes, if needed.

## 5 Conclusion

In this paper we proposed a strategy for curation and normalization of values in the facets of the VLO. We elaborated on the ways how to establish and sustain a workflow that combines systematic, automatic, transparent curation of the metadata with continuous input from human curators providing the mappings from actual values encountered in the metadata to recommended “normalized” values. Integral part of the process must be a suite of test cases that ensure the quality of the mappings and the whole curation process. Finally, all output of the curation (corrections & amended metadata records) must be recycled to the data providers in the hope of preventing further problems and the entire work cycle must repeat as new resources are added. Thus the need for metadata curation is perpetual.

## References

[Broeder et al. 2014] Broeder, D., Schuurman, I., & Windhouwer, M.. 2014.Experiences with the ISOcat Data Category Registry. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014), Reykjavik*.

[Broeder et al. 2012] Broeder, D., Windhouwer, M., Van Uytvanck, D., Goosen, T., & Trippel, T. 2012. CMDI: a component metadata infrastructure. In *Describing LRs with metadata: towards flexibility and interoperability in the documentation of LR workshop programme* (p. 1).

[Broeder et al. 2010] Broeder, D., Kemps-Snijders, M., Van Uytvanck, D., Windhouwer, M., Withers, P., Wittenburg, P., & Zinn, C. 2010. A data category registry-and component-based metadata framework. In *Seventh conference on International Language Resources and Evaluation [LREC 2010]* (pp. 43-47). European Language Resources Association (ELRA).

[Durco and Moerth, 2014] Ďurčo, M., & Moerth, K. 2014. Towards a DH Knowledge Hub - Step 1: Vocabularies. Poster at *CLARIN Annual Conference*, Soesterberg, Netherlands.

[Goosen et al., 2014] Goosen, T., Windhouwer, M.A., Ohren, O., Herold, A., Eckart, T., Durco, M., Schonefeld, O. 2014. CMDI 1.2: Improvements in the CLARIN Component Metadata Infrastructure. At the *CLARIN Annual Conference*. Soesterberg, The Netherlands, October 23 - 25.

[Haaf et al. 2014] Haaf, S., Fankhauser, P., Trippel, T., Eckart, K., Eckart, T., Hedeland, H., ... & Van Uytvanck, D.. 2014. CLARIN's Virtual Language Observatory (VLO) under scrutiny-The VLO taskforce of the CLARIN-D centres. *CLARIN*. http://ids-pub.bsz-bw.de/frontdoor/deliver/index/docId/3210/file/Haaf_Fankhauser_CLA-RINs_virtual_language_observatory_under_scrutiny_2014.pdf

[Lyse et al., 2014] Lyse, G. I., Meurer, P., & De Smedt, K. (n.d.). COMEDI: A New COmponent Metadata EDItor. Presented at the CLARIN Annual Conference 2014, Soesterberg, Netherlands. Retrieved from http://www.clarin.eu/sites/default/files/cac2014_submission_13_0.pdf

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

[Odijk, 2015] Jan Odijk. 2015. Metadata curation strategy. Internal document, unpublished.

[Van Uytvanck, 2010] Van Uytvanck, D., Zinn, C., Broeder, D., Wittenburg, P., & Gardelleni, M. 2010. Virtual language observatory: The portal to the language resources and technology universe. In *Seventh conference on International Language Resources and Evaluation [LREC 2010]* (pp. 900-903). European Language Resources Association (ELRA).

[Withers, 2012] Withers, P. 2012. Metadata management with Arbil. In *Proceedings of the Workshop Describing LRs with Metadata: Towards Flexibility and Interoperability in the Documentation of LR at LREC* (pp. 72–75).

# LAP: The CLARINO Language Analysis Portal

**Emanuele Lapponi, Stephan Oepen, Arne Skjærholt, and Erik Velldal**
University of Oslo
Department of Informatics
{emanuel|oe|arnskj|erikve}@ifi.uio.no

## 1 Introduction: High-Level Goals

This abstract describes the current state of the Language Analysis Portal (LAP) currently under development in the Norwegian CLARINO initiative. LAP provides users with a collection of state-of-the-art tools for natural language processing that are accessible via a unified, in-browser user interface. Built on top of the open-source Galaxy framework (Giardine et al., 2005; Blankenberg et al., 2010; Goecks et al., 2010), the system offers means to combine tools into workflows, keep track of their output, and deliver results to users in different formats.

Unlike related on-line processing environments such as Weblicht (Hinrichs et al., 2010), LAPPS (Ide et al., 2014) and Alveo (Estival and Cassidy, 2014), which predominantly instantiate a distributed architecture of web services, LAP achieves scalability to potentially very large data volumes through integration with the Norwegian national e-Infrastructure, and in particular job submission to a capacity compute cluster. This setup leads to tighter integration requirements and also calls for efficient, low-overhead communication of (intermediate) processing results with workflows. We meet these demands by coupling the data model of the Linguistic Annotation Framework (LAF) (Ide and Romary, 2001; Ide and Suderman, 2013) with a lean, non-redundant JSON-based interchange format and integration through an agile and performant NoSQL database—allowing parallel access from cluster nodes—as the central repository of linguistic annotation. While the utility of natural language processing tools is apparent for linguists (computational or not), the ultimate goal of LAP is to reduce technological barriers to entry for researchers from the social sciences and humanities (SSH).

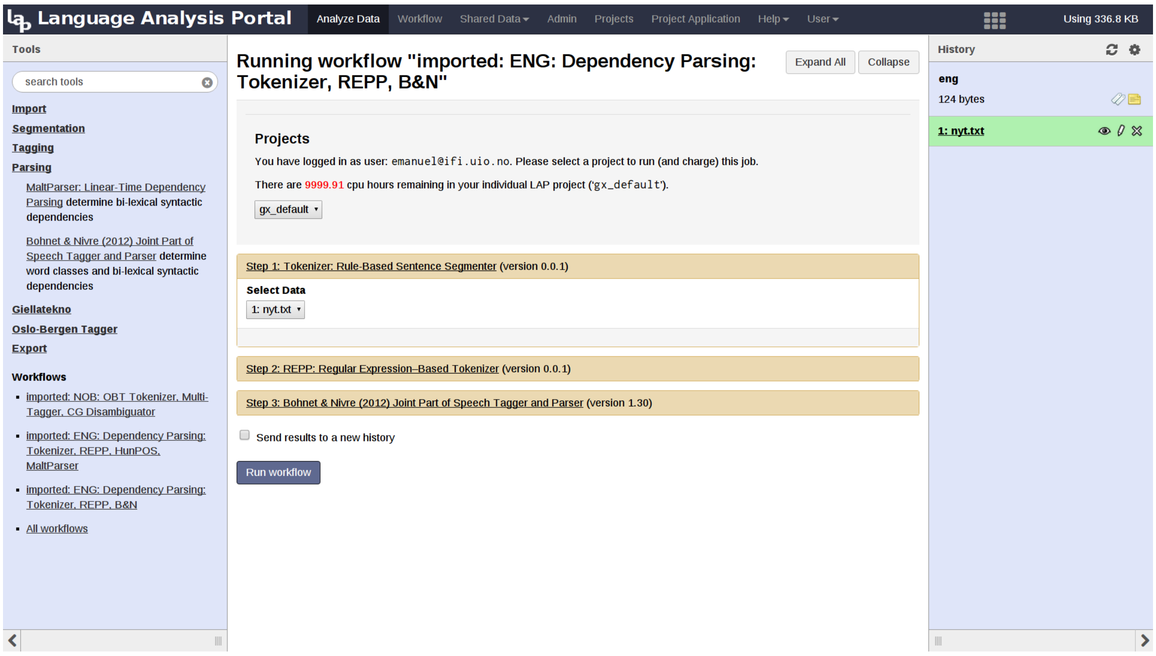
## 2 Design and Implementation: Galaxy and HPC

As described in Lapponi et al. (2013), LAP is built on top of Galaxy, a widely adopted framework for genome processing in the field of bioinformatics. Galaxy has proven to also be a suitable platform for natural language processing portals, and it has recently also been adopted by e.g. LAPPS[1] and Alveo.[2] Galaxy is an application that runs inside the browser, offering a graphical user interface to configure and combine analysis tools, upload, inspect and download data and share results and experiments with other users. A central part of the interface is a *workflow manager*, enabling the user to specify and execute a series of computations. For example, starting with a PDF document uploaded by the user, she might further want to perform content extraction, sentence segmentation, tokenization, POS tagging, parsing, and finally identification of subjective expressions with positive polarity—all carried out in a consecutive sequence. The output of each component provides the input to the next connected component(s) in the workflow, creating a potentially complex pipeline.

Rather than creating *ad-hoc* processing tools for LAP, existing NLP software is adapted and made available from within Galaxy. Adapting a tool to exist within the LAP ecosystem means making sure that it is compatible with the other installed tools: For instance, a dependency parser that requires sentence

[1] http://galaxy.lappsgrid.org/

[2] http://alveo.edu.au/help/analysing-data/transferring-data-to-galaxy-for-processing/



MaltParser: Linear-Time Dependency Parsing determine bi-lexical syntactic dependencies

Bohnet & Nivre (2012) Joint Part of Speech Tagger and Parser determine word classes and bi-lexical syntactic dependencies

- imported: NOB: OBT Tokenizer, Multi-Tagger, CG Disambiguator
- imported: ENG: Dependency Parsing: Tokenizer, REPP, HunPOS, MaltParser
- imported: ENG: Dependency Parsing: Tokenizer, REPP, B&N
- All workflows

Running workflow "imported: ENG: Dependency Parsing: Tokenizer, REPP, B&N"

Projects

You have logged in as user: emanuel@ifi.uio.no. Please select a project to run (and charge) this job.

There are 9999.91 cpu hours remaining in your individual LAP project ('gx_default').

gx_default

Step 1: Tokenizer: Rule-Based Sentence Segmenter (version 0.0.1)

Select Data: 1: nyt.txt

Step 2: REPP: Regular Expression–Based Tokenizer (version 0.0.1)

Step 3: Bohnet & Nivre (2012) Joint Part of Speech Tagger and Parser (version 1.30)

□ Send results to a new history

Figure 1: Screenshot of the LAP user interface: Preparing to launch a pre-defined workflow.

segmentation, tokenization, and part-of-speech annotations should be able to build its input from the output of other LAP annotators.

As different tools often use different and mutually incompatible representation formats for encoding input and output data, an important step in achieving interoperability is the implementation of an interchange format that can serve as a *lingua franca* among the components in a workflow. In LAP this interchange format is based on the LAF data model for representing annotations, as further described below.

## 3 Tool Interchange Format: LAF in LAP

LAF is a graph-based model for representing multi-modal linguistic annotations that aims at providing full interoperability among annotation formats. One of the fundamental principles that guided the development of LAF is that all annotation information should be explicitly represented, i.e., the interpretation of annotations should not require implicit knowledge about particular categories and relations (Ide and Suderman, 2013). Another fundamental principle is that one should observe a strict separation between annotation *structure* and annotation *content*. The focus of LAF is only on the structural part, and it is important to realize that LAF itself is not a format as such. It does not come with pre-specified linguistic labels or categories etc. Rather, it is a general framework for how to represent the annotation structure itself; an abstract data model specifying how to relate annotations to data and how to relate annotations to other annotations.

In LAP, annotations produced by tools are tied to both the original text and each other by means of the main components in a LAF graph: regions, nodes and edges. Regions describe the so-called base segmentation of a text in terms of character offsets, and might be associated to nodes containing annotation about the segment. For instance a tokenizer, which accepts free text as input, can produce both a target region and a paired node with an annotation containing the normalized token. A part-of-speech tagger produces both nodes containing the actual POS annotation and edges linking its nodes to the input tokens. Region, node and edge records are added to a MongoDB instance after each tool execution.

When the user runs another tool, only the subgraph containing the necessary annotations is invoked. For instance, if the user wants to run a new POS tagger on text that has already been tagged and parsed, LAP invokes only the part of the graph describing the relevant tokens. This strategy differs from other approaches adopted in other web interfaces to NLP tools. In Weblicht, for instance, annotations in the

TCF format (Heid et al., 2010) have to be parsed and re-encoded at each processing step. We believe that our approach is better suited to the modularity of LAP and provides a more malleable set-up for efficiently analyzing larger data-sets and running out-branching jobs on the cluster.

However, it is important to clarify that our implementation of the LAF data model, due to fundamental differences in goals and setup, does not stand in competition with TCF or richer formats for corpus annotations such as FoLiA (van Gompel and Reynaert, 2013) or TEI as an end-user format. The focus of our LAF implementation is its use as an internal interchange format. We have no expectations that future LAP users will find much use in downloading LAP annotations directly from the database. In keeping with the modular philosophy of LAP, we aim to provide users with LAP tools for importing from and exporting to other formats (imagine, for instance, a workflow starting with a TCF importer and ending in a TCF exporter, enabling compatibility with Weblicht workflows, or exporting to the FoLiA format for further manual annotations with the FLAT[3] annotation tool). A detailed description of the LAF data model as implemented in LAP is provided by Lapponi et al. (2014).

## 4 Tool Integration and Versioning: The LAP Tree

LAP integrates processing tools from many different sources, which are implemented in different programming languages. We have designed and populated a repository of tools, i.e. ready-to-run and pre-configured installations of processing components used by LAP, with replicability and relocatability as key design desiderata; this repository is dubbed the *LAP Tree*. Central installation of static versions of LAP tools on compute nodes is completely avoided. Instead, the LAP Tree is realized as a version-controlled repository, where individual components and all dependencies that transcend basic operating system functionality can (in principle) be checked out by an arbitrary user and into an arbitrary location, for immediate use (on all cluster nodes) through Galaxy. By default, Galaxy users are presented with the latest (stable) available version of tools, but the LAP Tree makes it possible for users to request individual versions of components directly in Galaxy, which ‘behind the scenes’ is implemented by means of a user-specific instance of (relevant parts of) the LAP Tree that is dynamically populated.

All code at the LAP interface layer is versioned jointly with the LAP Tree, such that we anticipate enabling users to back-date workflows to (in principle) arbitrary points in time of LAP evolution, thus taking an important step to reproducibility of experiments and replicability of results. Galaxy provides built-in functionality for sharing data with other users, including processing results, as well as complete workflows. In this context, a specific, per-user instantiation of a processing workflow can encompass all parameter choices (including tool versions) for all tools involved, i.e. the full ‘recipe’ that lead from a set of input data to a set of results. Once frozen as a workflow and shared with other users, the combination of versioning in the LAP Tree and standard Galaxy functionality, thus, promises to provide a framework that enables reproducibility (and, ultimately, also adaptation) by other researchers.

## 5 Reaching Out: Analyzing the European Parliament Debates

An important part of the motivation for CLARIN(O) as an infrastructure initiative is, of course, to facilitate the use of language technology among SSH researchers. This is also an important motivation for LAP. In attempting to facilitate and enable LT-based research in other fields, the importance of maintaining a bottom-up view on the process of how research questions are created and addressed in practice should not be underplayed (Zundert, 2012). Our position is that starting out from actual research questions, in direct collaboration with SSH researchers themselves, provides an ideal point of departure for surveying user requirements. In line with this we have focused on establishing contact with SSH researchers that might be interested in collaborating on using language technology in their own work. One such outreach effort has led to collaboration with researchers within political science interested in data-driven analysis of the legislative processes within the European Union. A joint ongoing project investigates whether an SVM classifier can be trained to predict the party affiliations of Members of the European Parliament on the basis of their speeches in the plenary debates, with a particular focus on the contribution of linguistically informed features. Preliminary results are presented by Høyland et al.

[3] `https://github.com/proycon/flat/`

(2014). The stages in the prediction pipeline here involve several layers of linguistic annotations, in addition to feature extraction, model estimation, and testing. An important future step is to fully integrate this entire pipeline within LAP itself, seeking to find the right balance between supporting all the requirements of this particular analysis task while also maintaining enough generality in the implementation to also make the involved components re-usable and applicable to other tasks. The work described by Høyland et al. (2014) ties in closely with the focus of the CLARIN-initiaited project *Talk of Europe: Travelling CLARIN Campus*[4] (ToE), focusing on processing and representing the European Parliament debates in a way that enables further analysis and collaboration by SSH researchers).

## 6 Talk of Europe: LAP Annotations in a Large RDF Store

The Talk of Europe project aims at curating the proceedings of the European Parliament Debates to linked data, so that it can be linked to and reused by other datasets and services. Our LAP-specific objective in this context is to contribute to the creation of a high- quality, multi-lingual corpus of European Parliament Proceedings, coupled with state-of-the-art syntactico-semantic analyses at the token and sentence levels—all encoded in the form of labeled directed graphs in the Resource Description Framework (RDF). LAP developers participated in the first ToE Creative Camp, where our contribution has targeted the textual (transcribed) content of European Parliament speeches. Leveraging in-house experience and technology for syntactic and semantic parsing of running text, primarily in English, we have worked with ToE colleagues at *Vrije Universiteit Amsterdam* to (a) help improve the content extraction pipeline and (b) establish consensus of a ToE-compatible RDF encoding of LAP annotations. We expect that making available a standardized collection of (automatically acquired) linguistic annotations of the raw text in the corpus, related to the resource at large through RDF links, will enable a variety of downstream analytical tasks and aid replicability and comparability of results.

The LAP workflow that produces these annotations consists of the CIS Tokenizer sentence segmenter, the REPP word tokenizer (Dridan and Oepen, 2012), and the Bonhet and Nivre part-of-speech tagger and dependency parser (Bohnet and Nivre, 2012). The resulting annotations are then exported to RDF triples as defined by a LAP ontology, and are linked to the relevant speech using named graphs in TriG syntax. As of September 2015, the ToE data consists of 292,379 speeches, amounting to roughly 63 million tokens and resulting in approximately 2.7 billion triples once annotated in LAP. An example showing both the form of the annotations and how to access them via a SPARQL endpoint can be found on the LAP development wiki pages. [5] Integration of this vast collection of LAP-derived triples with the main ToE store at Amsterdam remains to be negotiated.

## 7 Current State of Development and Future Plans

A LAP development instance has been available for trial use since late 2014,[6] with a production instance set to launch towards the end 2015. The tools currently installed in LAP target Norwegian and Sami as well as English, allowing users to process raw text, analyze it, and export the resulting annotations in different formats: various tabulator-separated formats in the tradition of the shared tasks of the Conference on Natural Language Learning; two common variants of the Constraint Grammar textual exchange format; and the RDF representation designed for the ToE use case. Documentation guiding new users through their first sessions in LAP is available in the LAP project pages.[7] Our short- to mid-term goals with regard to tool development include (a) broadening the range of processing types covered (for example to support language identification; ‘deeper’, semantic parsing; training and application of document- and substring-level classifiers; and others), (b) supporting parameterizable extraction of (relevant) linguistic content from various mark-up formats, as well as (c) developing import and export interfaces with other CLARINO platforms such as the Corpuscle and Glossa corpus search services.

[4] http://www.talkofeurope.eu/

[5] http://moin.delph-in.net/LapDevelopment/ToE

[6] https://lap.hpc.uio.no/

[7] http://www.mn.uio.no/ifi/english/research/projects/clarino/user/

The LAP trial instance currently supports authentication through the Norwegian national Feide federation and the world-wide eduGAIN interfederation, which in in late 2015 counts some 40 national federations among its members. In addition to the above, the forthcoming production instance will also allow authorization via the CLARIN IdP and Feide OpenIdP services (enabling new users to self-register), as it appears that some CLARIN member sites regrettably have yet to gain access to eduGAIN. Use of the Norwegian national supercomputing e-infrastructure 'behind the scenes' of LAP mandates full accountability, where the individual allocation of compute cycles will depend on user affiliations and the choice of authentication service. At present, (a) individual users (including international ones) can be granted a standard quota of up to 5,000 cpu hours per six-month period; (b) users or groups of users can apply for LAP projects, where a project typically can receive an allocation of up to 100,000 cpu hours per period; and (c) users or groups of users can independently acquire separate allocations through the Norwegian national Notur services, which they can transparently use for LAP computation.

For compatibility with other Galaxy-based portals operated at the University of Oslo (UiO) and in the national ELIXIR.NO network, the LAP trial instance currently still builds on a 2013 snapshot of the Galaxy code base, but the forthcoming production instance will be the first UiO portal running on a current (March 2015) Galaxy release. While the Galaxy framework continues to evolve dynamically, the so-called 'tool descriptions'—(mostly) declarative specifications of the interfaces and customization options for a specific processing tool—are often not fully portable across different Galaxy versions. With the launch of the LAP production instance, we expect to emphasize reproducibility and replicability (see §4 above). While the LAP Tree and API to the annotation database were designed with these goals in mind, the interface to Galaxy proper may, thus, pose technical challenges when future LAP instances migrate to newer Galaxy code; we optimistically expect that the tool description framework will evolve in a backwards-compatible manner, but in the extreme changes at this interface level might call for updates to not only current revisions of tool descriptions but also to historic ones.

As our choice of building the CLARINO Language Analysis Portal on Galaxy has recently been followed by the US-based LAPPS and Australian Alveo initiatives (see above), we plan to reach out to these developer communities and exchange experiences (as well as strengthen the recognition of NLP-specific needs in the Galaxy community), for example through a jointly organized workshop at the 2016 Galaxy Community Conference.

## References

Daniel Blankenberg, Gregory Von Kuster, Nathaniel Coraor, Guruprasad Ananda, Ross Lazarus, Mary Mangan, Anton Nekrutenko, and James Taylor. 2010. Galaxy. A web-based genome analysis tool for experimentalists. *Current Protocols in Molecular Biology*, page 19.10.1 – 21, January.

Bernd Bohnet and Joakim Nivre. 2012. A transition-based system for joint part-of-speech tagging and labeled non-projective dependency parsing. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Conference on Natural Language Learning*, page 1455 – 1465, Jeju Island, Korea.

Rebecca Dridan and Stephan Oepen. 2012. Tokenization. Returning to a long solved problem. A survey, contrastive experiment, recommendations, and toolkit. In *Proceedings of the 50th Meeting of the Association for Computational Linguistics*, page 378 – 382, Jeju, Republic of Korea, July.

Dominique Estival and Steve Cassidy. 2014. Alveo, a human communication science virtual laboratory. In *Australasian Language Technology Association Workshop 2014*, page 104.

Belinda Giardine, Cathy Riemer, Ross C. Hardison, Richard Burhans, Laura Elnitski, Prachi Shah, Yi Zhang, Daniel Blankenberg, Istvan Albert, James Taylor, Webb Miller, W. James Kent, and Anton Nekrutenko. 2005. Galaxy. A platform for interactive large-scale genome analysis. *Genome Research*, 15(10):1451 – 5, October.

Jeremy Goecks, Anton Nekrutenko, James Taylor, and The Galaxy Team. 2010. Galaxy. A comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology*, 11(8:R86), August.

Ulrich Heid, Helmut Schmid, Kerstin Eckart, and Erhard W. Hinrichs. 2010. A corpus representation format for linguistic web services. The D-SPIN Text Corpus Format and its relationship with ISO standards. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, page 494 – 499, Valletta, Malta.

Marie Hinrichs, Thomas Zastrow, and Erhard W Hinrichs. 2010. Weblicht: Web-based lrt services in a distributed escience infrastructure. In *LREC*.

Bjørn Høyland, Jean-François Godbout, Emanuele Lapponi, and Erik Velldal. 2014. Predicting party affiliations from European Parliament debates. In *Proceedings of the 52nd Meeting of the Association for Computational Linguistics: Workshop on Language Technologies and Computational Social Science*, page 56 – 60, Baltimore, MD, USA.

Nancy Ide and Laurent Romary. 2001. A common framework for syntactic annotation. In *Proceedings of the 39th Meeting of the Association for Computational Linguistics*, page 306 – 313, Toulouse, France, July.

Nancy Ide and Keith Suderman. 2013. The Linguistic Annotation Framework: A standard for annotation interchange and merging. *Language Resources and Evaluation*, (forthcoming).

Nancy Ide, James Pustejovsky, Christopher Cieri, Eric Nyberg, Denise DiPersio, Chunqi Shi, Keith Suderman, Marc Verhagen, Di Wang, and Jonathan Wright. 2014. The language application grid. *Proceedings of the Ninth International Language Resources and Evaluation (LREC14), Reykjavik, Iceland. European Language Resources Association (ELRA)*.

Emanuele Lapponi, Erik Velldal, Nikolay A. Vazov, and Stephan Oepen. 2013. Towards large-scale language analysis in the cloud. In *Proceedings of the 19th Nordic Conference of Computational Linguistics: Workshop on Nordic Language Research Infrastructure*, page 1 – 10, Oslo, Norway.

Emanuele Lapponi, Erik Velldal, Stephan Oepen, and Rune Lain Knudsen. 2014. Off-road LAF: Encoding and processing annotations in NLP workflows. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, page 4578 – 4583, Reykjavik, Iceland.

Maarten van Gompel and Martin Reynaert. 2013. FoLiA: A practical xml format for linguistic annotation - a descriptive and comparative study. *Computational Linguistics in the Netherlands Journal*, 3:63–81, 12/2013.

Joris van Zundert. 2012. If you build it, will we come? Large scale digital infrastructures as a dead end for digital humanities. *Historical Social Research*, 37(3).

# Cloning and porting the LINDAT DSpace for the CLARINO Bergen repository

**Rune Kyrkjebø**
University of Bergen Library
Bergen
Norway
rune.kyrkjebo@uib.no

**Hemed Al Ruwehy**
University of Bergen Library
Bergen
Norway
hemed.ruwehy@uib.no

**Øyvind Liland Gjesdal**
University of Bergen Library
Bergen
Norway
oyvind.gjesdal@uib.no

## Abstract

This poster and demo presents the architecture of the CLARINO Bergen repository. We describe the motivation for using the LINDAT repository as a model. We then describe the process of setting up the required functions by adapting the LINDAT software where needed.

## 1 Motivation

Every CLARIN Centre type B is required to run a dedicated language resource data repository in accordance with certain criteria for good practice and compatibility in the CLARIN infrastructure.[1] The University of Bergen Library (UBL) which participates in CLARINO[2] was assigned the task of implementing and running a repository to primarily manage the resources at the University of Bergen. The repository is also open to other partners in CLARINO and to the whole CLARIN community.

In 2013 the University of Bergen decided to use the open software application DSpace,[3] as modified by the Institute of Formal and Applied Linguistics at the Charles University in Prague for their CLARIN/LINDAT repository.[4] The motivations for this choice were the following.

UBL had some previous experience with DSpace for the implementation of the Bergen Open Research Archive.[5] This experience showed that DSpace is a functional and stable platform which is open source and well maintained by an active user community. It provides long term storage and linking, suitable authentication mechanisms, handling of licenses for downloading of resources, and metadata can be harvested at an OAI-PMH endpoint.

Furthermore, UBL was exposed to the LINDAT presentation at the CLARIN meeting in June 2013 in Utrecht where the Prague group was willing to share their software and knowledge. Some strengths of the CLARIN community is the fact that much of the software is open source and that mobility actions can be used to get assistance across borders.

For these reasons we decided to proceed directly with implementing DSpace.[6] We hereby present our experience as a good example of the value of sharing technical solutions within the CLARIN community.

[1] https://www.clarin.eu/sites/default/files/CE-2013-0095-B-centre-checklist-v4.pdf

[2] A Norwegian national infrastructure project contributing to CLARIN, http://clarin.b.uib.no

[3] http://www.dspace.org/

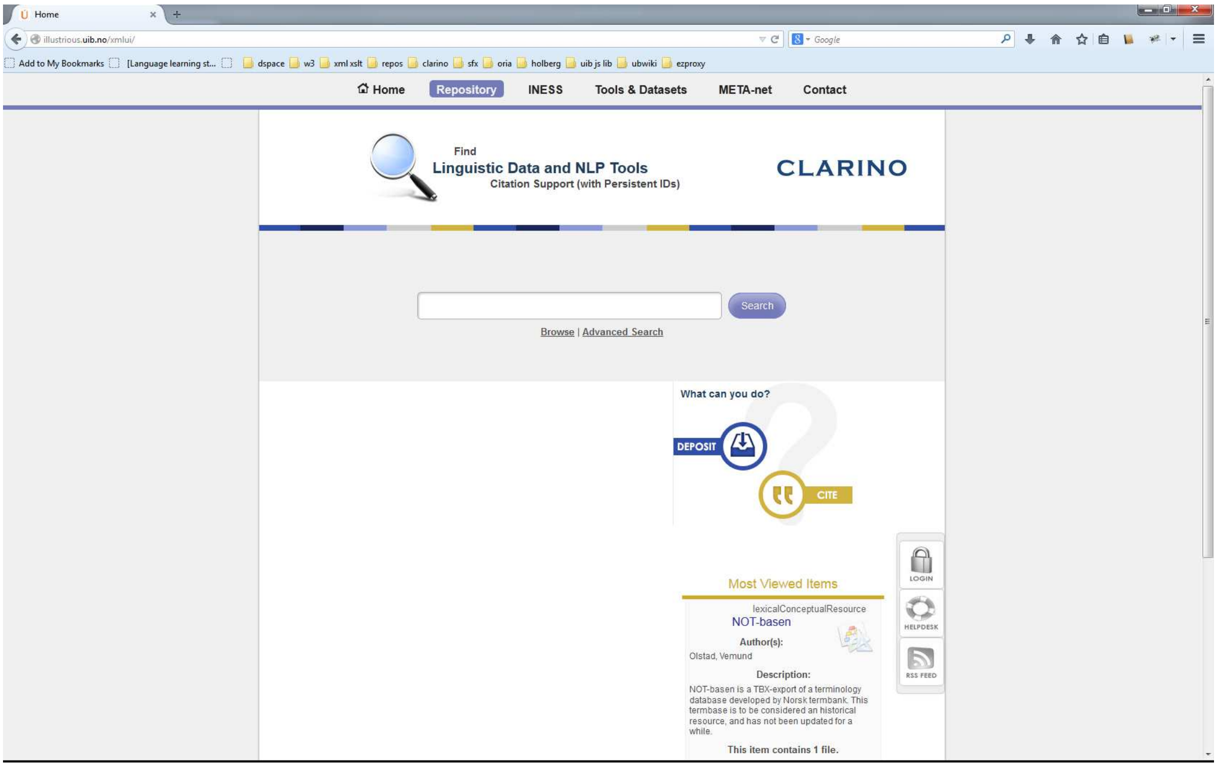
[4] https://lindat.mff.cuni.cz/repository/xmlui/

[5] http://bora.uib.no

[6] In the future we might look at how FEDORA is implemented both in the CLARIN community and elsewhere to build repository infrastructure.

## 2 Installation and adaptation

A mobility action funded by CLARIN enabled Jozef Mišutka to travel from Prague to Bergen in August 2013 in order to help set up the initial system. This mobility action was probably far more efficient than attempting to communicate at a distance. Indeed, within a few days, the first version of the installation was up and running. It was at first a clone of the LINDAT repository with minimal adaptations, as shown in Figure 1.



lexicalConceptualResource
NOT-basen
Author(s):
Olstad, Vemund
Description:
NOT-basen is a TBX-export of a terminology database developed by Norsk termbank. This termbase is to be considered an historical resource, and has not been updated for a while.
This item contains 1 file.

Figure 1: LINDAT clone, minimally adapted

The main features which had been added in LINDAT in order adapt DSpace for use in CLARIN are *CMDI metadata integration* and a method for license handling which adds the possibility of *signing licenses*. These are important functions for operation in a CLARIN context.

The LINDAT/CLARIN software is in an open source software repository at GitHub.[7]

Local modifications and configurations were necessary for the graphical profile, for authentication and for Persistent Identifiers (PIDs). A graphical profile for CLARINO was designed by Talan Memmott (University of Bergen), as shown in Figure 2. Federated single sign-on was installed by the UBL in cooperation with the IT-department at the University of Bergen, with helpful guidance from the Shibboleth setup guide by Sander Maijers, published by CLARIN ERIC.[8]

A Handle[9] configuration was set up by the UBL in order to assign PIDs to resources. We did not develop our own (PID) server, as it comes as a built-in feature in DSpace. The UBL bought the handle prefix and configured the PID server as described in the handle documentation.

From there, we have repeated the installation several times to be in sync with updates; we have also made further local customizations mostly to the user interface and we upgraded the repository to DSpace 5.2 in June 2015.[10] The front page of the site offers an immediate entry to the repository, but

[7] https://github.com/ufal/lindat-dspace

[8] Sander Maijers: *Your own Shibboleth Service Provider within the Service Provider Federation*. https://cdn.rawgit.com/clarin-eric/SPF-tutorial/master/Shib_SP_tutorial.html

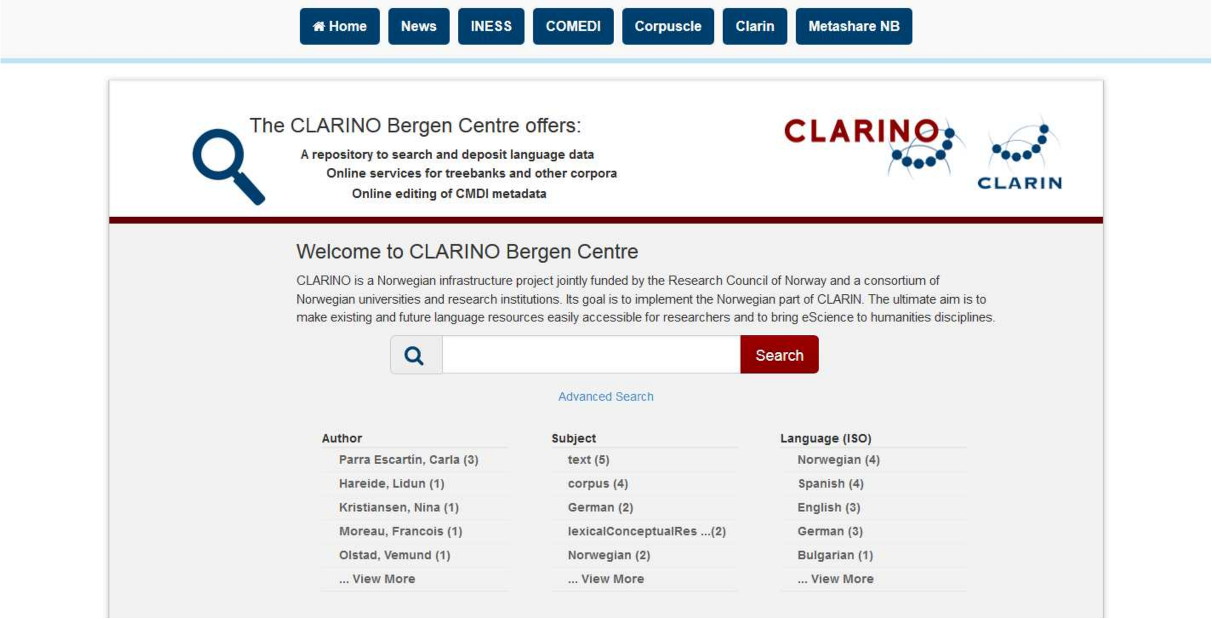
[9] http://handle.net

[10] The Bergen CLARINO repository is available at http://clarino.uib.no

it also has links to the other parts of CLARINO: the CLARINO news blog, the INESS[11] treebanking infrastructure, the COMEDI[12] component metadata editor, and the CORPUSCLE[13] advanced corpus management and search system. There are also links to the CLARIN ERIC website and to the Norwegian META-SHARE[14] node.

In terms of human resources our solution requires the library to have at least one programmer, who is backed up with planning and support from management and colleagues, and the university IT-department. At present we have two programmers in our digital systems section.

The time and effort spent to install the installation can be estimated to 5 man-months of programmer work on the library side. Graphical design and branding of the repository has  been done with the effort of the CLARINO community.



The CLARINO Bergen Centre offers:

A repository to search and deposit language data
Online services for treebanks and other corpora
Online editing of CMDI metadata

Welcome to CLARINO Bergen Centre

CLARINO is a Norwegian infrastructure project jointly funded by the Research Council of Norway and a consortium of Norwegian universities and research institutions. Its goal is to implement the Norwegian part of CLARIN. The ultimate aim is to make existing and future language resources easily accessible for researchers and to bring eScience to humanities disciplines.

| Author | Subject | Language (ISO) |
|---|---|---|
| Parra Escartín, Carla (3) | text (5) | Norwegian (4) |
| Hareide, Lidun (1) | corpus (4) | Spanish (4) |
| Kristiansen, Nina (1) | German (2) | English (3) |
| Moreau, Francois (1) | lexicalConceptualRes ...(2) | German (3) |
| Olstad, Vemund (1) | Norwegian (2) | Bulgarian (1) |
| ... View More | ... View More | ... View More |

Figure 2: CLARINO Bergen Center and repository homepage, September 2015 (http://clarino.uib.no)

## 3 Operation and perspectives

Once the repository was up and running, an initial set of metadata was batch imported after converting existing metadata from the META-SHARE format. We started uploading data for the resources. Further data and metadata is continuously being added. As of september 2015 the repository contains 12 items for 8 languages, and the types of the current items are *corpus* and *lexicalConceptualResource*.

Converting META-SHARE metadata to CMDI was achieved using XSLT transformations to map the META-SHARE schema to the DSpace built-in metadata schema. LINDAT has provided a CMDI component and an internal mapping in dspace to this component, and also the functionality for uploading CMDI metadata.

In our view, the CLARIN group in Prague has invested considerable efforts in making a solid repository system compatible with CLARIN and to make the system easy for others to adopt. Our installation and porting process has shown CLARIN partners in different countries have much to gain from sharing software and from cooperating through targeted mobility actions.

The present situation allows the emerging Bergen CLARINO center to apply for the Data Seal of Approval.

[11] The Norwegian Infrastructure for the Exploration of Syntax and Semantics, http://clarino.uib.no/iness/page
[12] A web-based editor for CMDI-conformant metadata, http://clarino.uib.no/comedi/page
[13] http://clarino.uib.no/korpuskel/page
[14] http://www.meta-net.eu/meta-share

However, a remaining challenge relates to the metadata flow. The LINDAT repository system is able to generate CMDI metadata, as required by CLARIN, from DSpace internal metadata fields. However, there are differences between the Norwegian (CLARINO) and Czech (LINDAT) metadata profiles. Furthermore, DSpace cannot accommodate arbitrary new CMDI profiles. A principal limitation is that DSpace local metadata are a flat structure while CMDI has a hierarchical structure with possible embedding of components in components.

Since the COMEDI metadata editor allows the production of metadata according to any CMDI profile, a current strategy is to produce metadata in COMEDI and to upload the CMDI metadata stream into the repository. However, these CMDI metadata are then not linked to the internal metadata of the repository system. Filling out metadata twice seems like a step to be avoided.

Currently depositing an item to CLARINO Bergen Repository for which metadata has been created using COMEDI needs manual upload of the file/item. In order to simplify the process, we would like to have a tighter integration of COMEDI in the depositing workflow. This could be achieved by automatically redirecting the user back to the CLARINO Bergen Repository when finished editing the metadata, and have the CLARINO Bergen Repository import the metadata using the JSON/REST API of COMEDI.

# Inforex – a web-based tool for corpora management, browsing and annotation

**Michał Marcińczuk**
Wrocław University of Technology
Wybrzeże Wyspiańskiego 27
Wrocław, Poland
michal.marcinczuk@pwr.edu.pl

**Jan Kocoń**
Wrocław University of Technology
Wybrzeże Wyspiańskiego 27
Wrocław, Poland
jan.kocon@pwr.edu.pl

**Marcin Oleksy**
Wrocław University of Technology
Wybrzeże Wyspiańskiego 27
Wrocław, Poland
marcin.oleksy@pwr.edu.pl

## 1 Introduction

Inforex is a web-based system for text corpora management, annotation and visualisation. We have been developing the system since 2010 and it was already used in several projects, i.e., a construction of Polish Corpus of Wrocław University of Technology called KWPr (Broda et al., 2012) within the NEKST[1] project, a construction of a Corpus of Economic News called CEN (Marcińczuk et al., 2013) within the SyNaT project and a construction of a Polish Corpus of Suicide Notes called PCSN[2] (Marcińczuk et al., 2011) guided by Monika Zaśko-Zielińska (2013). The system supports a wide range of common tasks related to text annotation, including text clean-up, annotation of named entities, word senses, semantic relations, anaphora, etc. In 2015 the system was incorporated into the CLARIN-PL infrastructure. The system was integrated with a data repository DSpace running as a part of CLARIN-PL. The integration allows users to import data from the repository to Inforex for data visualisation and further modifications.

## 2 Inforex as a Part of CLARIN-PL Infrastructure

Inforex[3] became a part of Polish Common Language Resources & Technology Infrastructure (CLARIN-PL[4]). The access to the system is granted to users after creating an account in the DSpace[5] repository. At the moment users use the same login and password as the one defined in the DSpace system. After login to the DSpace repository user can import a text corpus to its account in the Inforex system. The corpus must be processed to the CCL[6] format beforehand. This can be done by the "Process files to CCL" option. This option processes the corpus with a set of NLP tools, including text conversion, segmentation, tagging, word sense disambiguation, named entities recognition and temporal expressions recognition. When the processing is done, *Export to Inforex* option become available. After clicking the import button user is redirected to a page in Inforex, where the progress of the import process can be monitored. In the background, DSpace prepares a package with the data to import and moves it to a drive shared by DSpace and Inforex, and registers the request in the Inforex system. There is set of Inforex import daemons which are responsible for the import process. When a daemon receives a notification about a new request, it starts the import process by converting the CCL files into the Inforex document representation. The import status is updated after processing each CCL file in the package. The Inforex import daemons can be run in parallel and there are at least four processes. The number of daemons can be easily increased if needed. The import process is illustrated on Figure 1.

After importing the corpus to Inforex, the user can browse the content of documents, display the recognized named entities and temporal expressions, export a list of recognized annotations, verify the automatically recognized annotations and add new ones.

[1] http://nekst.ipipan.waw.pl/

[2] http://pcsn.uni.wroc.pl/

[3] http://inforex.clarin-pl.eu

[4] http://clarin-pl.eu/

[5] https://clarin-pl.eu/dspace/

[6] http://nlp.pwr.wroc.pl/redmine/projects/corpus2/wiki/CCL_format

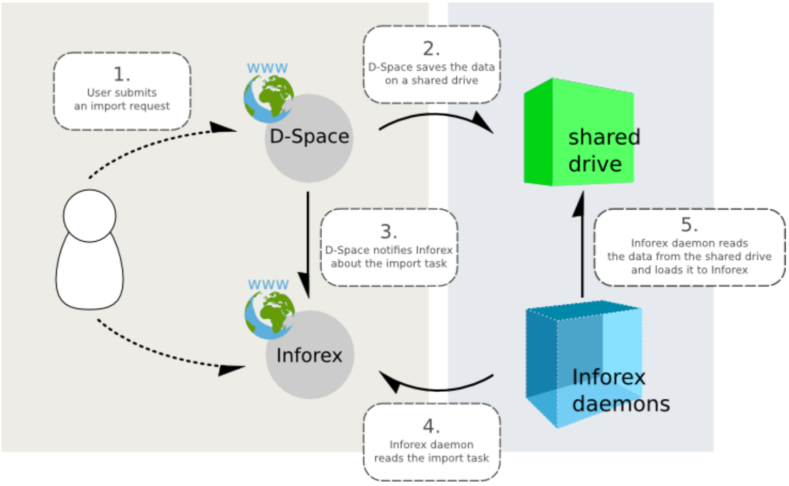


Figure 1: Inforex – import process

## 3 New Features

Comparing to the older version of Inforex (Marcińczuk et al., 2012) a set of new features was implemented. The major features are presented in the following subsections.

### 3.1 Wccl Match

Wccl Match (Marcińczuk and Radziszewski, 2013) is a formalism for text annotation. It allows to write rules which locate a defined sequence of tokens over a morphosyntactically and semantically tagged text, and annotate the matched sequence or its subsequences with the given labels. User can use Wccl Match rules to search for relevant text fragments in the imported corpora. The interface supports syntax highlighting and provides a toolbox with all Wccl Match operators. At the current version Inforex shows a list of sentences with the matched sequences. In the future we plan to implement two features based on the Wccl Match mechanism. The first one is an automatic text annotation with Wccl Match rules. The other is an export of sentences matched with the rules. Figure 2 shows Wccl Match feature in Inforex.

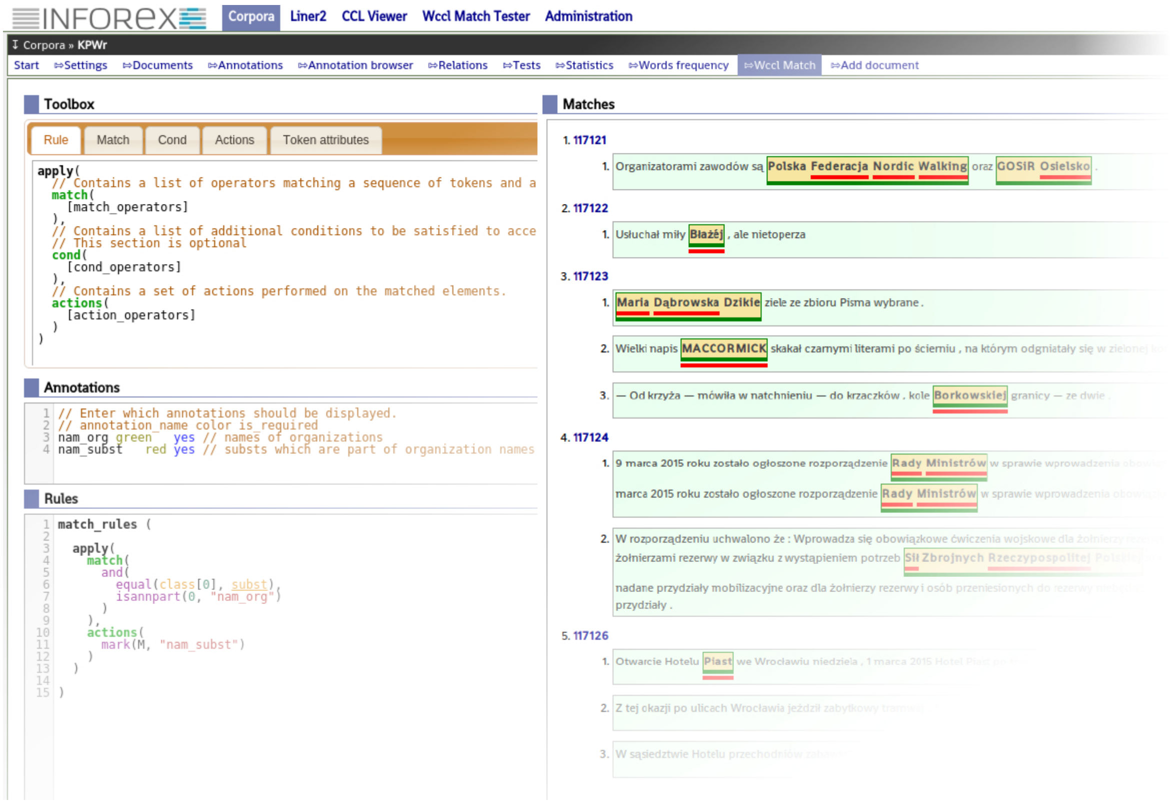
### 3.2 Wccl Match Tester

Wccl Match Tester is another application of the Wccl Match formalism. It allows to evaluate a set of Wccl Match rules which recognize a certain set of annotations (named entities, temporal expressions, etc.) on an annotated corpus. Wccl Match Tester is also an environment for rapid rule development. User interface displays the number of correctly, incorrectly and not recognized annotations for a predefined set of annotation categories and a list of those annotations. User can filter the list of annotations by their category and result type. Figure 3 shows Wccl Match Tester feature in Inforex.

### 3.3 CorpoGrabber

CorpoGrabber is the tool, which allows to get the most relevant content of the website, including all subsites. With this tool a user can build a big Web corpora, having only a list of websites as the input. The core tools composing CorpoGrabber were adapted to Polish, but most of the toolchain elements are language independent. The whole process includes the following tasks:

- downloading of the HTML subpages of each input page URL using HTTrack tool (Russell and Cohn, 2012)
- extracting of plain text from each subpage by removing boilerplate content (such as navigation links, headers, footers, advertisements from HTML pages) with JusText tool (Pomikálek, 2011)



```
apply(
  // Contains a list of operators matching a sequence of tokens and a
  match(
    [match_operators]
  ),
  // Contains a list of additional conditions to be satisfied to acce
  // This section is optional
  cond(
    [cond_operators]
  ),
  // Contains a set of actions performed on the matched elements.
  actions(
    [action_operators]
  )
)
```

Annotations

```
// Enter which annotations should be displayed.
// annotation_name color is_required
nam_org green   yes // names of organizations
nam_subst   red yes // substs which are part of organization names
```

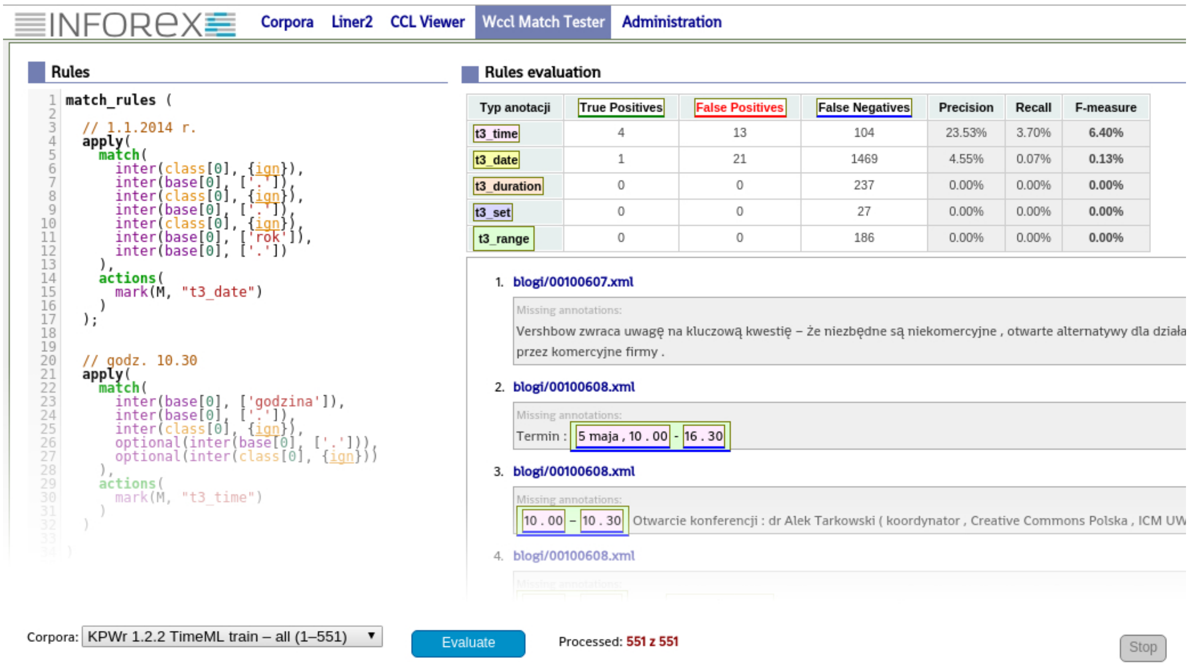
Rules

```
match_rules (

  apply(
    match(
      and(
        equal(class[0], subst),
        isannpart(0, "nam_org")
      )
    ),
    actions(
      mark(M, "nam_subst")
    )
  )

)
```

Matches

1. 117121
   1. Organizatorami zawodów są Polska Federacja Nordic Walking oraz GOSiR Osielsko .
2. 117122
   1. Usłuchał miły Błażéj , ale nietoperza
3. 117123
   1. Maria Dąbrowska Dzikie ziele ze zbioru Pisma wybrane .
   2. Wielki napis MACCORMICK skakał czarnymi literami po ścierniu , na którym odgniatały się w zielonej ko
   3. — Od krzyża — mówiła w natchnieniu — do krzaczków , kole Borkowskiej granicy — ze dwie .
4. 117124
   1. 9 marca 2015 roku zostało ogłoszone rozporządzenie Rady Ministrów w sprawie wprowadzenia obowią
      marca 2015 roku zostało ogłoszone rozporządzenie Rady Ministrów w sprawie wprowadzenia obowiązk
   2. W rozporządzeniu uchwalono że : Wprowadza się obowiązkowe ćwiczenia wojskowe dla żołnierzy rezerw
      żołnierzami rezerwy w związku z wystąpieniem potrzeb Sił Zbrojnych Rzeczypospolitej Polskiej
      nadane przydziały mobilizacyjne oraz dla żołnierzy rezerwy i osób przeniesionych do rezerwy nieb
      przydziały .
5. 117126
   1. Otwarcie Hotelu Piast we Wrocławiu niedziela , 1 marca 2015 Hotel Piast po
   2. Z tej okazji po ulicach Wrocławia jeździł zabytkowy tramwaj .
   3. W sąsiedztwie Hotelu przechodniów zabaw

Figure 2: Wccl Match



```
match_rules (

  // 1.1.2014 r.
  apply(
    match(
      inter(class[0], {ign}),
      inter(base[0], ['.']),
      inter(class[0], {ign}),
      inter(base[0], ['.']),
      inter(class[0], {ign}),
      inter(base[0], ['rok']),
      inter(base[0], ['.'])
    ),
    actions(
      mark(M, "t3_date")
    )
  );


  // godz. 10.30
  apply(
    match(
      inter(base[0], ['godzina']),
      inter(base[0], ['.']),
      inter(class[0], {ign}),
      optional(inter(base[0], ['.'])),
      optional(inter(class[0], {ign}))
    ),
    actions(
      mark(M, "t3_time")
    )
  )

)
```

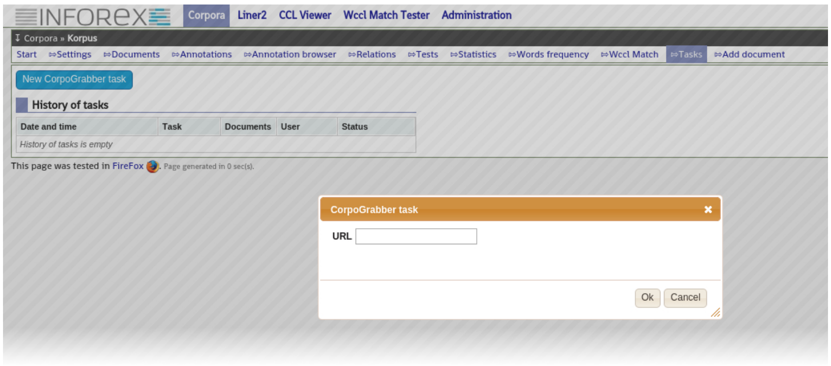
| Typ anotacji | True Positives | False Positives | False Negatives | Precision | Recall | F-measure |
|---|---|---|---|---|---|---|
| t3_time | 4 | 13 | 104 | 23.53% | 3.70% | **6.40%** |
| t3_date | 1 | 21 | 1469 | 4.55% | 0.07% | **0.13%** |
| t3_duration | 0 | 0 | 237 | 0.00% | 0.00% | **0.00%** |
| t3_set | 0 | 0 | 27 | 0.00% | 0.00% | **0.00%** |
| t3_range | 0 | 0 | 186 | 0.00% | 0.00% | **0.00%** |

1. blogi/00100607.xml
   Missing annotations:
   Vershbow zwraca uwagę na kluczową kwestię – że niezbędne są niekomercyjne , otwarte alternatywy dla działa
   przez komercyjne firmy .
2. blogi/00100608.xml
   Missing annotations:
   Termin : 5 maja , 10 . 00 - 16 . 30
3. blogi/00100608.xml
   Missing annotations:
   10 . 00 – 10 . 30 Otwarcie konferencji : dr Alek Tarkowski ( koordynator , Creative Commons Polska , ICM UW
4. blogi/00100608.xml
   Missing annotations:

Corpora: KPWr 1.2.2 TimeML train – all (1–551)

Processed: 551 z 551

Figure 3: Wccl Match Tester



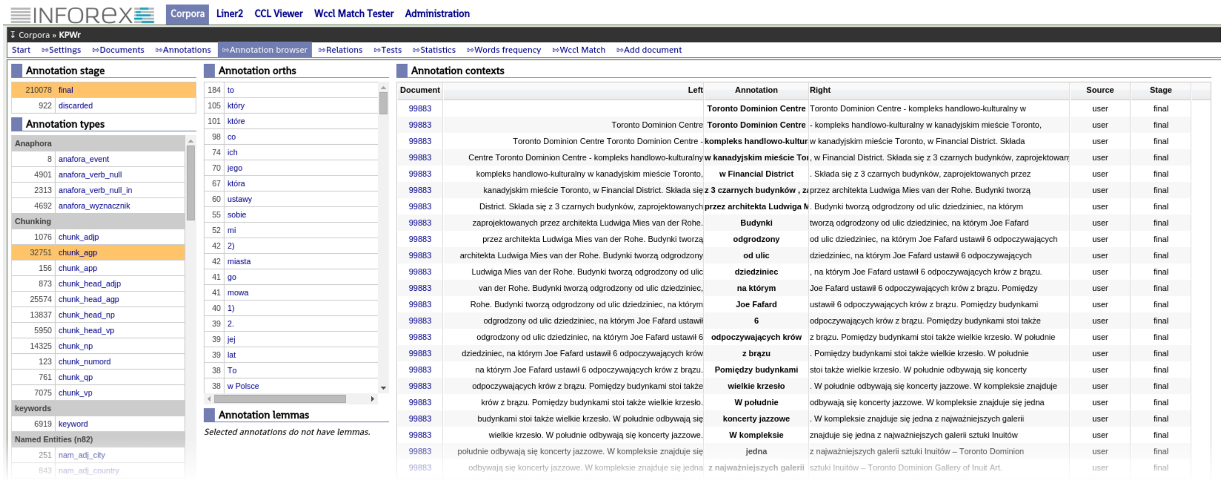
History of tasks

| Date and time | Task | Documents | User | Status |
|---|---|---|---|---|
| *History of tasks is empty* | | | | |

This page was tested in FireFox. Page generated in 0 sec(s).

URL

Figure 4: CorpoGrabber



Annotation stage

| | |
|---|---|
| 210078 | final |
| 922 | discarded |

Annotation types

| | |
|---|---|
| Anaphora | |
| 8 | anafora_event |
| 4901 | anafora_verb_null |
| 2313 | anafora_verb_null_in |
| 4692 | anafora_wyznacznik |
| Chunking | |
| 1076 | chunk_adjp |
| 32751 | chunk_agp |
| 156 | chunk_app |
| 873 | chunk_head_adjp |
| 25574 | chunk_head_agp |
| 13837 | chunk_head_np |
| 5950 | chunk_head_vp |
| 14325 | chunk_np |
| 123 | chunk_numord |
| 761 | chunk_qp |
| 7075 | chunk_vp |
| keywords | |
| 6919 | keyword |
| Named Entities (n82) | |
| 251 | nam_adj_city |
| 843 | nam_adj_country |

Annotation orths

| | |
|---|---|
| 184 | to |
| 105 | który |
| 101 | które |
| 98 | co |
| 74 | ich |
| 70 | jego |
| 67 | która |
| 60 | ustawy |
| 55 | sobie |
| 52 | mi |
| 42 | 2) |
| 42 | miasta |
| 41 | go |
| 41 | mowa |
| 40 | 1) |
| 39 | 2. |
| 39 | jej |
| 39 | lat |
| 38 | To |
| 38 | w Polsce |

Annotation lemmas

*Selected annotations do not have lemmas.*

Annotation contexts

| Document | Left | Annotation | Right | Source | Stage |
|---|---|---|---|---|---|
| 99883 | | **Toronto Dominion Centre** | Toronto Dominion Centre - kompleks handlowo-kulturalny w | user | final |
| 99883 | Toronto Dominion Centre | **Toronto Dominion Centre** | - kompleks handlowo-kulturalny w kanadyjskim mieście Toronto, | user | final |
| 99883 | Toronto Dominion Centre Toronto Dominion Centre - | **kompleks handlowo-kultur** | w kanadyjskim mieście Toronto, w Financial District. Składa | user | final |
| 99883 | Centre Toronto Dominion Centre - kompleks handlowo-kulturalny | **w kanadyjskim mieście Tor** | , w Financial District. Składa się z 3 czarnych budynków, zaprojektowan | user | final |
| 99883 | kompleks handlowo-kulturalny w kanadyjskim mieście Toronto, | **w Financial District** | . Składa się z 3 czarnych budynków, zaprojektowanych przez | user | final |
| 99883 | kanadyjskim mieście Toronto, w Financial District. Składa się | **z 3 czarnych budynków , za** | przez architekta Ludwiga Mies van der Rohe. Budynki tworzą | user | final |
| 99883 | District. Składa się z 3 czarnych budynków, zaprojektowanych | **przez architekta Ludwiga M** | . Budynki tworzą odgrodzony od ulic dziedziniec, na którym | user | final |
| 99883 | zaprojektowanych przez architekta Ludwiga Mies van der Rohe. | **Budynki** | tworzą odgrodzony od ulic dziedziniec, na którym Joe Fafard | user | final |
| 99883 | przez architekta Ludwiga Mies van der Rohe. Budynki tworzą | **odgrodzony** | od ulic dziedziniec, na którym Joe Fafard ustawił 6 odpoczywających | user | final |
| 99883 | architekta Ludwiga Mies van der Rohe. Budynki tworzą odgrodzony | **od ulic** | dziedziniec, na którym Joe Fafard ustawił 6 odpoczywających | user | final |
| 99883 | Ludwiga Mies van der Rohe. Budynki tworzą odgrodzony od ulic | **dziedziniec** | , na którym Joe Fafard ustawił 6 odpoczywających krów z brązu. | user | final |
| 99883 | van der Rohe. Budynki tworzą odgrodzony od ulic dziedziniec, | **na którym** | Joe Fafard ustawił 6 odpoczywających krów z brązu. Pomiędzy | user | final |
| 99883 | Rohe. Budynki tworzą odgrodzony od ulic dziedziniec, na którym | **Joe Fafard** | ustawił 6 odpoczywających krów z brązu. Pomiędzy budynkami | user | final |
| 99883 | odgrodzony od ulic dziedziniec, na którym Joe Fafard ustawił | **6** | odpoczywających krów z brązu. Pomiędzy budynkami stoi także | user | final |
| 99883 | odgrodzony od ulic dziedziniec, na którym Joe Fafard ustawił 6 | **odpoczywających krów** | z brązu. Pomiędzy budynkami stoi także wielkie krzesło. W południe | user | final |
| 99883 | dziedziniec, na którym Joe Fafard ustawił 6 odpoczywających krów | **z brązu** | . Pomiędzy budynkami stoi także wielkie krzesło. W południe | user | final |
| 99883 | na którym Joe Fafard ustawił 6 odpoczywających krów z brązu. | **Pomiędzy budynkami** | stoi także wielkie krzesło. W południe odbywają się koncerty | user | final |
| 99883 | odpoczywających krów z brązu. Pomiędzy budynkami stoi także | **wielkie krzesło** | . W południe odbywają się koncerty jazzowe. W kompleksie znajduje | user | final |
| 99883 | krów z brązu. Pomiędzy budynkami stoi także wielkie krzesło. | **W południe** | odbywają się koncerty jazzowe. W kompleksie znajduje się jedna | user | final |
| 99883 | budynkami stoi także wielkie krzesło. W południe odbywają się | **koncerty jazzowe** | . W kompleksie znajduje się jedna z najważniejszych galerii | user | final |
| 99883 | wielkie krzesło. W południe odbywają się koncerty jazzowe. | **W kompleksie** | znajduje się jedna z najważniejszych galerii sztuki Inuitów | user | final |
| 99883 | południe odbywają się koncerty jazzowe. W kompleksie znajduje się | **jedna** | z najważniejszych galerii sztuki Inuitów – Toronto Dominion | user | final |
| 99883 | odbywają się koncerty jazzowe. W kompleksie znajduje się jedna | **z najważniejszych galerii** | sztuki Inuitów – Toronto Dominion Gallery of Inuit Art. | user | final |

Figure 5: Annotation Browser

- de-duplication of the plain text using Onion tool (Pomikálek, 2011)
- removing of bad quality documents utilising Morphological Analysis Converter and Aggregator – MACA tool (Radziszewski and Śniatowski, 2011) (only Polish language)
- tagging of documents using Wroclaw CRF Tagger – WCRFT tool (Radziszewski, 2013) (only Polish language)

The result of the toolchain is uploaded to Inforex database as a text corpus. Each step of the processing is shown as separate task in Tasks section, with detailed information about the progress. Figure 4 shows CorpoGrabber feature in Inforex.

### 3.4 Annotation Browser

This new feature allows user to make concordances for annotations. User can see the context of the annotations of a particular (chosen by user) stage, type, text word and/or lemma. Inforex gives the ability to export selected annotations to CSV file. Figure 5 shows Annotation Browser feature in Inforex.

## 4 Summary and Future Work

We received an important and constructive feedback from scientists after three workshops on CLARIN-PL tools and resources. Compatibility of DSpace and Inforex and entirely browser-based access proved to be one of the most important features of the system. Another appreciated feature is the ability to support different text formats for uploading, thus making it easier for users to share and analyse their corpora. However, users expressed two main needs: the access to the features available only for Inforex administrator (like defining new sets of annotations) and facilitating the service (especially the user interface).

The common feature requested by users was data export, including frequency lists at the different levels (words, base forms, annotations, etc.) and a complete dump of the corpus. Those features are partially implemented but are accessible only through command-line scripts. This will be brought directly to the user interface.

## References

Bartosz Broda, Michał Marcińczuk, Marek Maziarz, Adam Radziszewski, and Adam Wardyński. 2012. KPWr: Towards a Free Corpus of Polish. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of LREC'12*, Istanbul, Turkey. ELRA.

Michał Marcińczuk, Jan Kocoń, and Bartosz Broda. 2012. Inforex – a web-based tool for text corpus management and semantic annotation. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 224–230, Istanbul, Turkey, May. European Language Resources Association (ELRA). ACL Anthology Identifier: L12-1233.

Michał Marcińczuk and Adam Radziszewski. 2013. WCCL Match – A Language for Text Annotation. In Mieczysław A. Kłopotek, Jacek Koronacki, Małgorzata Marciniak, Agnieszka Mykowiecka, and Sławomir T. Wierzchoń, editors, *Language Processing and Intelligent Information Systems*, volume 7912 of *Lecture Notes in Computer Science*, pages 131–144. Springer Berlin Heidelberg.

Michał Marcińczuk, Monika Zaśko-Zielińska, and Maciej Piasecki. 2011. Structure annotation in the polish corpus of suicide notes. In Ivan Habernal and Václav Matoušek, editors, *Text, Speech and Dialogue*, volume 6836 of *Lecture Notes in Computer Science*, pages 419–426. Springer Berlin Heidelberg.

Michał Marcińczuk, Jan Kocoń, and Maciej Janicki. 2013. Liner2 – a customizable framework for proper names recognition for Poli. In Robert Bembenik, Lukasz Skonieczny, Henryk Rybinski, Marzena Kryszkiewicz, and Marek Niezgodka, editors, *Intelligent Tools for Building a Scientific Information Platform*, pages 231–253.

Jan Pomikálek. 2011. Removing boilerplate and duplicate content from web corpora. *PhD en informatique, Masarykova univerzita, Fakulta informatiky*.

Adam Radziszewski and Tomasz Śniatowski. 2011. Maca — a configurable tool to integrate Polish morphological data. In *Proceedings of FreeRBMT11*. Software available at http://nlp.pwr.wroc.pl/redmine/projects/libpltagger/wiki.

Adam Radziszewski. 2013. A tiered CRF tagger for Polish. In H. Rybiński M. Kryszkiewicz M. Niezgódka R. Bembenik, Ł. Skonieczny, editor, *Intelligent Tools for Building a Scientific Information Platform: Advanced Architectures and Solutions*. Springer Verlag.

J. Russell and R. Cohn. 2012. *Httrack*. Book on Demand.

M. Zaśko-Zielińska. 2013. *Listy pożegnalne: w poszukiwaniu lingwistycznych wyznaczników autentyczności tekstu*. Quaestio.

# Consistent User Experience and Cross-referencing inside LINDAT/CLARIN infrastructure

**Jozef Mišutka, Ondřej Košarko, Amir Kamran, Michal Sedlák**
Institute of Formal and Applied Linguistics, Charles University in Prague
{misutka, kosarko, kamran, sedlak}@ufal.mff.cuni.cz

## Abstract

With an increasing amount of data and services available in the LINDAT/CLARIN infrastructure comes the need to properly cite, share, cross-reference and gather usage statistics. If done properly, the speed of the discover-inspect-use process should increase. The solution is to have a one size fits all framework that ensures a consistent look and the functionality mentioned above. In this paper, we present our solution and highlight several lessons learned during the development and initial deployment.

## 1 Introduction

LINDAT/CLARIN Language Research Infrastructure Centre in the Czech Republic (http://lindat.cz) provides technical background and assistance to institutions or researchers, who want to share, create and modernise their tools and data used for research in linguistics or related research fields. The centre provides an open digital repository and archive, open to all academics who want their work to be preserved, promoted and made widely available. It also hosts NLP tools (http://lindat.mff.cuni.cz/en/services) developed at Institute of Formal and Applied Linguistics (http://ufal.mff.cuni.cz/).

The centre must meet specific requirements by the funding agencies and must ensure sustainability and quality. These requirements can be met by proper policies and practices, but this puts part of the responsibility on the institutions and researchers.

If a new tool (or service) is to be included in centre's infrastructure, several prerequisites must be met; the main of them is to keep the look and feel consistent. Fulfilling these prerequisites should put no additional burden (e.g., complex changes or complete redesign) on the researchers, thus the centre aims to provide a ready-made solution. Furthermore, the tools or services often rely on specific 3rd party libraries, so the provided solution should avoid any conflicts. The more resources are part of the centre's infrastructure, the more complex and time consuming it is to deploy even a small change in the required look and feel. Therefore, the provided solution should be easy to update and it should support internationalization. At least one global language should be supported to reach potential users "world-wide"; however, as the centre is run by Czech institutions and a large portion of the data and/or tools involves Czech, it should also provide Czech version to improve the experience of native researchers.

Researchers often need to report "numbers" about their research. In the context of data and tools these usually are download counts, page views and other web metrics. A part of the centre's infrastructure is dedicated to collecting these statistics.

By utilising automated and easy to use technologies, we can meet the following basic requirements: unified look and feel of navigation and common elements, localised versions and consistent usage statistics. From the technical point of view, the most important requirements are to use precisely one generic solution (will be referred to as "common project") and to have the update process as automated as possible.

An important part of our research infrastructure is the digital repository. It is often the centre of resource discovery. LINDAT/CLARIN policy requires that the repository contains submissions describing every tool, service and data used in the infrastructure. Therefore, the repository is the ideal candidate for aggregating additional information on the data and tools inside their submission metadata. It must be noted that to fully exploit this approach the metadata must be published under a license permitting their reuse e.g., CC0[1].

## 2 Infrastructural Requirements

An important goal of the infrastructure is to improve the research impact of data and tools. In other words, the center should help data and services to be used and cited properly, to be easily discoverable, to be reused widely and the infrastructure should be able to measure basic indicators of usage. By using the common project, we can unify the experience across the services and help to reach the aforementioned goal.

In order to persuade the authors (researchers) to use the common project, it should be easily usable out of the box, should be standalone and should not conflict with other libraries. Automatic deployment can be achieved through a basic level of trust between the authors and the infrastructure where the authors create a repetitive task to update the common project automatically. This also means that the UI elements in the common project must have a generic design. From our experience, a simple mobile friendly localised header and footer with basic colour palette has worked remarkably well. The header contains elements to navigate across the infrastructure; the footer contains mainly acknowledgement and links to our partners. The footer includes simple JavaScript fragments counting usage statistics using an open-source analytics platform Piwik[2]. Using these we've met two important requirements from the funding agencies: acknowledgement and usage statistics.

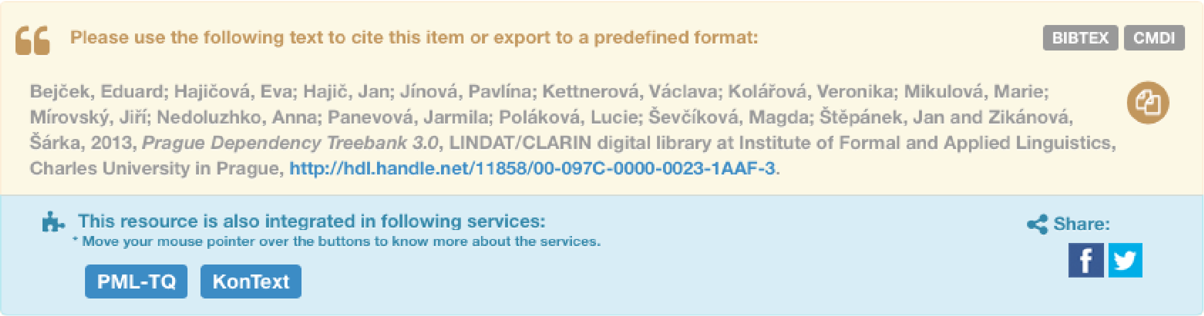
Our goal is to ensure that services and data are reused properly. This mostly means to make users cite the data but also to make them cite the data properly. We try to increase the awareness of correct data citations with Refbox - a shared UI element (see Fig. 1.).

It seems that research output has not avoided the recent trend. Following other repositories, we are supporting easy sharing via social media. If an institution or a researcher has a social media presence, they can e.g., announce a new version of a tool or a dataset to their followers with ease. This makes social sharing the next part of the Refbox.

We have been collecting datasets for some time now. Also, we have been creating, updating and hosting services that can process these datasets. The next step is to cross-reference the data and services. To give an example: there is  an engine to search corpora, there is also a treebank search service; both tools are loaded with datasets that are available in the repository; proper cross-referencing would make it possible to go from repository to one of the services and inspect the data there.  For the submitter of the data, this might improve the "discoverability" as the data are included in services that might have different user base than the repository. For the user, this allows to easily explore the data in the services and/or download them from the repository in a unified and consistent way across all the integrations. The only parameter Refbox requires is the persistent identifier of the resource, which it uses when querying the repository for metadata, so it is simple to add it into a hosted service.

[1] CC0 1.0 Universal (CC0 1.0) Public Domain Dedication, https://creativecommons.org/publicdomain/zero/1.0

[2] Piwik is a free and open source web analytics application, http://piwik.org/



Please use the following text to cite this item or export to a predefined format:

Bejček, Eduard; Hajičová, Eva; Hajič, Jan; Jínová, Pavlína; Kettnerová, Václava; Kolářová, Veronika; Mikulová, Marie; Mírovský, Jiří; Nedoluzhko, Anna; Panevová, Jarmila; Poláková, Lucie; Ševčíková, Magda; Štěpánek, Jan and Zikánová, Šárka, 2013, *Prague Dependency Treebank 3.0*, LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague, http://hdl.handle.net/11858/00-097C-0000-0023-1AAF-3.

This resource is also integrated in following services:

* Move your mouse pointer over the buttons to know more about the services.

**Figure 1. Refbox example with cross-referencing data and services**

## 3 Implementation details

Lindat-common[3] project is an attempt at keeping the look and feel of different LINDAT/CLARIN sites and services as consistent as possible. Earlier versions of the common project contained only a header, footer and their localised versions.

The common project was recently completely rewritten and updated. One of the reasons was that the project depended on Bootstrap[4]. The header and footer implemented with Bootstrap broke the design of specific services, which it should seamlessly overlay. Though, it was possible to include a standalone version via an iframe, most of the services were including the header and footer directly e.g., using 'php include'. The common project modified Bootstrap CSS classes in such a way that it was impossible to override them in general.

The new version, when deployed, does not depend on any external components or libraries. It uses Flexbox[5] CSS layout for more responsive user experience. The CSS classes are now prefixed with 'lindat-' prefix, which makes the integration into arbitrary CSS framework work seamlessly. It now integrates with the Angular[6] framework too.

One of the requirements was that the deployment should be automated. With properly set cron jobs, it is possible to distribute changes to the common project across all sites automatically.

Statistics are easily gathered by including usage tracking scripts into the footer which is used across the infrastructure.

To achieve all the goals, we decided to compile the project from source files. The result is that it is no longer an edit and commit project but development requires external projects (NodeJS[7] and Gulp[8]). On the other hand, the flexibility provided by compilation increased sustainability and usability. The source code uses LESS[9], which allows us to reduce code duplication by defining variables and using mixins[10]. The resulting JavaScript and CSS can be easily minified; this reduces the number of files that each service has to serve. Previously, there were five includes in the CSS resulting in five http requests (by default). Now, everything is inside one compact file.

The Refbox is an improvement over a former solution used to include citations of repository records in html pages, which was implemented as a standalone CSS/JavaScript library. It is now a part of the common project.. Making the cross-referencing a part of the common project seemed natural as all

[3] LINDAT/CLARIN Common Theme: current implementation of what is called "common project" in the text, https://github.com/ufal/lindat-common

[4] Bootstrap is a free and open-source collection of JavaScript and CSS tools for creating websites and web applications, http://getbootstrap.com

[5] Flexbox is a CSS box model, http://www.w3.org/TR/css3-flexbox

[6] AngularJS (commonly referred to as "Angular") is an open-source web application framework maintained by Google, https://angularjs.org

[7] Node.js is a platform built on Chrome's JavaScript runtime for easily building fast, scalable network applications, https://nodejs.org

[8] Gulp is a Task / Build runner, http://gulpjs.com

[9] LESS is a dynamic stylesheet language that can be compiled into CSS

[10] Mixins allow document authors to define patterns of property value pairs, which can then be reused in other rulesets

the services are using it anyway. It is also in accordance with the requirement of easy and automated deployment.

The Refbox uses REST API provided by our repository to get the required information. This makes it easy to programmatically access items and their metadata. The REST API lets us receive the response in JSON format, which is more convenient to work with inside JavaScript than exploiting OAI-PMH[11] protocol with customised metadata crosswalks as in the previous version. We parse the JSON and dynamically create the html fragment displaying the obtained data. The development cycle is usually shorter on the client side i.e., it should be easier, quicker and with minimum downtime.

The inclusion of the Refbox is very simple and it is designed in such a way that it does not interfere with the page that includes it. In case it does, the fallback plan is to use the raw JSON metadata to construct the Refbox by the service itself.

Let's take as an example the "Prague dependency treebank 3.0 (PDT)"[12]. This dataset is currently used in two LINDAT/CLARIN services, in KonText[13] and in PML-Tree Query[14]. Both services allow the user to search the dataset but on different levels. PML-Tree Query makes it possible to include the tree structure in your searches, whereas KonText uses only the "flat" textual data underlying the treebank. When using the Refbox in this particular case, the user can easily explore the data in one of the services and quickly navigate back or forth to download location.

Another prominent part of the Refbox is the social media sharing capabilities. At the moment, we are analysing various available social media integration libraries that fit our requirements.

Due to the fact that we compile the project, we can revisit our approach to localisation. Previously, there were different versions for the header and the footer for each language. It is trivial to see that it is not optimal. The new solution is to template one set of pages with constants translated into different languages and create localised versions during compilation.

## 4 Conclusion

We have introduced the common project used in the LINDAT/CLARIN infrastructure, which helps citing, sharing, cross-referencing and gathering statistics in a consistent way for data and services. Our goals are to improve the user experience and to motivate more researchers to share their data and tools.

We have described some of the caveats experienced while maintaining this common infrastructure and presented an improved version addressing these issues.

It will be interesting to observe whether our solution improves the data citations, whether the integration of social networks increases the accessibility of the data and services; and whether the researchers get motivation to actively promote their work. However, as the framework is going to be deployed in the next couple of days, it is not yet possible to evaluate if we meet all our expectations.

[11] The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) is a low-barrier mechanism for repository interoperability, https://www.openarchives.org/pmh

[12] Prague Dependency Treebank 3.0, http://hdl.handle.net/11858/00-097C-0000-0023-1AAF-3

[13] An interactive web demo for querying selected ÚFAL corpora based on the KonText project of the Institute of Czech National Corpus, http://hdl.handle.net/11234/1-1452

[14] System for querying annotated treebanks in PML format, http://hdl.handle.net/11858/00-097C-0000-0022-C7F6-3

# A comprehensive valence dictionary of Polish *Walenty* and its XML encoding

**Bartłomiej Nitoń, Tomasz Bartosiak, Elżbieta Hajnicz,**
**Agnieszka Patejuk, Adam Przepiórkowski, Marcin Woliński**
Institute of Computer Science, Polish Academy of Sciences

## 1 General statements

The aim of this paper is to present a current version of *Walenty*, a comprehensive valence dictionary of Polish developed at the Institute of Computer Science, Polish Academy of Sciences (ICS PAS), available for anyone by accessing the page `http://walenty.ipipan.waw.pl/`. It is created in several projects, but mainly as a part of CLARIN-PL. The dictionary is meant to be both human- and machine-readable; in particular, it is being employed by two parsers of Polish, Świgra[1] (Woliński, 2004) and POLFIE[2] (Patejuk and Przepiórkowski, 2012) developed as a part of the CLARIN infrastructure. The former, Świgra, is an implementation of the DCG grammar of Polish of (Świdziński, 1992). The latter, POLFIE, is an implementation of an LFG grammar of Polish. As these parsers are based on two rather different linguistic approaches, the valence dictionary must be sufficiently expressive to accommodate for the needs of both – and perhaps other to come. For this reason, *Walenty* exhibits a number of features which are rare or absent in other valence dictionaries.

The lexicon is represented in several formats. Its main internal format—used in a dedicated tool *Slowal* to introduce, modify and view its entries—is a database. The external formats are: textual (the most compact and poorly readable, intended mainly for parsers), PDF (intended only for human users) and XML (even less readable than text format, intended for communication with other tools and resources).

*Walenty* consists of two layers, syntactic and semantic, which are directly connected. Therefore, each lexical entry contains a number of syntactic valence schemata and semantic valence frames. Each schema is connected with at least one frame and each frame has at least one schema attached. Both schemata and frames are illustrated by at least one attested exemplary sentence, preferably from NKJP corpus `http://nkjp.pl/`; (Przepiórkowski et al., 2012). Each example is linked to a schema, and all phrase types filled in this sentence are marked.

## 2 Syntactic layer

Each schema is a set of syntactic positions (Szupryczyńska, 1996). *Walenty* uses coordination test to distinguish positions: if two morphosyntactically different phrases may occur coordinated, they belong to the same position. Usual phrase types are considered, such as nominal phrases (`np`), prepositional phrases (`prepnp`), adjectival phrases (`adjp`), clausal phrases (`cp`), etc. There are two labelled positions, subject and object.

This is exemplified in a schema (1) for the verb tłumaczyć 'explain', as used in (2) involving a coordinated phrase in the object position, consisting of an `np` (*najprostsze zasady* 'the most basic principles') and an interrogative clause (*dlaczego trzeba je stosować* 'why they should be adhered to'; marked here as `cp(int)`).The nominal realisation of the object is specified as structural, as it normally occurs in the accusative, unless the verb is nominalised or the object is in the scope of verbal negation, in which case it bears the genitive case (on the so-called Genitive of Negation in Polish, see (Przepiórkowski, 2000) and references therein). The structural case is used for the nominal subject for similar reasons. Other non-morphological cases are *partitive case* varying between accusative and genitive for partitive nouns,

[1] `http://zil.ipipan.waw.pl/Składnica`

[2] `http://zil.ipipan.waw.pl/LFG`

and for adjectives *agreeing case* (mainly for noun dependents) and *predicative case* (varying between instrumental and agreeing case on predicative position), cf. (Przepiórkowski et al., 2014a).

(1) `tłumaczyć: _: : imperf:subj{np(str)} + obj{np(str); cp(int)} + {np(dat)}`

(2) Musiałem im tłumaczyć najprostsze zasady i dlaczego trzeba je
had.1.SG.MASC they.DAT explain.INF simplest.ACC.PL principles.ACC.PL and why needs they.ACC
stosować.
apply.INF
'I had to explain to them the most basic principles and why they should be adhered to.'

Other specific phenomena we consider in *Walenty* are control (and raising), raised subject, non-nominal subjects, composed prepositions, semantically defined phrases (such as *manner* — `xp(mod)`) having separate lists of their possible surface realisations (Przepiórkowski et al., 2014a), and a rich phraseological component (Przepiórkowski et al., 2014b).

## 3 Semantic layer

Semantic frame consists of a list of semantic arguments, and each semantic argument is a pair ⟨semantic role, selectional preference⟩. Semantic roles are divided into two groups: main roles (Initiator, Theme, Stimulus, Experiencer, Instrument, Factor, Recipient, Result) and auxiliary roles (Condition, Attribute, Manner, Location, Path, Time, Duration, Measure, Purpose). These roles can be supplemented with attributes organised into pairs Foreground, Background (e.g., Initiator$^{\text{Foreground}}$, Initiator$^{\text{Background}}$ as in *someone buys something from someone, someone sells something to someone*) and Source, Goal (e.g., Location$^{\text{Source}}$, Location$^{\text{Goal}}$ as in *to go from somewhere to somewhere*).

Selectional preferences are based on PlWordNet synsets and can be represented as a list of synsets and relations to other arguments. Each semantic frame is connected with a PlWordNet lexical unit. Semantic representation of schema (1) is given in (3).

(3)

| tłumaczyć-1 | Initiator | Theme | Recipient |
|---|---|---|---|
| | osoba-1 | KONCEPCJA | osoba-1 |

osoba-1 'person' is a PlWordNet synset (represented by a corresponding lexical unit), used here to express a preference concerning humans. KONCEPCJA 'concept' is a predefined list of synsets representing human thoughts and ideas. The arguments in frame (3) are linked to positions in schema (1): Initiator to `subj`, Theme to `obj` and Recipient to `{np(dat)}`.

## 4 XML structure

CLARIN infrastructure prefers resources to be encoded in LMF format. However, *Walenty* structure is too baroque to be encoded in the rather rigid LMF structure. Therefore, we decided to use the more flexible TEI format (TEI, 2008).

Main structure of the XML file is based on TEI *Dictionary* module.[3] For each dictionary entry, we created several interconnected layers. They correspond to possible meanings of the entry (`meaningsLayer`), syntax (`syntacticLayer`), semantics (`semanticLayer`), connections between the previous two (`connectionsLayer`) and finally examples (`examplesLayer`). Each of the listed layers is defined using highly customisable TEI feature structures.[4] The connection and meanings layers are not presented in this abstract due to space limitations.

A general XML structure for `entry` (for the verb TŁUMACZYĆ 'explain') is presented in (4). All layers will be described in more details below.

(4)
```
<entry xml:id="wal_4904-ent">
  <form><orth>tłumaczyć</orth><pos>verb</pos></form>
  <fs type="syntacticLayer"><!-- presented in (5) --></fs>
  <fs type="examplesLayer"><!-- presented in (7) --></fs>
```

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

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

```
  <fs type="semanticLayer"><!-- presented in (6) --></fs>
  <fs type="meaningsLayer"><!-- not presented in this abstract --></fs>
  <fs type="connectionsLayer"><!-- not presented in this abstract --></fs>
</entry>
```

The `syntacticLayer` serves for representing the list of all schemata associated with the entry (cf. (5) for a part of schema (1)). Each `schema` consist of general characteristics (e.g., `selfMark`, `aspect`) and a list of syntactic positions (`positions`). Those positions are represented by lists of possibly coordinating phrases. Phrase type determines the possible features describing it.

(5)

```
<fs type="syntacticLayer">
  <f name="schemata">
    <vColl org="list">
      <fs xml:id="wal_4904.2598-sch" type="schema">
        <f name="opinion"><symbol value="cer"/></f>
        <f name="selfMark"><binary value="true"/></f>
        <f name="aspect"><symbol value="imperf"/></f>
        <f name="negativity"><symbol value="_"/></f>
        <f name="positions">
          <vColl org="list">
            <fs xml:id="wal_4904.2598.2-psn" type="position">
              <f name="function"><symbol value="obj"/></f>
              <f name="phrases">
                <vColl org="list">
                  <fs xml:id="wal_4904.2598.2.1-phr" type="np">
                    <f name="case">str</f>
                  </fs>
                  <fs xml:id="wal_4904.2598.2.2-phr" type="cp">
                    <f name="type">int</f>
            </fs></vColl></f></fs>
            <!-- other syntactic positions in the schema -->
      </vColl></f></fs>
      <!-- other schemata -->
</vColl></f></fs>
```

Similarly, `semanticLayer` serves for representing the list of all semantic frames of a verb, see encoding of part of frame (3) in (6). Each frame consists of two features: list of meanings (`meanings`, being references to meanings defined in `meaningsLayer`) and list of semantic arguments (`arguments`). Each `argument` is described by semantic role (features `role` and `roleAttribute`) and selective preferences (`selPrefs`), which are divided into three groups: `synsets`, predefined preferences (`predefs`) and relational preferences (`relations`).

(6)

```
<fs type="semanticLayer">
  <f name="frames">
    <vColl org="list">
      <fs xml:id="wal_4904.13-frm" type="frame">
        <f name="meanings">
          <vColl org="list">
            <fs sameAs="#wal_4904.1-mng" type="lexicalUnit"/>
        </vColl></f>
        <f name="arguments">
          <vColl org="list">
            <fs xml:id="wal_4904.13.2-arg" type="argument">
              <f name="role"><symbol value="Theme"/></f>
              <f name="roleAttribute"/>
              <f name="selPrefs">
                <fs type="selPrefsGroups">
                  <f name="synsets"><vColl org="list"/></f>
                  <f name="predefs">
                    <vColl org="list"><symbol value="KONCEPCJA"/></vColl>
                  </f>
                  <f name="relations"><vColl org="list"/></f>
            </fs></f></fs>
            <!-- other semantic arguments -->
      </vColl></f></fs>
      <!-- other semantic frames -->
</vColl></f></fs>
```

Finally, examplesLayer serves for representing the list of examples attached to the entry (cf. (7)). Structure for particular examples (example) includes meaning which points to a definition from the meaningsLayer and a list of pointers to phrases defined in syntacticLayer which are illustrated by the sentence quoted in the sentence feature.

(7)
```
<fs type="examplesLayer">
  <f name="examples">
    <vColl org="list">
      <fs xml:id="wal_4904.687-exm" type="example">
        <f name="meaning">
          <fs sameAs="#wal_4904.1-mng" type="lexicalUnit"/>
        </f>
        <f name="phrases">
          <vColl org="list">
            <fs sameAs="#wal_4904.2598.2.1-phr" type="phrase"/>
            <fs sameAs="#wal_4904.2598.2.2-phr" type="phrase"/>
        </vColl></f>
        <f name="sentence">
          Musiałem im tłumaczyć najprostsze zasady i dlaczego trzeba je stosować.
        </f>
        <f name="source"><symbol value="NKJP1800M"/></f>
        <f name="opinion"><symbol value="dobry"/></f>
</fs></vColl></f></fs>
```

## 5 Conclusions

This presentation aims at showing the XML representation of a valence dictionary of Polish *Walenty*. This representation has to take into account all phenomena represented in the dictionary together with their interdependencies. This demands, e.g., using a number of nested lists. Thus, the XML structure becomes very complicated, hence we were not able to presented it here with all details. Nevertheless, it enables us to represent the whole structure of *Walenty* in a precise and flexible way.

## References

[Patejuk and Przepiórkowski2012] Agnieszka Patejuk and Adam Przepiórkowski. 2012. Towards an LFG parser for Polish: An exercise in parasitic grammar development. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012*, pages 3849–3852, Istanbul, Turkey. ELRA.

[Przepiórkowski et al.2012] Adam Przepiórkowski, Mirosław Bańko, Rafał L. Górski, and Barbara Lewandowska-Tomaszczyk, editors. 2012. *Narodowy Korpus Języka Polskiego*. Wydawnictwo Naukowe PWN, Warsaw, Poland.

[Przepiórkowski et al.2014a] Adam Przepiórkowski, Elżbieta Hajnicz, Agnieszka Patejuk, Filip Skwarski, Marcin Woliński, and Marek Świdziński. 2014a. Walenty: Towards a comprehensive valence dictionary of Polish. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC-2010)*, pages 2785–2792, Reykjavík, Iceland. ELRA.

[Przepiórkowski et al.2014b] Adam Przepiórkowski, Elżbieta Hajnicz, Agnieszka Patejuk, and Marcin Woliński. 2014b. Extended phraseological information in a valence dictionary for NLP applications. In *Proceedings of the Workshop on Lexical and Grammatical Resources for Language Processing (LG-LP 2014)*, pages 83–91, Dublin, Ireland.

[Przepiórkowski2000] Adam Przepiórkowski. 2000. Long distance genitive of negation in Polish. *Journal of Slavic Linguistics*, 8:151–189.

[Szupryczyńska1996] Maria Szupryczyńska. 1996. Problem pozycji składniowej. In Krystyna Kallas, editor, *Polonistyka toruńska Uniwersytetowi w 50. rocznicę utworzenia UMK. Językoznawstwo*, page 135–144. Wydawnictwo Naukowe Uniwersytetu Mikołaja Kopernika, Toruń.

[Świdziński1992] Marek Świdziński. 1992. *Gramatyka formalna języka polskiego*. Rozprawy Uniwersytetu Warszawskiego. Wydawnictwa Uniwersytetu Warszawskiego, Warsaw, Poland.

[TEI2008] 2008. TEI P5: Guidelines for electronic text encoding and interchange. Internet.

[Woliński2004] Marcin Woliński. 2004. *Komputerowa weryfikacja gramatyki Świdzińskiego*. PhD thesis, Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland.

# Using PaQu for language acquisition research

**Jan Odijk**
Utrecht University
j.odijk@uu.nl

## 1 Introduction

In this paper I illustrate the use of the PaQu application for carrying out linguistic research. I describe the results of a small experiment in linguistic research into first language acquisition. The experiment was carried out mainly to test the functionality of the PaQu application, but I selected an example related to a linguistic problem that fascinates me and that has been used earlier to guide developments in the CLARIN infrastructure.

The major findings of this paper are: (1) PaQu is very useful for aiding researchers in better and more efficient manual verification of hypotheses; (2) PaQu can even be used for automatic verification of hypotheses, provided some care is exercised; (3) the Dutch CHILDES data are too small to address certain research questions; and (4) the data found suggest several new hypotheses on the acquisition of lexical properties that should be further explored.

In the final paper I will sketch the structure of the paper here.

## 2 Background

CLARIN-NL has made available a wide range of web applications for search in and analysis of linguistically annotated corpora for Dutch. These include GrETEL, OpenSONAR, FESLI, COAVA, and MIMORE.[1]

However, each of these interfaces applies to a specific set of text corpora only. Many linguists want to have such search and analysis opportunities also for the text corpora they are currently working with. To that end, the CLARIN-NL project commissioned the development of *PaQu*.

This application is available in the CLARIN infrastructure and can be accessed via the CLARIN-NL portal.

## 3 Basic facts

The three words *heel*, *erg* and *zeer* are (near-)synonyms meaning 'very'. Of these, *heel* can modify adjectival (A) predicates only, while *erg* and *zeer* can modify not only adjectival, but also verbal (V) and adpositional (P) predicates. This is illustrated in example (1)

(1) a. Hij is daar heel / erg / zeer blij over
he is there very / very / very glad about
'He is very happy about that'

b. Hij is daar *heel / erg / zeer in zijn sas mee
he is there very / very / very in his lock with
'He is very happy about that'

c. Dat verbaast mij *heel / erg / zeer
That surprises me very / very / very
'That surprises me very much'

[1] For more information on these (and other) web applications and links to them, see the CLARIN-NL portal faceted search for resources.

In (1a) the adjectival predicate *blij* ‘glad’ can be modified by each of the three words. In (1b) the (idiomatic) prepositional predicate *in zijn sas* can be modified by *zeer* and *erg* but not by *heel*. The same holds in (1c) for the verbal predicate *verbaast*.[2] In English, something similar holds for the word *very*: it can only modify adjectival predicates. For verbal and prepositional predicates one cannot use *very* but one can use the expression *very much* instead. There is a lot more to say about these data, and there are more relevant data to consider and some qualifications to be made. I refer to (Odijk, 2011) and (Odijk, 2014) for details.

## 4 Assessment of the facts

In this section I will show that the facts described cannot be derived from independent properties of grammar, language acquisition, the Dutch language, or the relevant lexical items.

I conclude that these facts must be acquired by learners of Dutch.

## 5 Research questions

The simple facts described in the previous sections raise many research question related to language acquisition. I will list several examples, but focus on the question of what kind of evidence children have access to for acquiring such properties.

In order to address these research questions, relevant data are needed, Fortunately, such data exist, and the Dutch CHILDES corpora are the most important ones in this context.

## 6 Dutch CHILDES corpora

The Dutch CHILDES corpora(MacWhinney, 2015) are accessible via the CLARIN Virtual Language Observatory (VLO) or directly via Talkbank and contain relevant material to investigate the research questions formulated in section 5. They contain transcriptions of dialogues between children acquiring Dutch on the one hand, and adults (mostly parents) and in some cases other children on the other hand, and a lot of additional information about the context, setting, age of the child, etc.

However, a serious problem for the investigation is that the words being investigated are, as any decent word in natural language, highly ambiguous. Here I will illustrate that *heel* is 6-fold ambiguous, *erg* is 4-fold ambiguous, and *zeer* is 3-fold ambiguous.

The Dutch CHILDES corpora do not contain any information about the meanings of their word occurrences. However, most of the ambiguities can be resolved by taking into account morpho-syntactic and syntactic properties of the word occurrences. More specifically, the syntactic information on these words suffices for my purposes. Unfortunately, the Dutch CHILDES corpora do NOT have (reliable) morpho-syntactic information (part of speech tags) and they do not contain syntactic information for the utterances at all.

For this reason, (Odijk, 2014, 91) carried out a manual analysis in terms of morpho-syntactic and syntactic information to disambiguate the occurrences of *heel*, *erg* en *zeer* in adult utterances of the Dutch CHILDES Van Kampen subcorpus. With PaQu, however, one can largely automate the disambiguation process, and I carried out a small test to investigate how well this works.

## 7 PaQu

PaQu is a web application developed by the University of Groningen. It enables one to upload a Dutch text corpus. This text corpus is either already parsed by Alpino, or if not, PaQu can have it automatically parsed by Alpino. After this, it is available in the word relations search interface of PaQu (an extension of the Groningen Word Relations Search application originally developed by (Tjong Kim Sang et al., 2010), as well as via PaQu’s XPATH interface.

In order to evaluate the quality of the automatically generated parses for this research, I selected all transcriptions of adult utterances in the Van Kampen corpus that contain one of the words *heel*, *erg*, or

[2] or maybe the whole VP *verbaast mij*.

*zeer*. The transcriptions contain all kinds of mark-up that Alpino cannot deal with, so I wrote a script to clean out this mark-up. I will provide more details about the cleaning script in the full paper.

## 8 Evaluation

The assignment of parses by the Alpino parser is a fully automated process, and therefore the results will most likely contain errors. As a first small experiment to investigate the quality of the parses produced by Alpino in this domain, I used the manually tagged Van Kampen adult utterances as gold standard to compare the output of Alpino with.[3]

When comparing the Alpino results with the gold standard, two types of errors were found in the manual annotations, which were corrected, resulting in an original and in a revised gold standard.
A striking first finding is that *heel* occurs much more often than *erg*, which occurs much more often than *zeer*. This holds for the words themselves (in the proportion 76% - 17% -7%) and even more for their use as a modifier of A,V, or P (proportion 90% - 9% - 1%). The results of the comparison with the reference annotations are provided in (2). It specifies the accuracy of the Alpino parser in characterizing the words *heel*, *erg* and *zeer* correctly compared to the original gold standard (Acc) and the revised gold standard (RevAcc):

(2)

| **word** | **Acc** | **RevAcc** |
|---|---|---|
| *heel* | 0.94 | 0.95 |
| *erg* | 0.88 | 0.91 |
| *zeer* | 0.21 | 0.21 |

The results for *heel* and *erg* are very good with over 90% accuracy compared to the revised gold standard. The results of *zeer* appear to be very bad. Further analysis reveals that most errors are made for the construction *zeer doen*, lit. *pain do*, 'to hurt', which Alpino really does not know how to analyze. I will discuss this case in more detail in the full paper.

Since the bad results for *zeer* are mainly caused by one type of construction, which can be easily identified in PaQu, as I will show in the full paper, the results of PaQu are still very useful.

I will discuss here in the full paper that care must exercised in genereralizing these results.

I actually also evaluated the performance of Alpino on the children's utterances. Similar figures are found, though the accuracy figures are lower. The relative proportion of *heel*, *erg* and *zeer* shows a distribution similar to the adults utterances, though with an even higher proportion for *heel*: (93% - 5% - 2%) and (96% - 4% - 0%) for their use as a modifier of A,V, or P. For the accuracy figures, see Table (3):

(3)

| **word** | **Acc** |
|---|---|
| *heel* | 0.90 |
| *erg* | 0.73 |
| *zeer* | 0.17 |

Here I will analyze and discuss these results in more detail in the full paper

## 9 Analysis for all Dutch CHILDES corpora

The results of an analysis of the words *heel*, *erg* and *zeer*, based on an automatic parse of all adult utterances in the Dutch CHILDES corpora are given in (4).[4] It specifies, for each of the three words, the counts of their occurrences in specific grammatical roles that concern us here, the counts of their occurrences in other grammatical roles (*other*), and of cases where the grammatical role could not be determined (*unclear*).[5]

[3] If one logs in into the PaQu application, one actually finds the parsed corpora with the cleaned Van Kampen adult sentences, since I shared the corpora with everyone. They are called *VanKampenHeel*, *VanKampenerg*, and *VanKampenZeer*, resp.

[4] I use the following notation in the table: *Mod X* means that the word can modify a word of category X; *predc* stands for *can occur as predicative complement*.

[5] For example, in incomplete or ungrammatical utterances.

(4)

| Results | mod A | mod N | Mod V | mod P | predc | other | unclear | Total |
|---|---|---|---|---|---|---|---|---|
| *heel* | 881 | 51 | 2 | 2 | 14 | 0 | 2 | **952** |
| *erg* | 347 | 27 | 109 | 0 | 187 | 5 | 0 | **675** |
| *zeer* | 7 | 1 | 83 | 0 | 19 | 21 | 7 | **138** |

The proportion of *heel*, *erg* and *zeer* shows a similar distribution as in the Van Kampen subcorpus, though the frequency of *erg* is much higher than in the Van Kampen Corpus: 54% - 38% - 8%. In their use as a modifier of A,V, or P the proportions are 65% - 34% - 1%.[6]

Most striking in these data is the overwhelming number of cases where *heel* modifies an adjective. This covers over 92% of the examples with *heel* found. Modification of V and P by *heel* hardly occurs, and in fact the four examples all involve wrong parses. The mod V cases actually involve adjectives, as will be shown in the full paper. The Mod P examples involve wrong parses as well.

A second observation is that there are quite some examples in which *erg* modifies an adjective or a verb.

A third observation is that there are very few examples involving *zeer* modifying an adjective. In only 6 out of the 83 examples of Mod V, *zeer* indeed modifies a verb. In one case it modifies an adjective. All other examples where it modifies V according to Alpino are in fact wrong parses involving *zeer doen* 'to hurt', discussed above). The scarcity of the examples of *zeer* as a modifier of A,V, or P can plausibly be attributed to its more formal pragmatic nature, so that it will be less used in spoken parent - young child interactions.

The table suggests that there are no examples of *erg* and *zeer* modifying prepositional phrases at all. In fact, there are a few (4 occurrences), but they involve idiomatic expressions such as *op prijs stellen*[7] (modified by *erg* once and by *zeer* twice) and *in de smaak vallen*[8] (modified by *zeer* once) in which Alpino has analyzed them as modifying the verb.

## 10 Conclusions

We can draw two types of conclusions from the work presented in this paper: conclusions with regard to the linguistic problem, and conclusions with regard to PaQu as a research tool.

Starting with the linguistics, any conclusions here must be very preliminary, given the small scale of the research done here. Nevertheless, the observations made in the preceding section are suggestive of further research. For example, they suggest that the overwhelmingness of the occurrence of *heel* as a modifier of an adjective in comparison to its occurrence as a modifier of a verb (881 v. 2) might play a role in fixing the modification potential of this word to adjectives. In contrast, the occurrences of the word *erg* as a modifier of adjectives and verbs are more balanced: 347 v. 129.

The fact that there are hardly any examples for *zeer* make it difficult to draw any conclusions. This most probably means that the current CHILDES samples are insufficiently large as a sample of first language acquisition.[9]

Turning to the research tool PaQu, it can be safely concluded from this paper that PaQu is very useful for aiding researchers in better and more efficient manual verification of hypotheses. Because of its fully automated nature, it applies blindly and automatically and is in this respect usually more consistent than humans (who err). But of course, the parses generated by PaQu (via Alpino) are fully automatically generated and will contain errors. Nevertheless, as shown in this paper, in some cases, its fully automatically generated parses and their statistics can reliably be used directly (though care is required!), and one frequently occurring error described in this paper turned out to be so systematic that the relevant examples can be easily identified using PaQu queries.

[6] These figures are based on reassignments for wrong parses of *heel* and *zeer*, see below.

[7] Lit. at price set, 'appreciate'

[8] lit. in the taste fall, 'like' (with arguments reversed).

[9] A rough count shows that the Dutch CHILDES corpora dealt with here contain 534 k utterances and approx. 2.9 million inflected word form occurrences ('tokens').

## 11 Future Work

It is obvious that the work reported on in this paper is just the beginning. There is a lot of work that can (and should) be done in the near future. I will describe this in the full paper. Most of the possible future work mentioned here is actually planned in the CLARIAH-CORE project or in the Utrecht University project *AnnCor*.

## References

[MacWhinney2015] Brian MacWhinney. 2015. Tools for analyzing talk, electronic edition, part 1: The CHAT transcription format. Technical report, Carnegie Mellon University, Pittsburg, PA, April27. `http://childes.psy.cmu.edu/manuals/CHAT.pdf`.

[Odijk2011] Jan Odijk. 2011. User scenario search. internal CLARIN-NL document, `http://www.clarin.nl/sites/default/files/User%20scenario%20Serach%20110413.docx`, April 13.

[Odijk2014] Jan Odijk. 2014. CLARIN: What's in it for linguists?, March 27. Uilendag Lecture, Utrecht, `http://dspace.library.uu.nl/handle/1874/295277`.

[Tjong Kim Sang et al.2010] Erik Tjong Kim Sang, Gosse Bouma, and Gertjan van Noord. 2010. LASSY for beginners. Presentation at CLIN 2010, February 5.

# POLFIE: an LFG grammar of Polish accompanied by a structure bank

**Agnieszka Patejuk and Adam Przepiórkowski**
Institute of Computer Science, Polish Academy of Sciences,
ul. Jana Kazimierza 5, 01-248 Warszawa, Poland
{aep,adamp}@ipipan.waw.pl

## Abstract

This paper presents POLFIE, an LFG grammar of Polish, by discussing the resources it builds upon, the ways in which the grammar is being made available to the public and the LFG structure bank of Polish created using the grammar.

## 1 Introduction

POLFIE is a grammar of Polish written in the framework of LFG,[1] implemented in the XLE[2] platform. It produces a syntactic representation of sentences which consists of 2 parallel structures: c(onstituent)-structure (tree) and f(unctional)-structure (attribute-value matrix).

POLFIE is designed as a resource which maximises the use of existing resources for Polish: the grammar was written on the basis of two other implemented formal grammars of Polish, it uses converted valence information from Walenty and it uses Morfeusz2 as the morphological analyser (see below for references to these resources).

POLFIE is made available in two ways: as a stand-alone package which requires the local installation of XLE and as a web service using XLE-Web, a part of INESS system (`http://iness.uib.no/`; Rosén et al. 2007), which makes it possible to use the grammar without a local installation of XLE.

The coverage of the LFG grammar as well as the quality of the analyses it offers is evaluated by building an LFG structure bank.

Finally, structures created by POLFIE serve as input for semantic processing (not discussed here).

## 2 Resources used

Creating an LFG grammar consists of two major tasks: writing the rules and building the lexicon. In case of POLFIE, both tasks are accomplished by using existing resources for Polish.

The rules of the current grammar were written on the basis of two previous implemented grammars of Polish: GFJP2, a DCG[3] (based on Świdziński 1992) used by the Świgra parser (`http://zil.ipipan.waw.pl/Świgra/`; Woliński 2004), which provided a basis for c-structure rules of POLFIE, and FOJP, a small-scale but linguistically very sophisticated HPSG[4] grammar. C-structure rules of the current LFG grammar are based on context-free grammar rules of GFJP2 – these rules were annotated with instructions on how to build an additional level of structure on top of trees, namely the f-structure. This provides a representation employing grammatical functions, which is considered more universal across languages than c-structure, which is subject to much variation (such as word order, for instance). The f-structure annotation was inspired by both grammars: while the analysis of many linguistic phenomena offered by the original DCG grammar could often be translated into the f-structure representation almost unchanged, there are, however, some significant differences, especially in the area of agreement, case assignment and negation, where the LFG analysis draws broadly from existing HPSG analyses of these phenomena (Przepiórkowski, 1999; Przepiórkowski et al., 2002). The resulting LFG

[1] *Lexical Functional Grammar* (Bresnan, 1982; Bresnan, 2000; Dalrymple, 2001).

[2] *Xerox Linguistic Environment* (`http://www2.parc.com/isl/groups/nltt/xle/`; Crouch et al. 2011).

[3] *Definite Clause Grammar* (Warren and Pereira, 1980).

[4] *Head-driven Phrase Structure Grammar* (Pollard and Sag, 1987; Pollard and Sag, 1994).

grammar extends the original DCG grammar vertically, by adding the level of f-structure to the c-structure offered by the DCG grammar, and horizontally, by covering a wider range of phenomena.

The next step is the creation of the lexicon which provides information about morphosyntax and valence. Morphosyntactic information can be taken from a variety of sources, including XML files from the National Corpus of Polish (NKJP; `http://nkjp.pl/`; Przepiórkowski et al. 2010; Przepiórkowski et al. 2012) and Składnica (`http://zil.ipipan.waw.pl/Składnica/`; Woliński et al. 2011), a treebank of parses created using Świgra – it is worth noting that the information from these resources is disambiguated (manually or automatically). An alternative is to use Morfeusz, the state-of-the-art morphological analyser for Polish (`http://sgjp.pl/morfeusz/`; Woliński 2006; Saloni et al. 2012; Woliński 2014) – while its output is undisambiguated, it is very convenient for interactive parsing: a grammar library transducer is used to convert the output of Morfeusz. Valence information is taken from Walenty (`http://zil.ipipan.waw.pl/Walenty/`; Przepiórkowski et al. 2014b; Przepiórkowski et al. 2014a), a new valence dictionary for Polish built within CLARIN-PL project, which has many features which make it particularly useful for the LFG grammar: it distinguishes the subject and object grammatical functions, it explicitly accounts for coordination within one argument (unlike category coordination) and it provides valence schemata not only for verbs, but also for nominals, adjectives and adverbs; finally, it takes phraseological information into account. Information from Walenty is converted into XLE/LFG constraints using a dedicated script, which is in turn used when creating the lexicon.

## 3 Availability

The grammar is open source and it is made available according to the terms of GNU General Public License version 3 (`http://www.gnu.org/licenses/gpl-3.0.en.html`).

The grammar can be obtained at `http://zil.ipipan.waw.pl/LFG/` as a package intended for use with Linux distributions (32- and 64-bit). It requires a local installation of XLE (not provided in the package) as well as Morfeusz and Walenty (both bundled in the package).

There is, however, a convenient alternative to the local installation – it is the web service version of the grammar, made available at `http://iness.mozart.ipipan.waw.pl/iness/xle-web` via XLE-Web, a web interface to XLE, which is a part of INESS system. The XLE-Web version does not require a local installation of XLE, it is immediately available via a user-friendly interface and it does not require login – it can be accessed using a variety of web browsers. The grammar available via XLE-Web uses Morfeusz2 with a grammar library transducer, so in principle it can analyse any segment known to Morfeusz2. Two versions of the grammar are available for use: `POLFIE-Morfeusz2`, which is the current stable version of the grammar, and `POLFIE-Morfeusz2-OT`, which uses OT (*Optimality Theory*) mechanisms to disambiguate the output of POLFIE by selecting the most optimal solution according to the so-called OT marks placed in the grammar. In both versions the user can disambiguate the output using discriminants.

## 4 Quality control

The evaluation of the LFG grammar of Polish is performed against three independent measures: constructed testsuites, authentic sentences from the treebank (Składnica), and authentic sentences from the corpus (NKJP1M, a manually annotated subcorpus of NKJP containing 1.2 million segments). The aim of testing using constructed examples is to ensure that the grammar correctly models particular linguistic phenomena. Testing based on sentences from the treebank checks the level of compatibility with the grammar which provided the original c-structure (GFJP2). Finally, testing on sentences from the corpus checks the grammar for robustness and real-life coverage (currently 33% of sentences have a parse).

## 5 LFG structure bank

An LFG structure bank is currently being created with the help of the INESS infrastructure for building structure banks (Rosén et al., 2007). It is based on Składnica in the sense that sentences from Składnica are reparsed using the LFG grammar with a lexicon created from disambiguated morphosyntactic interpretations taken from Składnica (which in turn were taken from NKJP1M, the manually annotated subcorpus of NKJP) and valence information provided by schemata from Walenty (converted to XLE/LFG constraints). It contains almost 6500 sentences (with a parse), which have been manually disambiguated.

Sentences parsed with XLE are uploaded to INESS, where they are disambiguated by human annotators: each sentence is disambiguated independently by two people who cannot see each other's solution nor comments. Annotators can, however, communicate with each other and the grammar writer using the dedicated mailing list to discuss issues related to disambiguation.

Annotators perform the disambiguation by choosing discriminants which may apply to different parts of structures: there are lexical, c-structure and f-structure discriminants. Annotators of Polish LFG structure bank are asked to choose f-structure discriminants whenever possible as these are considered less likely to change across grammar versions.

When there is no good analysis or the analysis offered could be improved, annotators write comments pointing to the problematic fragment and sometimes also identify the problem at hand. Comments report problems related to the grammar (e.g. some coordinate structures are not accounted for), to the valence schemata (a missing argument – a problem with Walenty; an argument not classified properly – usually an issue with Walenty, sometimes with the conversion) and to the morphosyntactic annotation (wrong case marking, wrong choice of part of speech), making it possible to correct these problems at the source. Currently, there are almost 3000 comments.

After a round of annotation is completed, comments created by annotators are inspected by the grammar writer, who responds to each of them using the mailing list (after they have been anonymised) – is the comment right (sometimes explaining what is happening in the relevant case) or is it wrong (explaining why it is wrong). The purpose of this review is to give feedback to annotators (improving their skills by making them aware of certain linguistic issues, encouraging them to write comments) and to improve the grammar as well as the valence dictionary. Comments are anonymised to avoid discouraging people who might fear that their comment could be wrong.

Subsequently relevant comments containing confirmed issues are passed together with responses (and additional comments, if needed) to the developers of relevant resources. Developers of Walenty are asked to inspect relevant entries and introduce appropriate changes if need be. Issues related to the conversion are handled by the grammar writer. Finally, comments related to problems in the grammar are collected and passed to the grammar writer to introduce appropriate modifications to improve the treatment of relevant phenomena.

After relevant changes have been introduced in Walenty and the grammar, a new lexicon is created, sentences are reparsed and a new version of parses is added to INESS so that discriminants can be reapplied from the previous disambiguated version of the structure bank. Discriminant choices are reapplied only if the relevant discriminant choice is still available in the new version. It is a very useful feature of INESS since it makes it possible to maximally reuse previous disambiguation work rather than start from scratch. After discriminants have been reapplied, annotators are asked to return to sentences from their annotation set which did not have a complete good solution in the previous version, return to their comments and check if the relevant problem has been solved in the current version.

The LFG structure bank is going to be released officially this year (on an open source licence), making it possible to view the chosen parses and search the entire structure bank.

## 6 Conclusion

This paper presented how POLFIE, an LFG grammar of Polish, it is being developed and made available to the public. It also discussed the LFG structure bank of Polish constructed using INESS, which is being created parallel to the development of the grammar. Both resources have been presented to potential users during CLARIN-PL workshops and there was considerable interest, which was reflected in comments.

## References

[Bresnan1982] Joan Bresnan, editor. 1982. *The Mental Representation of Grammatical Relations*. MIT Press Series on Cognitive Theory and Mental Representation. The MIT Press, Cambridge, MA.

[Bresnan2000] Joan Bresnan. 2000. *Lexical-Functional Syntax*. Blackwell Textbooks in Linguistics. Blackwell.

[Calzolari et al.2014] Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors. 2014. *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*, Reykjavík, Iceland. ELRA.

[Crouch et al.2011] Dick Crouch, Mary Dalrymple, Ron Kaplan, Tracy King, John Maxwell, and Paula Newman. 2011. XLE documentation. `http://www2.parc.com/isl/groups/nltt/xle/doc/xle_toc.html`.

[Dalrymple2001] Mary Dalrymple. 2001. *Lexical Functional Grammar*. Academic Press.

[Pollard and Sag1987] Carl Pollard and Ivan A. Sag. 1987. *Information-Based Syntax and Semantics, Volume 1: Fundamentals*. Number 13 in CSLI Lecture Notes. CSLI Publications, Stanford, CA.

[Pollard and Sag1994] Carl Pollard and Ivan A. Sag. 1994. *Head-driven Phrase Structure Grammar*. Chicago University Press / CSLI Publications, Chicago, IL.

[Przepiórkowski1999] Adam Przepiórkowski. 1999. *Case Assignment and the Complement-Adjunct Dichotomy: A Non-Configurational Constraint-Based Approach*. Ph.D. dissertation, Universität Tübingen, Germany.

[Przepiórkowski et al.2002] Adam Przepiórkowski, Anna Kupść, Małgorzata Marciniak, and Agnieszka Mykowiecka. 2002. *Formalny opis języka polskiego: Teoria i implementacja*. Akademicka Oficyna Wydawnicza EXIT, Warsaw.

[Przepiórkowski et al.2010] Adam Przepiórkowski, Rafał L. Górski, Marek Łaziński, and Piotr Pęzik. 2010. Recent developments in the National Corpus of Polish. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC 2010*, Valletta, Malta. ELRA.

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

[Przepiórkowski et al.2014a] Adam Przepiórkowski, Elżbieta Hajnicz, Agnieszka Patejuk, Marcin Woliński, Filip Skwarski, and Marek Świdziński. 2014a. Walenty: Towards a comprehensive valence dictionary of Polish. In Calzolari et al. (Calzolari et al., 2014), pages 2785–2792.

[Przepiórkowski et al.2014b] Adam Przepiórkowski, Filip Skwarski, Elżbieta Hajnicz, Agnieszka Patejuk, Marek Świdziński, and Marcin Woliński. 2014b. Modelowanie własności składniowych czasowników w nowym słowniku walencyjnym języka polskiego. *Polonica*, XXXIII:159–178.

[Rosén et al.2007] Victoria Rosén, Paul Meurer, and Koenraad De Smedt. 2007. Designing and implementing discriminants for LFG grammars. In Miriam Butt and Tracy Holloway King, editors, *The Proceedings of the LFG'07 Conference*, pages 397–417, Stanford, CA. CSLI Publications.

[Saloni et al.2012] Zygmunt Saloni, Marcin Woliński, Robert Wołosz, Włodzimierz Gruszczyński, and Danuta Skowrońska. 2012. *Słownik gramatyczny języka polskiego*. Warsaw, 2nd edition.

[Świdziński1992] Marek Świdziński. 1992. *Gramatyka formalna języka polskiego*, volume 349 of *Rozprawy Uniwersytetu Warszawskiego*. Wydawnictwa Uniwersytetu Warszawskiego, Warsaw.

[Warren and Pereira1980] D. H. D. Warren and Fernando C. N. Pereira. 1980. Definite clause grammars for language analysis — a survey of the formalism and a comparison with augmented transition networks. *Artificial Intelligence*, 13:231–278.

[Woliński2004] Marcin Woliński. 2004. *Komputerowa weryfikacja gramatyki Świdzińskiego*. Ph.D. dissertation, Institute of Computer Science, Polish Academy of Sciences, Warsaw.

[Woliński2006] Marcin Woliński. 2006. Morfeusz — a Practical Tool for the Morphological Analysis of Polish. In Mieczysław Kłopotek, Sławomir T. Wierzchoń, and Krzysztof Trojanowski, editors, *Intelligent information processing and web mining*, pages 503–512. Springer-Verlag.

[Woliński2014] Marcin Woliński. 2014. Morfeusz reloaded. In Calzolari et al. (Calzolari et al., 2014), pages 1106–1111.

[Woliński et al.2011] Marcin Woliński, Katarzyna Głowińska, and Marek Świdziński. 2011. A preliminary version of Składnica—a treebank of Polish. In Zygmunt Vetulani, editor, *Proceedings of the 5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, pages 299–303, Poznań, Poland.

# PICCL: Philosophical Integrator of Computational and Corpus Libraries

**Martin Reynaert**[1 2] **Maarten van Gompel**[2] **Ko van der Sloot**[2] **Antal van den Bosch**[2]
TiCC / Tilburg University[1] CLST / Radboud University Nijmegen[2]
The Netherlands
mreynaert|M.vanGompel|K.vanderSloot|a.vandenbosch@let.ru.nl

## Abstract

CLARIN activities in the Netherlands in 2015 are in transition between the first national project CLARIN-NL and its successor CLARIAH. In this paper we give an overview of important infrastructure developments which have taken place throughout the first and which are taken to a further level in the second. We show how relatively small accomplishments in particular projects enable larger steps in further ones and how the synergy of these projects helps the national infrastructure to outgrow mere demonstrators and to move towards mature production systems. The paper centers around a new corpus building tool called PICCL. This integrated pipeline offers a comprehensive range of conversion facilities for legacy electronic text formats, Optical Character Recognition for text images, automatic text correction and normalization, linguistic annotation, and preparation for corpus exploration and exploitation environments. We give a concise overview of PICCL's components, integrated now or to be incorporated in the foreseeable future.

## 1 Introduction

The transition from CLARIN-NL (Odijk, 2010) to its successor CLARIAH[1] offers an opportunity to assess past and future CLARIN activities in the Netherlands. We give an overview of text infrastructure developments which have taken place throughout the first and are set to be taken to a further level in the second. The demonstrator built in CLARIN-NL Call 4 project @PhilosTEI is now in CLARIAH to grow into the full-fledged production system PICCL. This acronym stands for 'Philosophical Integrator of Computational and Corpus Libraries', the first term of which signifies 'well-considered' and is hoped for its users to grow to mean 'practical'.

## 2 PICCL: an overview

PICCL constitutes a complete workflow for corpus building. It is to be the integrated result of developments in the CLARIN-NL project @PhilosTEI, which ended November 2014, further work in NWO 'Groot' project Nederlab[2], which continues up to 2018, and in CLARIAH, which runs until 2019.

### 2.1 What went before

At Tilburg University, the Netherlands, work was started on building web applications and services for the CLARIN infrastructure in 2011 in project CLARIN-NL TICCLops (Reynaert, 2014b). In this CLARIN-NL Call 1 project the idea to provide text normalization and spelling/OCR post-correction facilities as an 'online processing service' – hence the -ops in the project name – spawned the idea of building a generic system for turning linguistic command-line applications into RESTful web services and web applications. This begat CLAM, the Computational Linguistics Application Mediator (van Gompel and Reynaert, 2014), which TICCLops builds upon. CLAM[3] has been adopted widely within the CLARIN-NL community and underlays the Dutch-Flemish cooperation in the CLARIN-NL infrastructure project

[1] http://www.clariah.nl/en/

[2] https://www.nederlab.nl/onderzoeksportaal/

[3] https://proycon.github.io/clam

TTNWW, in which available tools for both text as well as speech are turned into web services and subsequently united in a workflow management system (Kemps-Snijders et al., 2012).

The storage and exchange of linguistically annotated resources requires a modern and expressive format capable of encoding a wide variety of linguistic annotations. A solution has been devised in the form of FoLiA, short for "Format for Linguistic Annotation" (van Gompel and Reynaert, 2013). FoLiA provides a generic single-solution XML format for a wide variety of linguistic annotations, including lemmata, part-of-speech tags, named-entity labels, shallow and deep syntactic structure, spelling and OCR variation, etc. Furthermore, it provides an ever-expanding software infrastructure to work with the format. The format was adopted by the large corpus building effort for Dutch, the SoNaR project (Oostdijk et al., 2013) in the STEVIN programme, as well as other other projects. In order to provide uniform linguistic annotations for this 540 million word token reference corpus of contemporary, written Dutch, Frog[4] (Van den Bosch et al., 2007), a suite of various natural language processing tools for Dutch based on the TIMBL classifier (Daelemans et al., ), was further developed.

### 2.2 PICCL: system overview

PICCL aims to provide its users with the means to convert their textual research data into an easily accessible, researchable corpus in a format fit for the future. A schematic overview of the PICCL pipeline and some of the planned extensions is shown in Figure 2.2.

Input can be either images or text. Images may be e.g. the scanned pages of a book in DjVu, PDF or TIFF formats. Text images are converted into electronic text by Tesseract[5]. Text may be plain, in various word-processing formats, embedded in PDF, or in OCR engine output formats; i.e. hOCR HTML, Page XML, or Alto XML. Alto XML is the major text format in the large digital text collections aggregated by the Dutch National Library (KB). The conversion tool FoLiA-alto developed for the Nederlab project allows for direct harvesting from the KB. To overcome the severe acceptable input format limitations of (Jongejan, 2013), PICCL is to be equipped with convertors for a panoply of document formats. We intend to incorporate OpenConvert[6], another CLARIN-NL web service. FoLiA XML is PICCL's pivot format. The workflow can handle texts in a broad range of –currently– European languages. Provisions are available for dealing with old print or diachronical language variation.

Output text is in FoLiA XML[7]. The pipeline will therefore offer the various software tools that support FoLiA. Language categorization may be performed by the tool FoLiA-langcat at the paragraph level. TICCL or 'Text-Induced Corpus Clean-up' performs automatic post-correction of the OCRed text. Dutch texts may optionally be annotated automatically by Frog, i.e. tokenized, lemmatized and classified for parts of speech, named entities and dependency relations. The FoLiA Linguistic Annotation Tool (FLAT)[8] will provide for manual annotation of e.g. metadata elements within the text – for later extraction. FoLiA-stats delivers $n$-gram frequency lists for the texts' word forms, lemmata, and parts of speech. Colibri Core[9] allows for more efficient pattern extraction, on text only, and furthermore can index the text, allowing comparisons to be made between patterns in different (sub)corpora. BlackLab[10] and front-end WhiteLab[11], developed in the OpenSoNaR project[12] (Reynaert et al., 2014), allow for corpus indexing and querying. Convertors to other formats, e.g. TEI XML, for allowing scholars to build critical editions of books, will be at hand.

PICCL is to be available to all researchers in the CLARIN infrastructure and is hosted by certified CLARIN Centre INL in Leiden. PICCL is to have a highly intuitive user-friendly interface in order to allow even the most computer-weary user to obtain texts in a corpus-ready, annotated format. Its predecessor, the @PhilosTEI system, provides two distinct interfaces: the more generic interface type

[4] http://ilk.uvt.nl/frog

[5] https://github.com/tesseract-ocr

[6] https://github.com/INL/OpenConvert

[7] https://proycon.github.io/folia

[8] https://github.com/proycon/flat

[9] https://proycon.github.io/colibri-core

[10] https://github.com/INL/BlackLab/wiki

[11] https://github.com/TiCCSoftware/WhiteLab

[12] http://opensonar.clarin.inl.nl

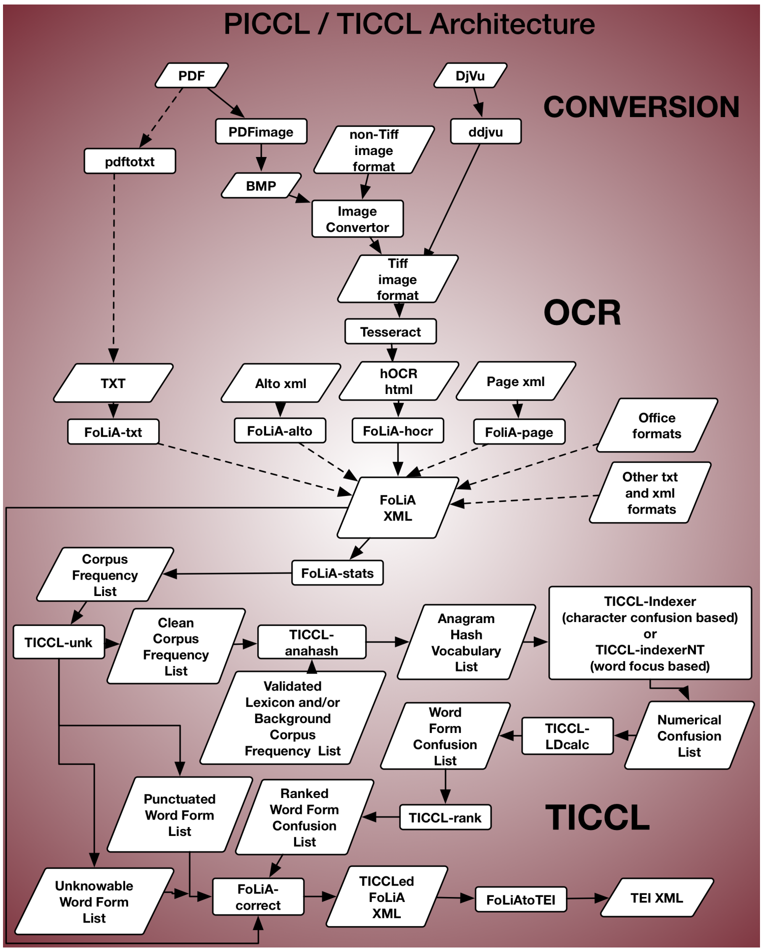


Figure 1: A schematic overview of the PICCL pipeline, dotted lines signify future extensions

that comes with CLAM [13] as well as a more end-user-oriented web interface that was custom made for the @PhilosTEI project, according to the specifications of the end users, i.e. philosophers[14].

PICCL continues to be developed with the users firmly in mind. We aim to make the user-friendly system available as a large black box that processes a book's images into a digital version with next to no user intervention or prior knowledge required. At the same time we also want to equip PICCL with the necessary interface options to allow more sophisticated users to address any sub-module or combination of sub-modules individually at will.

Future developments in CLARIAH are that Frog is to be made more easily retrainable, e.g. for older varieties of Dutch. It is also to be trained for both present-day English and German.

### 2.3 PICCL in comparison

In contrast to the CLARIN-NL TTNWW workflow (Kemps-Snijders et al., 2012) in the Taverna[15] framework, PICCL is implemented as a single and efficient pipeline, rather than a collection of many interconnected webservices. PICCL's focus is on the end-user who has an interest in the pipeline as a whole rather than its individual parts. This approach avoids network overhead, which can be a significant bottleneck in dealing with large corpus data. It still allows for distributional use of the available hardware through load-balancing, and still allows for the whole to be available as a RESTful webservice, through CLAM, for

[13] http://ticclops.clarin.inl.nl
[14] http://philostei.clarin.inl.nl
[15] http://www.taverna.org.uk

automated connectivity. Another major difference between TTNWW and PICCL is that the latter allows for better control over and handling of the text document flow. A TTNWW workflow offers sequential handling of documents by the various web services only, i.e. every single input file is processed in turn by each service and passed on to the next. In contrast, the PICCL wrapper allows for flexible handling of numbers of input/output files, taking e.g. $x$ PDF input files apart into $y$ (where $y \geq x$) image files to be sent to the OCR engine Tesseract, then presenting the $y$ OCRed files as a single batch to TICCL which eventually corrects the $y$ FoLiA XML files to be collated into a single output FoLiA XML and also, if the user so desires, a TEI XML output e-book.

Another solution for NLP workflows is provided by the Weblicht project (Hinrichs et al., 2010), developed in the scope of CLARIN-D. Both Taverna and Weblicht are generic workflow frameworks and are more suited for a larger number of interconnected webservices. PICCL is a single, more monolithic, workflow, albeit heavily parametrised. Weblicht also centers around their own TCF file format whilst our solutions are deliberately FoLiA-based because it can better express spelling correction and lexical normalization.

## 3 TICCL: an overview

A major component of the PICCL pipeline is Text-Induced Corpus Clean-up or TICCL, a system for unsupervised spelling correction and lexical normalisation or post-correction of OCRed corpora. TICCL is now multilingual and diachronic. In contrast to e.g. Vobl et al. (2014), TICCL aims at fully automatic post-correction. It was shown to outperform VARD2 (Baron and Rayson, 2008) in Reynaert et al. (2012) in the task of spelling normalization of historical Portuguese.

### 3.1 TICCL: current implementation

TICCL currently consists of a wrapper (written in Perl) around efficient multithreaded modules (in C++). Two of these modules, respectively the first and last of the TICCL pipeline, work on and require FoLiA XML input. The intermediate modules are TICCL-specific, and do not work on running text but rather on lists containing either words and frequency information or anagram hash values derived from corpus and lexicon words. The main publication on how TICCL operates is (Reynaert, 2010). Reynaert (2014a) offers more details on its current implementation and an in-depth evaluation on historical Dutch. This shows that when equipped with the most comprehensive historical lexicons and name lists, as well as with the word frequency information derived from a large background corpus of contemporary books, TICCL achieves high precision and useful recall. After fully automated correction, the word accuracy of the gold standard book experimentally corrected was raised from about 75% to 95%.

### 3.2 TICCL and language-specific lexical resources

TICCL relies on word and $n$-gram frequencies derived from the corpus to be cleaned. It can also be provided with further word form frequencies derived from e.g. another – possibly very large – background corpus. For languages other than Dutch the system is currently equipped with open source lexicons only. More importantly, PICCL will allow its users to equip the system with their own lexical resources of choice through the simple expedient of uploading them.

## 4 Conclusion

We have given an overview of work delivered and ongoing on PICCL, a comprehensive corpus building work flow. The system is geared to be equipped with the best available solutions for the sub-problems it is meant to solve. It is highly user-friendly, shielding the user to the highest extent from the intricacies of the many software modules it is composed of, asking only for the most minimal user input possible. The Nederlab project as prime user of the system is set to 'piccl' a great many diachronic corpora of Dutch. We hope PICCL will enable anyone to build their own personal text corpora and to derive the utmost benefit from them.

## Acknowledgements

The authors acknowledge having been funded by several projects in CLARIN-NL and to be funded by the Dutch national CLARIN project CLARIAH. Martin Reynaert further acknowledges being funded by NWO in the project Nederlab.

## References

Alistair Baron and Paul Rayson. 2008. VARD2: A tool for dealing with spelling variation in historical corpora. In *Proceedings of the Postgraduate Conference in Corpus Linguistics*.

Walter Daelemans, Jakub Zavrel, Ko Van der Sloot, and Antal Van den Bosch. TiMBL: Tilburg Memory Based Learner, version 6.3, Reference Guide, year = 2010. Technical Report ILK 10-01, ILK Research Group, Tilburg University.

Marie Hinrichs, Thomas Zastrow, and Erhard W. Hinrichs. 2010. Weblicht: Web-based LRT Services in a Distributed eScience Infrastructure. In Nicoletta et al. Calzolari, editor, *LREC*. European Language Resources Association.

Bart Jongejan. 2013. Workflow Management in CLARIN-DK. In *Proceedings of the workshop on Nordic language research infrastructure at NODALIDA 2013*, volume 089 of *NEALT*, pages 11–20.

Marc Kemps-Snijders, Matthijs Brouwer, Jan Pieter Kunst, and Tom Visser. 2012. Dynamic web service deployment in a cloud environment. In Nicoletta Calzolari et al., editor, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 2941–2944, Istanbul, Turkey. ELRA.

Jan Odijk. 2010. The CLARIN-NL project. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC-2010*, pages 48–53, Valletta, Malta.

Nelleke Oostdijk, Martin Reynaert, Véronique Hoste, and Ineke Schuurman. 2013. The construction of a 500-million-word reference corpus of contemporary written Dutch. In *Essential Speech and Language Technology for Dutch: Results by the STEVIN-programme*, chapter 13. Springer Verlag.

Martin Reynaert, Iris Hendrickx, and Rita Marquilhas. 2012. Historical spelling normalization. A comparison of two statistical methods: TICCL and VARD2. In Francesco Mambrini, Marco Passarotti, and Caroline Sporleder, editors, *Proceedings of ACRH-2*, pages 87–98. Lisbon: Colibri.

Martin Reynaert, Matje van de Camp, and Menno van Zaanen. 2014. OpenSoNaR: user-driven development of the SoNaR corpus interfaces. In *Proceedings of COLING 2014: System Demonstrations*, pages 124–128, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Martin Reynaert. 2010. Character confusion versus focus word-based correction of spelling and OCR variants in corpora. *International Journal on Document Analysis and Recognition*, 14:173–187.

Martin Reynaert. 2014a. Synergy of Nederlab and @PhilosTEI: diachronic and multilingual Text-Induced Corpus Clean-up. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. ELRA.

Martin Reynaert. 2014b. TICCLops: Text-Induced Corpus Clean-up as online processing system. In *Proceedings of COLING 2014: System Demonstrations*, pages 52–56, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Antal Van den Bosch, Gertjan Busser, Sander Canisius, and Walter Daelemans. 2007. An efficient memory-based morpho-syntactic tagger and parser for Dutch. In P. Dirix et al., editor, *Computational Linguistics in the Netherlands: Selected Papers from the Seventeenth CLIN Meeting*, pages 99–114, Leuven, Belgium.

Maarten van Gompel and Martin Reynaert. 2013. FoLiA: A practical XML Format for Linguistic Annotation - a descriptive and comparative study. *Computational Linguistics in the Netherlands Journal*, 3.

Maarten van Gompel and Martin Reynaert. 2014. CLAM: Quickly deploy NLP command-line tools on the web. In *Proceedings of COLING 2014: System Demonstrations*, pages 71–75, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Thorsten Vobl, Annette Gotscharek, Ulrich Reffle, Christoph Ringlstetter, and Klaus Schulz. 2014. PoCoTo - An Open Source System for Efficient Interactive Postcorrection of OCRed Historical Texts. In *Proceedings of Datech 2014*. ACM.

# CLARIN Concept Registry: the new semantic registry

**Ineke Schuurman**
Utrecht University
University of Leuven
ineke@ccl.kuleuven.be

**Menzo Windhouwer**
Meertens Institute
menzo.windhouwer@
meertens.knaw.nl

**Oddrun Ohren**
National Library of
Norway
oddrun.ohren@nb.no

**Daniel Zeman**
Charles University
in Prague
zeman@
ufal.mff.cuni.cz

## 1 Introduction

One of the foundations of the CLARIN Component Metadata Infrastructure (CMDI; Broeder et al. 2012; clarin.eu/cmdi) is a semantic layer (Durco and Windhouwer, 2013) formed by references from CMDI components or elements to entries in various semantic registries. Popular have been references to the metadata terms provided by the Dublin Core Metadata Initiative (DCMI; dublincore.org) and the data categories provided by ISO Technical Committee 37's Data Category Registry (DCR; ISO 12620, 2009), ISOcat (isocat.org). Although using ISOcat has been encouraged by CLARIN, it has its drawbacks. As pointed out by Broeder et al. (2014) and Wright et al. (2014), ISOcat, with its rich data model combined with a very open update strategy has proved too demanding, at least for use in the CLARIN context. Among other things, confusion on how to judge whether a candidate ISOcat entry adequately represents the semantics of some CMDI component or element, has led to proliferation far beyond the real need, resulting in a semantic layer of questionable quality. Therefore, when ISOcat, due to strategic choices made by its Registration Authority, had to be migrated and became, for the time being, static, CLARIN decided to look for other solutions to satisfy the needs of the infrastructure. As a result the Meertens Institute is now hosting and maintaining this new CLARIN semantic registry.

This paper motivates and describes the new semantic registry, the CLARIN Concept Registry (CCR), its model, content and access regime, indicating the differences from ISOcat where appropriate. Proposed management procedures for CCR are also outlined, although not in detail.[1]

## 2 An OpenSKOS registry

In CLARIN-NL the Meertens Institute had already developed (and continues hosting) the CLAVAS vocabulary service based on the open source OpenSKOS software package (Brugman and Lindeman, 2012; openskos.org), which was originally created in the Dutch CATCHPlus project. The OpenSKOS software provides an API to access, create and share thesauri and/or vocabularies, and also provides a web-based editor for most of these tasks. The software is used by various Dutch cultural heritage institutes. The Meertens Institute joined them to collectively maintain and further develop the software.

Based on the experiences with ISOcat OpenSKOS was evaluated to see if it would meet the needs of the CLARIN community and infrastructure. The major aim was to improve the quality of the concepts by having a much simpler data model and a less open, and also less complicated, procedure for adding new concepts or changing existing ones and recommending them to the community. In addition, certain technological requirements of the CLARIN infrastructure had to be met. Based on this evaluation the Meertens Institute extended the OpenSKOS software in various ways:

- Concepts in the CCR get a handle as their Persistent IDentifier (PID);
- The CCR can easily be accessed by the CLARIN community via a faceted browser;
- Support for SKOS collections;
- Shibboleth-based access to the CCR.

Currently these extensions reside in a private Meertens Institute source code repository, but as part of the CLARIN-PLUS project these extensions (and more) will be integrated with the next version of OpenSKOS now under development.

## 3 Representing CCR concepts in the SKOS model

The data model supported by OpenSKOS is a substantial part of the Simple Knowledge Organisation Scheme (SKOS) recommendation by W3C (w3.org/skos). SKOS is typically used to represent thesauri, taxonomies and other knowledge organisation systems. At the Meertens Institute support for collections was added and currently Picturae, a Dutch service provider within the cultural heritage world and the original developer of OpenSKOS, works on supporting the extended labels of SKOS-XL.

The work done by the CLARIN community in ISOcat was made available in the CCR by importing selected sets of data categories as new concepts (see Section 4). This made it possible to start a round of clean-up and creating a coherent set of recommended concepts (see Section 5). This import is not lossless as data category specific properties like the data category type and data type are lost. However, these properties have turned out to be one of the main causes of confusion and proliferation in the use of ISOcat (Broeder et al., 2014; Wright et al., 2014). In general SKOS appears to be a suitable model for the CCR. Each CCR concept may be assigned preferred labels (at most one per language), alternative labels, definitions, examples and various kinds of notes. Moreover, the ISOcat thematic domains and data category selections could be maintained by importing them to SKOS concept schemes and collections, respectively. Only one import decision turned out to be problematic: converting the data category identifier into a concept notation. SKOS notations are required to be unique within their concept scheme, whereas this constraint did not apply to data category identifiers in the DCR data model. A first clean-up round to remedy this has been finished successfully.

The SKOS model provides the possibility to express semantic relationships between concepts, e.g. broader than, narrower than and related to. In contrast, the DCR data model did only contain relationships based on the data category types, e.g., a simple data category belonged to the value domain of one or more closed data categories. These domain-range relationships do not correspond well to any of the SKOS relationship types. Careful manual inspection would be needed to determine if any mapping can be made, hence, for now these relationships have not been imported into the CCR. At a later date these facilities of the SKOS model and OpenSKOS can be exploited and could eventually take over the role originally envisioned for RELcat (Windhouwer, 2012). However, for now the initial focus is on the concepts themselves.

Neither SKOS itself nor OpenSKOS does yet provide an extensive versioning model, i.e., concepts can be expired but there is no explicit link to a superseding concept. This is now on the wishlist for the next version of OpenSKOS.

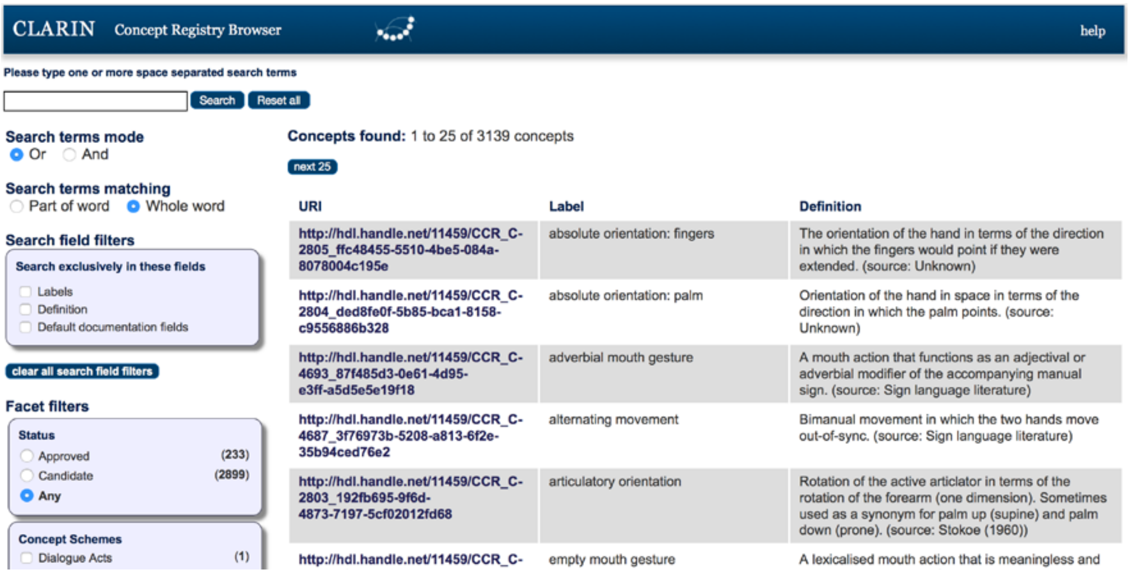
Being RDF-based SKOS also brings the potential to more easily join forces with linked data and semantic web communities.

## 4 The CCR content

In the past few years, many national CLARIN teams made an effort to enter their data in ISOcat. This work has not been useless as all entries deemed to be worthwhile for a specific CLARIN group were inserted in CCR. Leaving out redundant entries already means a considerable reduction in number of entries (from over 5000 in ISOcat (Broeder et al., 2014) to 3139 in CCR (see Figure 1; clarin.eu/conceptregistry, June 2015)). Although the imported concepts got new handles care was taken to retain a link with their ISOcat origin, so automated mapping is possible and can be used to convert references to ISOcat data categories into references to CCR concepts. A mapping tool for this has been developed and used for the CMDI components, but is generally applicable and available to the CLARIN community (github.com/TheLanguageArchive/ISOcat2CCR).

## 5 Maintaining the CCR – procedures and actors

Just like ISOcat the CCR can be browsed and searched by anyone, member of the CLARIN community or not, and anyone can refer to the concepts. However, contrary to ISOcat, only specifically appointed users, namely the national CCR content coordinators are given rights to update the CCR. These coordinators were assigned by CLARIN national consortia (see clarin.eu/content/concept-registry-coordinators) when the problems with the usage of ISOcat became apparent, and their mission was to improve the quality of the data categories (now concepts) used


CLARIN Concept Registry Browser
help
Search
Reset all
next 25
clear all search field filters

Please type one or more space separated search terms

**Search terms mode**
● Or ○ And

**Search terms matching**
○ Part of word ● Whole word

**Search field filters**

Search exclusively in these fields
□ Labels
□ Definition
□ Default documentation fields

**Facet filters**

Status
○ Approved (233)
○ Candidate (2899)
● Any

Concept Schemes
□ Dialogue Acts (1)

**Concepts found:** 1 to 25 of 3139 concepts

| URI | Label | Definition |
|---|---|---|
| http://hdl.handle.net/11459/CCR_C-2805_ffc48455-5510-4be5-084a-8078004c195e | absolute orientation: fingers | The orientation of the hand in terms of the direction in which the fingers would point if they were extended. (source: Unknown) |
| http://hdl.handle.net/11459/CCR_C-2804_ded8fe0f-5b85-bca1-8158-c9556886b328 | absolute orientation: palm | Orientation of the hand in space in terms of the direction in which the palm points. (source: Unknown) |
| http://hdl.handle.net/11459/CCR_C-4693_87f485d3-0e61-4d95-e3ff-a5d5e5e19f18 | adverbial mouth gesture | A mouth action that functions as an adjectival or adverbial modifier of the accompanying manual sign. (source: Sign language literature) |
| http://hdl.handle.net/11459/CCR_C-4687_3f76973b-5208-a813-6f2e-35b94ced76e2 | alternating movement | Bimanual movement in which the two hands move out-of-sync. (source: Sign language literature) |
| http://hdl.handle.net/11459/CCR_C-2803_192fb695-9f6d-4873-7197-5cf02012fd68 | articulatory orientation | Rotation of the active articlator in terms of the rotation of the forearm (one dimension). Sometimes used as a synonym for palm up (supine) and palm down (prone). (source: Stokoe (1960)) |
| http://hdl.handle.net/11459/CCR_C- | empty mouth gesture | A lexicalised mouth action that is meaningless and |

Figure 1. The CCR browser (clarin.eu/conceptregistry)

within CLARIN. With the CCR in place the national CCR content coordinators have teamed up more actively and established procedures around the CCR to fulfil this mission.

To deal with the ISOcat legacy the coordinators are doing a round of clean up with the aim to expire low quality concepts and recommend high quality concepts. Notice that, just like in ISOcat, expired concepts remain accessible, i.e., their semantic descriptions are not lost, but their active usage is discouraged. The main focus is on providing good definitions. A good definition should be "as general as possible, as specific as necessary" and should therefore be:

1. Unique, i.e., not a duplicate of another concept definition in the CCR;
2. Meaningful;
3. Reusable, i.e., refrain from mentioning specific languages, theories, annotation schemes, projects;
4. Concise, i.e., one or two lines should do;
5. Unambiguous.

As far as point 5 is concerned, a concept used in the entry of another concept, should be referred to using its handle. Detailed guidelines are under development by the coordinators and will become generally available in due course. Apart from defining best practice for the coordinator group, such guidelines will benefit users directly, enabling them to issue informed requests to the CCR coordinators (see below).

The changes the coordinators can do to existing concepts are limited, i.e., they should not change the meaning. Only typos, awkward formulations, etc. can be remedied. Otherwise a new concept has to be created, and the original one may be expired.

All the coordinators or their deputies are involved with these changes. In cases where they do not agree a vote might take place and the change will be performed if 70% or more of the coordinators agree. A book keeping of the results of votes is maintained at the CCR section of the CLARIN intranet. The time frame within which the discussions and possibly a vote have to reach a decision is 2 weeks. In the holiday seasons and during the initial startup phase a longer time period can be agreed upon by the coordinators.

Members of the CLARIN community wanting new concepts or changes to existing ones need to contact their national CCR content coordinator. Users from countries with no content coordinator should use the general CCR email address (ccr@clarin.eu) to file their requests. These requests will then be discussed within the national CCR content coordinators forum as described above. Note that in OpenSKOS any changes made to concepts are directly public. Therefore new entries or changes will only be entered after their content has been approved by the content coordinator forum. This procedure will take some time, but should result in better quality concepts, less proliferation and eventually a higher level of trust of the CCR content than was the case for ISOcat. One can also expect that the need for new concepts will diminish over time due to the CCR covering more and more of the domain.

## 6 Conclusions and future work

Although CLARIN just started working on the new OpenSKOS-based CLARIN Concept Registry and there is still a lot of ISOcat legacy to deal with, the new registry looks promising. Our feeling is that it will be able to provide a more sustainable and higher quality semantic layer for CMDI. An important lesson from the ISOcat experience is that technology is not always the main problem, although a complicated data model or interface never helps. What we do believe in, is establishing robust yet simple management procedures, as outlined in Section 5. These rely on good teamwork in the national CCR content coordinators forum, together with active involvement of the user community.

## Acknowledgements

The authors like to thank the national CCR content coordinators forum for the fruitful discussions on the procedures around the CCR. They also like to thank the Max Planck Institute for Psycholinguistics, CLARIN-NL and the Meertens Institute for their support to realize a smooth transition from ISOcat to the CCR.

## Reference

[Broeder et al. 2012] Daan Broeder, Menzo Windhouwer, Dieter van Uytvanck, Twan Goosen, and Thorsten Trippel. 2012. *CMDI: a Component Metadata Infrastructure*. Describing LRs with Metadata: Towards Flexibility and Interoperability in the Documentation of LR Workshop.

[Broeder et al. 2014] Daan Broeder, Ineke Schuurman, and Menzo Windhouwer. 2014. *Experiences with the ISOcat Data Category Registry*. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014), Reykjavik, Iceland.

[Brugman and Lindeman 2012] Hennie Brugman and Mark Lindeman. 2012. *Publishing and Exploiting Vocabularies using the OpenSKOS Repository Service*. Proceedings of the Describing Language Resources with Metadata workshop (LREC 2012), Istanbul, Turkey.

[Durco and Windhouwer 2013] Matej Durco and Menzo Windhouwer. *Semantic Mapping in CLARIN Component Metadata*. 2013. In E. Garoufallou and J. Greenberg (eds.), Metadata and Semantics Research (MTSR 2013), CCIS Vol. 390, Springer.

[ISO 12620 2009] ISO 12620. *Specification of data categories and management of a Data Category Registry for language resources*. 2009. International Organization for Standardization, Geneve.

[Windhouwer 2012] Menzo Windhouwer. *RELcat: a Relation Registry for ISOcat data categories*. 2012. Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012), European Language Resources Association (ELRA), Istanbul, Turkey.

[Wright et al. 2014] Sue Ellen Wright, Menzo Windhouwer, Ineke Schuurman and Daan Broeder. 2014. *Segueing from a Data Category Registry to a Data Concept Registry*. Proceedings of the 11th international conference on Terminology and Knowledge Engineering (TKE 2014), Berlin, Germany.

# DMPTY – A Wizard For Generating Data Management Plans

**Thorsten Trippel** and **Claus Zinn**
Seminar für Sprachwissenschaft
Universität Tübingen
firstName.lastName@uni-tuebingen.de

## Abstract

To optimize the sharing and reuse of existing data, many funding organizations now require researchers to specify a management plan for research data. In such a plan, researchers are supposed to describe the entire life cycle of the research data they are going to produce, from data creation to formatting, interpretation, documentation, short-term storage, long-term archiving and data re-use. To support researchers with this task, we built DMPTY, a wizard that guides researchers through the essential aspects of managing data, elicits information from them, and finally, generates a document that can be further edited and linked to the original research proposal.

## 1 Introduction

All research depends on data. To address a research question, scientists may need to collect, interpret and analyse data. Often the first phase of scientific activity, data collection, is the most decisive, and also a time-consuming and resource-intensive task. It must be planned well enough so that a significant number of data points are available for subsequent inspection so that underlying research questions can be analysed thoroughly. When the analysis of data is yielding results of significant interest, the study is described in scientific parlance, and then submitted to a scientific conference or journal. Once reviewed and accepted for publication, the resulting article constitutes the formal act of sharing research results with the scientific community, and most articles in reputable publication outlets are archived for posteriority. While the results are now public, the underlying research data often remains private, and usually stays with the individual researcher or the research organization. This makes it hard for other researchers to find and to get access to the data, and limits the opportunity for them to reproduce the results, or to base secondary studies on the same data. In brief, the sharing and long-term archiving has these four main benefits:

**Reproducibility** One of the main principles of the scientific method is reproducibility: it shall be possible to replicate experimental results, in preference by redoing the analysis on the existing data rather than on newly collected data. This discourages fraud and tempering with research data.

**Facilitation of secondary studies** With researchers having access to existing data sets, there is no need for a costly collection of new data, and therefore, it becomes easier for researchers to explore similar research questions, contrastive studies, or meta studies.

**Attribution** It should be good scientific practise to give an explicit acknowledgement of ownership or authorship to the one who has collected the data. Scientific reputation shall not only be merited by findings, but also by having acquired underlying data.

**Economy** Funding money and researchers' time shall not be wasted for collecting data sets if comparable data already exist. Open access to existing data also allows researchers to add to existing data sets, and hence might contribute towards a "Wikipedia effect", yielding increasingly rich resources.

To reap these benefits, research data should be accessible in public repositories, properly documented, with generous access rights, and possibly, in an easy-to-read, non-proprietory data format.

Funding agencies increasingly require grant applicants to complement their research plan with a plan for managing and sharing the research data that is going to be created during their research. In the United Kingdom, the site `https://www.dcc.ac.uk/resources/data-management-plans/funders-requirements` lists the major British funding bodies and their data policies. In Germany, the German Research Foundation expects data management plans, at least for larger collaborative research projects, and often funds personnel and necessary hard- and software to ensure that data is sustainably archived. The situation is similar in Switzerland where the Swiss National Science Foundation requires researchers that ”the data collected with the aid of an SNSF grant must be made available also to other researchers for secondary research and integrated in recognised scientific data pools” [1]. Some universities see data management issues as an integral part to good scientific practise. In the *Guidelines for Research Integrity of the ETH Zurich*, Article 11 is about the collection, documentation and storage of primary data, primarily, with the purpose that all ”research results derived from the primary data can be reproduced completely” [2]. The *Research Data Management Policy* of the University of Edinburgh states that ”All new research proposals [...] must include research data management plans or protocols that explicitly address data capture, management, integrity, confidentiality, retention, sharing and publication” [3].

Successful research proposals are centred around one or more research questions, the applicants apply sound research methodologies, and in empirically-driven research, this usually includes a convincing approach to gather and analyse research data to address the research questions. Individual research projects, by their very nature, take a short-term view: at the end of the projects research questions have been addressed in the best possible way, and results properly written up and published. Usually, there is no long-term view, in particular, with respect to the research data gathered. What is the research data life cycle, how will data be stored (*e.g.*, using which format) and adequately documented? How can the data be cited and made accessible for the scientific community for the long term (archival)? With regard to accessibility, how is personal or confidential data be taken care of? Which licence should be chosen for the data to ensure that other researchers can access the data in the future? Does the data collection ensure that research results can be reproduced, or that follow-up studies can use the data with ease?

We must not expect researchers to handle such questions, which are secondary to the research question, entirely on their own. Taking the long-term perspective requires a different set of skills, and it calls for a cooperation between researchers and a permanent research infrastructure. It is advisable that such cooperation is initiated at the early stage of a research project so that all aspects of the data life cycle are properly taken care of. Infrastructures such as CLARIN can assist researchers in managing their data.

The remainder of this paper is structured as follows. In Sect. 2 we describe the common elements of data management plans. Sect. 3 sets data management in the CLARIN context and defines the division of labour and shared responsibilities between data producer and data archive. In Sect. 4, we present the DMPTY wizard for data management planning. And in Sect. 5, we discuss related work and conclude.

## 2 Common elements of data management plans

In Britain, the Digital Curation Centre (DCC) ”provides expert advice and practical help to anyone in UK higher education and research wanting to store, manage, protect and share digital research data” (see `http://www.dcc.ac.uk/about-us`). The DCC, for instance, has a good summary page that overviews and links to data management plan requirements of a number of British funding agencies (see [4]). The DCC has also published a checklist for devising a data plan [5]. The checklist seems to be an amalgamation of the various plans, and with its broad base takes into account requirements from different scientific fields such as the Natural Sciences or the Humanities. The checklist is divided into eight different parts, and covers all of the essential aspects for managing research data:

1. **Administrative Data:** nature of research project, research questions, purpose of data collection, existing data policies of funder or research institution;

2. **Data Collection:** type, format and volume of data; impact on data sharing and long-term access; existing data for re-use; standards and methodologies, quality assurance; data versioning;

3. **Documentation and Metadata:** information needed for the data to be read and interpreted in the future; details on documenting data acquisition; use of metadata standards;

4. **Ethics and Legal Compliance:** consent for data preservation and sharing; protection of personal data; handling of sensitive data; data ownership; data license;

5. **Storage and Backup:** redundancy of storage and backup; responsibilities; use of third party facilities; access control; safe data transfer;

6. **Selection and Preservation:** criteria for data selection; time and effort for data preparation; foreseeable research uses for the data, preservation time frame; repository location and costs;

7. **Data Sharing:** identification of potential users; time frame for making data accessible; use of persistent identifiers, data sharing via repository and other mechanisms;

8. **Responsibilities and Resources:** for DMP implementation, review, & revision at plan and item level, potentially shared across research partners; use of external expertise; costs.

## 3 Data Management in the CLARIN Infrastructure

The CLARIN shared distributed infrastructure aims at making language resources, technology and expertise available to the Humanities and Social Sciences research communities. To streamline the inclusion of new data and tools, and to help researchers with managing their data, CLARIN-D now offers advice on data management plans and supports their execution. The CLARIN-D plan template mirrors the structure of the DCC checklist, but has a number of adaptations to best profit from the CLARIN infrastructure. As a first step, researchers are invited to select the CLARIN-D centre whose expertise matches best the type of resource being created during the project. This aims at ensuring that researchers get the best possible advice from a CLARIN-D centre of their choice.[1] With DMPTY, researchers are asked to contact their CLARIN centre of choice when starting to devise their research data plan.

With regards to the DCC plan, our plan adjusts to the CLARIN infrastructure as follows:

**Data Collection:** a policy on preferred non-proprietary data formats for all types of language-related resources (in line with the CLARIN-D User Guide, see `http://www.clarin-d.net/de/sprachressourcen-und-dienste/benutzerhandbuch`).

**Documentation and Metadata:** the selection of existing CMDI-based metadata schemes, or if necessary, the adaption of existing ones to best describe the research data.

**Ethics and Legal Compliance:** encouraging the use of the CLARIN License Category Calculator, see `https://www.clarin.eu/content/clarin-license-category-calculator`.

**Responsibilities and Resources:** a budget estimate that accounts for all the personnel and financial resources, and which are shared between data producer and CLARIN-D archive.

Moreover, there is a ninth plan component that describes a time schedule that data producer and data archivist agree on: when is the research data (i) described with all metadata, (ii) ingested in a data repository, and (iii) made accessible to interested parties or the general public? Moreover, it defines how long the research data should be held, and when, if applicable, it should be deleted. The data management plan shall also be complemented with a *pre-contractual agreement* between the data producer and the CLARIN centre for archiving, and which captures the rights and obligations of each partner.

[1] For identifying the most suitable CLARIN-D centre, researchers can consult the link `http://www.clarin-d.net/de/aufbereiten/clarin-zentrum-finden`.

## 4 The DMPTY Wizard for the Generation of CLARIN-supported DMPs

DMPTY is a browser-based wizard available at the German CLARIN-D portal.[2] The wizard makes use of the Javascript framework *AngularJS*, see `angularjs.org`, where each of the nine steps is presented to the user as an HTML form, and where navigation between the nine forms is easily possible so that information can be provided in flexible order. Associated with the forms is an HTML document that represents the data management template. Whenever the user enters information into one of the web form elements, the underlying plan is instantiated appropriately. At any given time (but within a browser session), it is possible to generate the plan as a text document in one of the formats Word, rtf and LaTeX, using the *pandoc* tool, see `pandoc.org`. Researchers can then edit the document to enter additional information to address, for instance, institution-special ethical or legal questions, or to state a cooperation with third parties, or to respond to specific requirements of funding agencies that we have not anticipated. Researchers may also want to add cross-links to relevant parts of the corresponding research proposal, change the formatting, *etc*.

At the time of writing, a beta version of DMPTY is publicly available; the wizard generates plans in German only, and it only lists CLARIN-D centres as cooperation partners. Upon a successful evaluation, DMPTY will also be capable of generating plans in English, and listing all CLARIN centres as archiving partner. So far, the scope of DMPTY is restricted to its application within the CLARIN world.

We are currently preparing an evaluation of DMPTY and seek interested researchers to participate in our evaluation study. We also intend do make DMPTY available on an HTTPS-enabled server to protect the privacy and integrity of all data exchanged between the web browser and the server.

## 5 Related Work and Conclusion

With data management plans becoming increasingly necessary to attract funding, there are now a number of tools available that help researchers to address all relevant data management issues. The DCC provides a web-based wizard to help researchers devising data management plans, see `https://dmponline.dcc.ac.uk`. Once researchers have selected a funding organisation for their research proposal, a corresponding data management plan template is created, which the wizard then follows step by step. The second DINI/nestor workshop was devoted to data management plans [6]. The workshop's programme featured, among others, a talk from the German Research Foundation on policy issues, several presentations on the nature and benefits of data management plans, and also talks on examples of successful data management, *e.g.*, for biological data and earth system science data. During the workshop, there were also a number of tools presented: the DMP Webtool from the University of Bielefeld (which is based on the WissGrid checklist [7]), and the TUB-DMP tool of the Technical University of Berlin. DMPTY is unique because its aims at matching data producers with research infrastructure partners at an early stage, ensuring that the entire life cycle of research data is properly taken care of.

Data management plans are becoming an accepted part of good scientific practice[3], and researchers must take into account all questions concerning their research data at an early stage of their research projects. Clearly, specifying and executing data management plans consumes resources, but the investment will pay off. DMPTY lowers the burden for researchers to develop their own plan, it guides them through all relevant aspects of such plans, and helps streamlining the cooperation with the CLARIN infrastructure. With CLARIN involved, researchers get support for the management of their data during the data's entire life cycle; touching base at regular intervals with CLARIN guarantees that the plans are up to their needs and properly executed. It also ensures that the appropriate resources (personnel, equipment) are accounted for. As a result, the number of high-quality accessible research data is bound to increase, which makes it easier for researchers to reap the benefits of sustainable data archiving and data re-use.

[2] See `http://www.clarin-d.net/de/aufbereiten/datenmanagementplan-entwickeln`.

[3] And managing research data at the grande scale poses significant challenges for research infrastructures, see [8].

## References

[1] Guidelines at the Swiss National Science Foundation. See http://www.snf.ch/sitecollectiondocuments/allg_reglement_e.pdf,Article44(b).

[2] Guidelines for Research Integrity at the ETH Zurich. See https://www.ethz.ch/content/dam/ethz/main/research/pdf/forschungsethik/Broschure.pdf.

[3] Research Data Management Policy at the University of Edinburgh. See http://www.ed.ac.uk/schools-departments/information-services/research-support/data-management/data-management-planning.

[4] Funder Requirements at the Digital Curation Centre (DCC). See http://www.dcc.ac.uk/resources/data-management-plans/funders-requirements.

[5] Checklist for a Data Management Plan. v.4.0, Edinburgh: Digital Curation Centre, 2013. Available online: http://www.dcc.ac.uk/resources/data-management-plans.

[6] Second DINI/nestor Workshop, Berlin, 2015. See http://www.forschungsdaten.org/index.php/DINI-nestor-WS2.

[7] J. Ludwig and H. Enke (Eds.) Leitfaden zum Forschungsdaten-Management. Verlag Werner Hülsbusch, Glückstadt, 2013. See http://www.wissgrid.de/publikationen/Leitfaden_Data-Management-WissGrid.pdf.

[8] B. Almas *et al.*, Data Management Trends, Principles and Components - What Needs to be Done Next? V6.1. EUDAT, 2015. See http://hdl.handle.net/11304/f638f422-f619-11e4-ac7e-860aa0063d1f.

*All links were accessed on September 24, 2015.*

# Enriching a grammatical database with intelligent links to linguistic resources

| **Ton van der Wouden**<br>Meertens Institute<br>The Netherlands<br>Ton.van.der.wouden@meertens.knaw.nl | **Gosse Bouma, Matje van de Kamp, Marjo van Koppen, Frank Landsbergen, and Jan Odijk** |
|---|---|

## Abstract

*We describe goals and methods of CLARIN-TPC, a project to enrich the on-line Taalportaal (Language Portal) grammatical database with intelligent links in the form of annotated queries to a variety of interfaces to on-line corpora and an on-line linguistic morphophonological database.*

## 1 Introduction

We describe how the on-line Taalportaal (Language Portal) grammatical database is enriched with intelligent links in the form of annotated queries to a variety of interfaces to on-line corpora and an on-line linguistic morphophonological database. This contributes to the use of the CLARIN research infrastructure, since

- It provides users with actual corpus examples for linguistic phenomena described in Taalportaal;
- It points out the existence and usefulness of search interfaces developed in the CLARIN infrastructure such as PaQu, GrETEL and OpenSONAR to linguists;
- By redirecting the user to the front-ends, it stimulates the further use of these applications in the CLARIN infrastructure for modifying queries or submitting new queries. Together with the multiple interfaces of most of these applications, this may also have a significant educational role.

## 2 Background

Linguistic data is everywhere. The working linguist is confronted with data any moment he/she reads a newspaper, talks to their neighbour, watches television, switches on the computer. To overcome the volatility of many of these data, digitized corpora have been compiled since the 1960 for languages all around the globe. These days, there is no lack of natural language resources. Large corpora and databases of linguistic data are amply available, both in raw form and enriched with various types of annotation, and often free of charge or for a very modest fee.

There is no lack of linguistic descriptions either: linguistics is a very lively science area, producing tens of dissertations and thousands of scholarly articles in a small country as the Netherlands only. An enormous amount of this linguistic knowledge, however, is stored in paper form: in grammars, dissertations and other publications, both aimed at scholarly and lay audiences. The digitization of linguistic knowledge is only beginning, online grammatical knowledge is relatively scarce in comparison with what is hidden in the bookshelves of libraries and studies.

Of course, there are notable exceptions. One such exception is the Taalportaal (Language Portal) project, that is currently developing an online portal containing a comprehensive and fully searchable digitized reference grammar, an electronic reference of Dutch and Frisian phonology, morphology and syntax. With English as its meta-language, the Taalportaal aims at serving the international scientific community by organizing, integrating and completing the grammatical knowledge of both languages.

To enhance the Taalportaal's value, the CLARIN project described here (NL-15-001: TPC) seeks to enrich the grammatical information within the Taalportaal with links to linguistic resources. The idea is that the user, while reading a grammatical description or studying a linguistic example, is offered the possibility to find both potential examples and counterexamples of the pertinent constructions in a range of annotated corpora, as well as in a lexical database containing a wealth of morphophonological data on Dutch. We explicitly focus on resources with rich linguistic annotations, since we want to do more than just string searches: searching for construction types and linguistic annotations themselves is one way to reduce the problem of the massive ambiguity of natural language words.

In light of the restricted resources, in terms both of time and money, this CLARIN project is not aiming at exhaustivity, that is, not all grammatical descriptions and not all examples will be provided with query links. TPC is thus explicitly to be seen as a pilot project, aiming for a proof of concept by showing the feasibility of efficient coupling of grammatical information with queries in a number of corpora.

## 3 The Taalportaal

The Taalportaal project (www.taalportaal.org) is a collaboration of the Meertens Institute, the Fryske Akademy, the Institute of Dutch Lexicology and Leiden University, funded, to a large extent, by the Netherlands Organisation for Scientific Research (NWO). The project is aimed at the development of a comprehensive and authoritative scientific grammar for Dutch and Frisian in the form of a virtual language institute. The Taalportaal is built around an interactive knowledge base of the current grammatical knowledge of Dutch and Frisian. The Taalportaal's prime intended audience is the international scientific community, which is why the language used to describe the language facts is English. The Language Portal will provide an exhaustive collection of the currently known data relevant for grammatical research, as well as an overview of the currently established insights about these data. This is an important step forward compared to presenting the same material in the traditional form of (paper) handbooks. For example, the three sub-disciplines syntax, morphology and phonology are often studied in isolation, but by presenting the results of these sub-disciplines on a single digital platform and internally linking these results, the Language Portal contributes to the integration of the results reached within these disciplines.

Technically, the Taalportaal is an XML-database that is accessible via any internet browser. Organization and structure of the linguistic information will be reminiscent of, and is is to a large extend inspired by, Wikipedia and comparable online information sources. An important difference, however, is that Wikipedia's democratic (anarchic) model is avoided by restricting the right to edit the Taalportaal information to authorized experts.

## 4 Enriching the Taalportaal with links to linguistic resources

CLARIN-NL15-001 is a collaborative effort of the Meertens Institute, the Institute of Dutch Lexicology, the Universities of Groningen and Utrecht, and Taalmonsters. In this pilot project, a motivated selection of Taalportaal texts will be enriched with links that encompass queries in corpus search interfaces. Queries will be linked to

- Linguistic examples
- Linguistic terms
- Names or descriptions of constructions

The queries are embedded in the Taalportaal texts as standard hyperlinks. Clicking these links brings the user to a corpus query interface where the specified query is executed – or, if it can be foreseen that the execution of a query takes a lot of time, the link may also connect to an internet page containing the stored result of the query. In general, some kind of caching appears to be an option worth investigating.

Two tools are available for queries that are primarily syntactic in nature:

- The PaQU web application
- The GrETEL web application (cf. Augustinus et al. 2013)

Both can be used to search largely the same syntactically annotated corpora (viz. the Dutch spoken corpus CGN (cf. van der Wouden et al. 2003) and the (written) LASSY corpus (cf. van Noord et al. 2006)), but they offer a slightly different functionality. Both applications offer dedicated user-friendly query interfaces (word pair relation search in PaQu and an example-based querying interface in GrETEL) as well as XPATH as a query language (cf. https://en.wikipedia.org/wiki/XPath), so that switching between these tools is trivial. Moreover, it is to be foreseen that future corpora of Dutch (and hopefully for Frisian as well) will be embedded in the very same CLARIN infrastructure, using the same architecture type of interface, allowing for the reuse of the queries on these new data.

Translation of a linguistic example, a linguistic term, or a name or description of a construction is not a deterministic task that can be implemented in an algorithm. Rather, the queries are formulated by student assistants. After proper training, they get selections of the Taalportaal texts to read, interpret and enrich with queries where appropriate. The queries are amply annotated with explanations concerning the choices made in translating the grammatical term or description or linguistic example into the corpus query. When necessary, warnings about possible false hits, etc. can be added. The student assistant's work is supervised by senior linguists. The page http://www.clarin.nl/node/2080 already contains a small example of a Taalportaal fragment adorned with a few query links using PaQu.

Next to the annotated corpora mentioned above, access to two more linguistic resources will be investigated in TPC. On the one hand, there is the huge SONAR corpus (cf. Oostdijk et al. 2013). The size of this corpus (> 500 M tokens) makes it potentially useful to search for language phenomena that are relatively rare. In this corpus, however, (morpho-)syntactic annotations (pos-tags, inflectional properties, lemma) are restricted to tokens (i.e., occurrences of inflected word forms). It comes with its own interface (OpenSONAR), which allows queries in (a subset of) the Corpus Query Processing Language and via a range of interfaces of increasing complexity. The current interface is not directly suited for linking queries as proposed here. For that reason, an update of this interface has been made to make the relevant queries possible. This updated version is available (as a beta version) on http://zilla.taalmonsters.nl:8080/whitelab/search/simple.

As the corpora dealt with so far offer little or no morphological or phonological annotation, they cannot be used for the formulation of queries to accompany the Taalportaal texts on morphology and phonology. There is, however, a linguistic resource that is in principle extremely useful for precisely these types of queries, namely the CELEX lexical database (cf. Baayen et al. 1995) that offers morphological and phonological analyses for more than 100.000 Dutch lexical items. This database is currently being transferred from the Nijmegen Max Planck Institute for Psycholinguistics (MPI) to the Leiden Institute for Dutch Lexicology (INL). It has its own query language, which implies that Taalportaal queries that address CELEX will have to have yet another format, but again, the Taalportaal user will not be bothered with the gory details.

As was mentioned above, the Frisian language – the other official language of the Netherlands, next to Dutch – is described in the Taalportaal as well, parallel to Dutch. Although there is no lack of digital linguistic resources for Frisian, internet accessibility is lagging behind. This makes it difficult at this point to enrich the Frisian parts of the Taalportaal with queries. It is hoped that this CLARIN project will stimulate further efforts to integrate Frisian language data in the research infrastructure.

## 5 Concrete Examples

The final paper will contain several concrete examples of Taalportaal descriptions and the links to their associated queries; we hope to demonstrate the system at the conference.

Since the links with the queries always go via the corpus search applications' *front-ends*, the Taalportaal user will, when a link has been clicked, be redirected not only to actual search results but also to a corpus search interface. The user can, if desired, adapt the query to better suit his/her needs, change the corpus being searched, search for constructions or sentences that diverge in one or more aspects (features) from the original query, or enter a completely new one. Since most applications used (viz. PaQu, GrETEL, and OpenSONAR) have multiple interfaces differing in pre-supposed background knowledge of the user, we believe that such options will actually be used. In this way, the enrichment of the Taalportaal as described here not only provides linguist users with actual corpus examples of linguistic phenomena, but may also have an educational effect of making the user acquainted with the existing corpus search interfaces.

## 6 Concluding remarks

We have described goals and methods of CLARIN-NL15-001, a co-operation project to enrich the online Taalportaal (Language Portal) grammatical database with intelligent links that take the form of annotated queries in a number of on-line language corpora and an on-line linguistic morphophonological database. The project will contribute to the research infrastructure for linguistics and related scientific disciplines, since

- It provides users with actual corpus examples for linguistic phenomena described in Taalportaal
- It points out the existence and usefulness of search interfaces developed in the CLARIN infrastructure such as PaQu, GrETEL and OpenSONAR to linguists
- By redirecting the user to the front-ends, it stimulates the further use of these applications in the CLARIN infrastructure for modifying queries or submitting new queries. Together with the multiple interfaces of most of these applications, this may also have a significant educational role.

## References

Liesbeth Augustinus, Vincent Vandeghinste, Ineke Schuurman, and Frank Van Eynde. (2013). Example-Based Treebank Querying with GrETEL – now also for Spoken Dutch. In: *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013).* NEALT Proceedings Series 16. Oslo, Norway. pp. 423-428.

R. H. Baayen, R. Piepenbrock & L. Gulikers, The CELEX Lexical Database (CD-ROM). Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA, 1995.

Frank Landsbergen, Carole Tiberius, and Roderik Dernison: Taalportaal: an online grammar of Dutch and Frisian. In Nicoletta Calzolari et al. (eds.): *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14*). Reykjavik, Iceland, European Language Resources Association (ELRA), 2014, 26-31.

Gertjan van Noord, Ineke Schuurman, Vincent Vandeghinste. Syntactic Annotation of Large Corpora in STEVIN. In: LREC 2006 (http://www.lrec-conf.org/proceedings/lrec2006/) .

Nelleke Oostdijk, Marc Reynaert, Veronique Hoste, Ineke Schuurman, (2013) The Construction of a 500 Million Word Reference Corpus of Contemporary Written Dutch in: *Essential Speech and Language Technology for Dutch: Results by the STEVIN-programme* (eds. P. Spyns, J. Odijk), Springer Verlag.

Ton van der Wouden, Ineke Schuurman, Machteld Schouppe, and Heleen Hoekstra. 2003. Harvesting Dutch trees: Syntactic properties of spoken Dutch. In *Computational Linguistics in the Netherlands 2002. Selected Papers from the Thirteenth CLIN Meeting*, ed. by Tanja Gaustad, 129-141. Amsterdam/New York: Rodopi.

# How can big data help us study rhetorical history?

**Jon Viklund**
Department of Literature
Uppsala University, Sweden
jon.viklund@littvet.uu.se

**Lars Borin**
Språkbanken/Dept. of Swedish
University of Gothenburg, Sweden
lars.borin@svenska.gu.se

## Abstract

Rhetorical history is traditionally studied through rhetorical treatises or selected rhetorical practices, for example the speeches of major orators. Although valuable sources, these do not give us the answers to all our questions. Indeed, focus on a few canonical works or the major historical key figures might even lead us to reproduce cultural self-identifications and false generalizations. In this paper we will try to demonstrate the usefulness of large-scale corpus studies ('text mining') in the field of rhetorical history, and hopefully point to some interesting research problems and how they can be analyzed using "big-data" methods.

## 1 Introduction

In the last decades digital humanities have grown bigger and more influential. Big data has been a buzzword for some time, and as larger amounts of texts are digitized and made searchable, we are able to see and investigate abstract patterns in large text masses that produce, partly, a new kind of knowledge. In 2010 an American team of researchers published a study based on the enormous amount of texts made public by Google – a corpus of over 5 million books – naming the new field of study *Culturomics* (Michel et al., 2011). This and other studies are both tantalizing and disappointing. One can hardly deny their potential, but most often their results affirm what we already knew. Moreover, the studies are generally constructed as showcases: they point to the advantages of different kind of 'text mining' in large collections of data, but so far they have not produced significant results that answer important research questions in the different fields of the humanities.

In this paper we will try to demonstrate the usefulness of such large-scale computational methods in the field of rhetorical history, and hopefully point to some interesting research problems and how they can be analyzed. In rhetorical studies, and more so in the field of rhetorical history, these quantitative, statistical methods have barely been tested. Our main historical interest lies in Swedish 19th century rhetorical culture, and we will demonstrate how quantitative analysis can be used to learn more about ideas of and attitudes towards eloquence in this period.

These issues are not easily investigated through traditional research perspectives used by historians of rhetoric. Usually the materials of investigation are either rhetorical treatises, or selected rhetorical practices, for example the speeches of major orators. These are valuable sources, but they do not give us the answers to all our questions. And in the worst case, focus on a few canonical works or the major historical key figures might even lead us to reproduce cultural self-identifications and false generalizations (Malm, 2014).

## 2 The research question: *doxa* in everyday discourse

Obviously, some questions are better suited than others for text mining. Thus we believe that it might be productive to raise issues that concern expressions of *doxa* (roughly: belief and opinion) in everyday discourse: What did people talk about, and believe was true, good, or beautiful? For instance, what were people's attitudes towards *eloquence* as a cultural phenomenon? How did people talk about eloquence, what ideas where brought forwards in relation to rhetoric, what kind of stereotypes were used, and how were they transformed?

Previous work has showed the cultural importance of of eloquence in Sweden up until the 18th century, as well as the centrality of rhetorical theory in schooling, aesthetics, sermons, political discourse etc. (e.g., Johannesson 2005). There is a general conception that the importance of rhetoric as a discipline diminishes during the 19th century, only to increase again in our days. However, we have no significant studies that confirm this general image, and probably we need to revise our view (Viklund, 2013). As demonstrated by Fischer (2013) in his study of the status of rhetoric and eloquence in the 18th century debate, the talk of rhetoric's 'demise' or 'death' in the period is based on an anachronistic view; one didn't conceive of rhetoric in those terms. The status of rhetoric can therefore best be evaluated through studies of the manner in which one talked about eloquence and rhetorical issues in general.

These issues are better investigated taking as the point of departure the large mass of everyday public discourse than through these singular and exceptional books and speeches. The literary scholar Franco Moretti (2013) calls these computational analyses of large masses of texts "distant reading", as opposed to "close reading", generally used in the humanities. Through abstraction and reduction, these quantitative analyses reveal patterns that only emerge from a distance. Of course, the results are not meaningful in themselves. The data are not explanations; these you need to supply yourself.

## 3 Methodology

The digital infrastructure used for this investigation is an advanced corpus search tool called *Korp* (Borin et al., 2012).[1] It is developed and maintained by Språkbanken (the Swedish Language Bank), a national center for collecting and processing Swedish text corpora, so that they are made available for researchers and the public. Språkbanken is the coordinating node of SWE-CLARIN, the Swedish CLARIN ERIC member organization,[2] and Korp is a central component of the Swedish CLARIN infrastructure, a mature corpus infrastructure with modular design and an attractive and flexible web user interface, which is also used by other national CLARIN consortia.

Språkbanken offers access to a large amount of annotated corpora and Korp offers the opportunity to make simple word searches as well as more complicated combined searches. The results are presented in three different result views: as a list with hits with context (KWIC); as statistical data, which for example give you the opportunity to create a trend graph of one or several words or lemmas; and, thirdly, as a 'word picture', which shows the most typical fillers of selected syntactic dependency relations of a word (most typical subjects and objects of a verb, most typical adjectival modifiers of a noun, etc.). Although originally devised for the purposes of linguistic analysis of texts, the word picture can be used as a kind of abstract topical maps that guide you to closer readings of the corpus. The corpus used in this study is a collection of historical newspapers from the late 18th to early 20th century, containing almost 70 million sentences. On the one hand this is small in comparison with the Google Books dataset, but our corpus is annotated with linguistic information, including lemmatization made using high-quality Swedish lexical resources (modern and historical), which goes a long way towards compensating for the smaller size of the corpus by providing much greater accuracy (Borin and Johansson, 2014; Tahmasebi et al., 2015).

## 4 Preliminary findings

So how can this search tool help us to learn more about the history of rhetoric? One of the great things about it is that it facilitates studies of historical transformations. When did certain words come into use, and when did they disappear? How does the interest in certain topics change over time?

The trend graph in Figure 1 shows the use of the word *talare* 'speaker' between 1800 and the 1920s in the Swedish press corpus. The nine peaks between the 1830s and 1860s coincide with the parliament sessions held every third year. In 1866 the old parliament where only the upper strata of society were represented, was dissolved, in favor of a new, more egalitarian parliamentary system. Apparently the newspapers were very keen on reporting the discussions in the old parliament, but less so thereafter, when the parliament met annually. The graph prompts a number of interesting questions: Why the sudden rise of interest in speakers around 1835, and why did the interest in the debates diminish? And still one

[1] See <http://spraakbanken.gu.se/korp/#?lang=en>.

[2] See <http://sweclarin.se>.

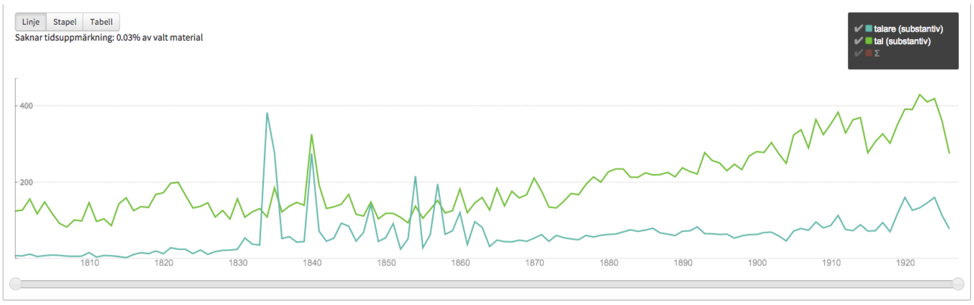


Figure 1: *Talare* ‘speaker’ and *tal* ‘speech’, 1800–1920s

notices a slow but steady increase of interest in the ‘speaker’ and in ‘speech’ in newspapers from the 1860s to the 1920s. Why is that?

This last question might be answered by a hypothesis, which certainly seems plausible: We ought to find a correlation between on the one hand democratization and the rising interest in politics, and, on the other, an increasing interest in rhetorical practices: oral performances and debate. As simple way of testing this one might see in what degree people talked about ‘politics’ and ‘democracy’. That is, through these simple searches one can get an idea of the topical changes in the newspapers. The graph in Figure 2 seems to confirm that there is an increasing interest in politics towards the end of the investigated period.

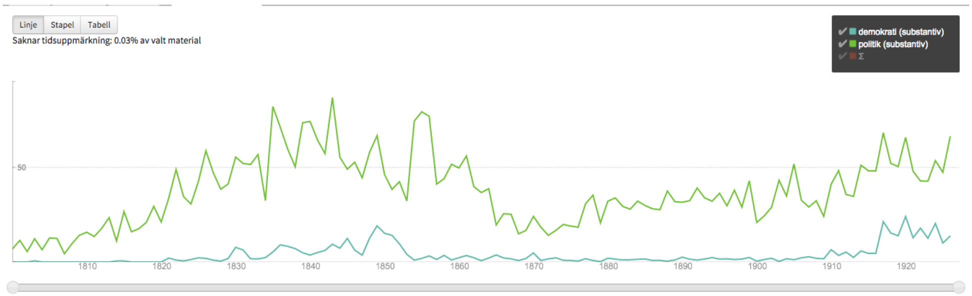


Figure 2: *Demokrati* ‘democracy’ and *politik* ‘politics’, 1800–1920s

## 5 ‘Eloquence’ in 19th century Swedish public discourse

All these results point to an increase in talk about rhetorical practices. If we investigate the word *vältalighet* ‘eloquence’, we see that it has a relatively stable trend curve. What is more interesting here is what words go together with this noun.

If we now use the word picture function of Korp to study the different modifiers, generally adjectives, we observe a pattern of notions that describe certain qualities of ‘eloquence’. They can be divided into three main groups, plus an ‘other’ category:

- **Genre:** ‘parliamentary’, ‘spiritual’, ‘political’, ‘Roman’, ‘academic’, ‘marital’, etc.
- **Truthfulness and naturalness:** ‘real’, ‘true’, ‘right’, etc.; ’manly’, ’natural’, ‘artless’, ‘unpretentious’, ‘simple’, etc.
- **Conceptual metaphors:** ‘fiery’, ‘glowing’, ‘burning’ (eloquence is *fire*); ‘flowing’, ‘fluent’, ‘pouring’ (eloquence is a *river/stream*)
- **Other:** ‘mute’, ‘normal’, ‘mighty’, ‘great’, ‘irresistible’, ‘bold’

This helps us to better understand the contexts in which one talked about ‘eloquence’, as well as the attitude towards rhetoric displayed in the corpus. One might for example investigate positive vs. negative connotations to the concept of eloquence, or one might look into gender differences in relations to the various examples and the categories produced by the computational reading.

We found the conceptual metaphors interesting. It was surprising to see that the majority of metaphors connected to the notion of 'eloquence' so clearly could be divided into two main categories: either *eloquence is fire* or *eloquence is a stream*. Methodologically we here used all the context material retrieved in connection to the word 'eloquence' and searched for the most common words in the conceptual metaphor clusters, and in that way ended up with many examples of these metaphorical expressions that were not necessarily associated with the specific word 'eloquence'.

So what can be learnt about how 'eloquence' was perceived from these clusters of words?

One topic concerns what *values* that are emphasized through these expressions. The fire metaphors generally express great pathos, burning heart, passion and energy. Also (and this is not the case with the flood metaphors) the fire metaphors are clearly coded as a male feature – there are no women described or speaking in this category.

As for 'eloquence' seen as a stream, the semantic orientation has more to do with rhetorical ability than degree of pathos. Both thoughts and feelings can come 'streaming from the speaker', but in general a person characterized with flowing eloquence has an ability to speak naturally and seemingly without effort, with rhetorical skill although without too much art. This metaphor is more often than the other used negatively: If someone speaks too fluently, with too many words and a shortage of genuine thoughts and feelings, then the metaphorical frame can be used ironically.

If the author wants to intensify the description, the two metaphorical clusters can be joined, as when a famous eloquent politician's speech is compared to a 'stream of lava', or when another's eloquence is described as 'a volcano of flowing eloquence'.

Another topic concerns the metaphor's *orientational frames*, more or less explicitly expressed. A fire is burning from the soul, and a flame, a bolt of lightning or an electric spark is coming from the speaker to the audience and sets the listeners's mind on fire. In the same way, the stream of eloquence is sometimes described as coming – with an often used cliché – from the heart of a speaker.

So, this closer look at conventional metaphors points to the importance of pathos, rhetorical force and oral virtuosity, as well as the value of naturalness and an eloquence that genuinely persuades, from heart to heart.

## 6 Conclusions and future work

Above we have described some initial experiments where a very large historical corpus of Swedish newspaper text and the state-of-the-art corpus infrastructure of SWE-CLARIN have been brought to bear on research questions in the field of rhetorical history.

A central and important purpose of this work was to investigate the method itself: Does it produce new knowledge? Yes and no. But also if the results only confirm old knowledge it is worthwhile: if the method reveals results that confirm old knowledge then it might be able to see also new things which until now have not been acknowledged. The method is promising, and we see several natural directions in which this work can be continued, e.g.:

We have not yet had the opportunity to see how the verbal conventions discussed above transform over time, but of course they do. Today, for one, *vältalighet*'eloquence' is rarely used – the word *retorik* 'rhetoric' has taken its place – and both these words appear in different contexts when looking at modern corpora such as newspapers and blogs from the last four decades, as compared to the 19th century material. Today we find almost exclusively two kinds of modifiers: type of rhetoric (for example: 'political', 'feminist', 'religious', 'social democratic') or negative qualities (for example: 'empty', 'made up', 'aggressive'), and of course the two categories combined (for example: 'racist', 'populist', 'scholastic'). With a corpus that covered also the 20th century it would be possible to study in detail the transformation of attitudes toward eloquence and rhetoric as expressed in Swedish public discourse over the last two centuries.

Finally, the present corpus infrastructure – intended mainly for linguistically oriented research – should be complemented by interfaces more geared towards supporting more general digital humanistic inquiry. In particular, to the form-oriented search useful to linguists we would like to add *content-oriented* search modes, as well as interfaces that allow the user to move easily and effortlessly between various forms of macro view visualization ("distant reading") and individual instances ("close reading"). This we believe to be a crucial – even necessary – feature of any such tool.

## References

Lars Borin and Richard Johansson. 2014. Kulturomik: Att spana efter språkliga och kulturella förändringar i digitala textarkiv. In Jessica Parland-von Essen and Kenneth Nyberg, editors, *Historia i en digital värld*.

Lars Borin, Markus Forsberg, and Johan Roxendal. 2012. Korp – the corpus infrastructure of Språkbanken. In *Proceedings of LREC 2012*, pages 474–478, Istanbul. ELRA.

Otto Fischer. 2013. *Mynt i Ciceros sopor. Retorikens och vältalighetens status i 1700-talets svenska diskussion*, volume 1 of *Södertörn Retoriska Studier*. Södertörns högskola, Huddinge.

Kurt Johannesson. 2005. *Svensk retorik. Från medeltiden till våra dagar*. Norstedts, Stockholm.

Mats Malm. 2014. Digitala textarkiv och forskningsfrågor. In Jessica Parland-von Essen and Kenneth Nyberg, editors, *Historia i en digital värld*.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science*, (331).

Franco Moretti. 2013. *Distant reading*. Verso, London/New York.

Nina Tahmasebi, Lars Borin, Gabriele Capannini, Devdatt Dubhashi, Peter Exner, Markus Forsberg, Gerhard Gossen, Fredrik Johansson, Richard Johansson, Mikael Kågebäck, Olof Mogren, Pierre Nugues, and Thomas Risse. 2015. Visions and open challenges for a knowledge-based culturomics. *International Journal on Digital Libraries*, 15(2–4):169–187.

Jon Viklund. 2013. Performance in an age of democratization: The rhetorical citizen and the transformation of elocutionary manuals in Sweden ca. 1840–1920. Paper presented at ISHR [International Society for the History of Rhetoric] biannual conference in Chicago.

# Research Data Workflows: From Research Data Lifecycle Models to Institutional Solutions

**Tanja Wissik**
ACDH-OEAW
Vienna, Austria
tanja.wissik@oeaw.ac.at

**Matej Ďurčo**
ACDH-OEAW
Vienna, Austria
matej.durco@oeaw.ac.at

## Abstract

In this paper we will present a research data workflow model covering the whole lifecycle of the data and showcases the implementation of the model in a specific institutional context. The main challenge addressed is how to harmonize existing processes and systems and reach a clear division of roles and workable, sustainable workflow in dealing with research data.

## 1 Introduction

Institutions like universities and academies have an increasing obligation to manage and share research data. For the majority of scholars these endeavours, especially in the humanities, are relatively new and not deeply integrated into their existing working practices: for example only recently funding bodies started to request a data management plan as part of a project proposal. Therefore it becomes necessary to present models that synchronise smoothly with existing workflows and identify the type that fits this environment best. This analysis is based on already existing lifecycle or workflow models taking into account the existing working practice and institutional requirements. Therefore research workflows and the related data management processes vary not only from discipline to discipline but also from one institutional context to another. Once a workflow model is in place it can serve also as a quality assurance mechanism.

In this paper we will present a case study from Austrian Centre for Digital Humanities of the Austrian Academy of Sciences, where a research data workflow model is being implemented. Starting from already existing (research) data lifecycle models we develop an institutional research data management model. The context-specific challenge for this undertaking has been to bring all the stakeholders together. Another challenge has been to be general enough to be applicable to different scenarios including national and international contexts due to the heavy involvement of the institute in the infrastructure consortium CLARIN-ERIC, most notably in its role as national coordinator and as service provider of the CLARIN B Centre with a domain specific repository providing depositing services for language resources on national level, the CLARIN Centre Vienna[1]

## 2 Research data lifecycle models

Data lifecycle models describe in different forms at a high level the data lifecycle. There are a wide range of data lifecycle models, each with a different focus or perspective. In this section we describe and discuss existing data lifecycle models.

### 2.1 Linear data lifecycle models

An example of the linear type is the USGS Science Data Lifecycle Model (see figure 1). This model describes the data lifecycle from the perspective of research projects and the activities that are performed, from the planning phase over the data collection and process phase to analysing the data and then to preserve the data at the end of a project and to publish and share them. In addition to these

[1] http://clarin.oeaw.ac.at

activities there are other activities that must be performed continually across all phases of the lifecycle. One activity is the documentation of the activities and the annotation of metadata, the other is quality assurance and backup and prevention of physical loss of data (see Faundeen et al., 2013).

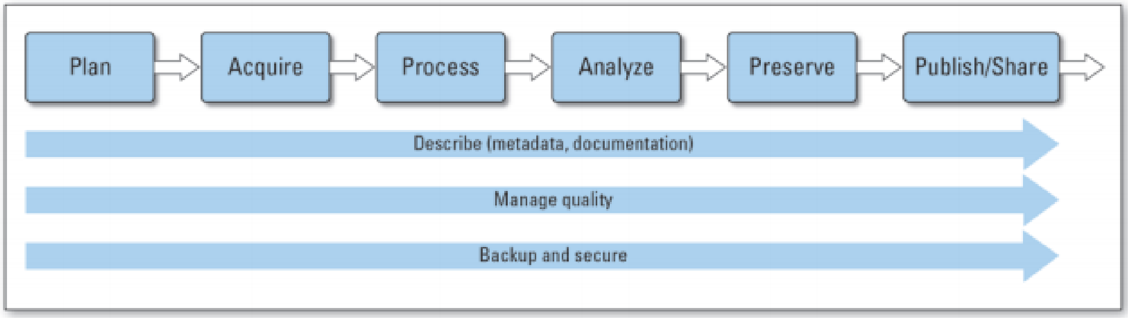


*Figure 1: USGS Science Data Lifecycle Model (Faundeen et al. 2013: 2).*

## 2.2 Circular lifecycle models

There are also circular models which try to reflect the iterative nature of the research process where each step builds on existing material. Circular models seem better suited to describe current research practices increasingly relying on sharing and reuse of data (beyond one researcher or group). For example the e-research and data and information lifecycle (Allan, 2009) with a focus on sharing of data and information.

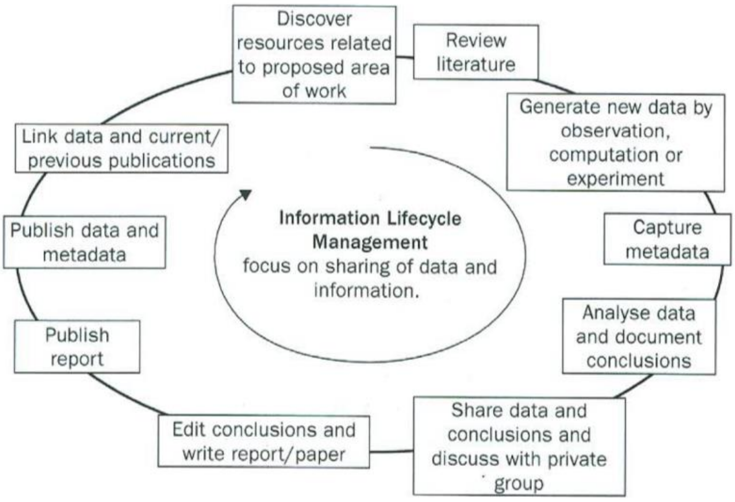

Discover resources related to proposed area of work
Review literature
Generate new data by observation, computation or experiment
Capture metadata
Analyse data and document conclusions
Share data and conclusions and discuss with private group
Edit conclusions and write report/paper
Publish report
Publish data and metadata
Link data and current/ previous publications
Information Lifecycle Management
focus on sharing of data and information.

*Figure 2: E-research and data and information lifecycle (Allan, 2009:22).*

Other lifecycle models concentrate on a specific step or phase, for example the DCC Curation Lifecycle Model (Higgings, 2008). It describes the different stages of data curation in detail but does not locate the curation process within a research project lifecycle.

## 2.3 Other models

There are also other types of lifecycle models or workflow models that do not fit into this classification, for example the non-linear GLBPM model (Barkow et al., 2013).

Also the OAIS – Reference Model, the Open Archival Information System Reference Model (Lavoie, 2004) does not fit in the classification above. The OASIS Reference Model is a concept model for digital repository and archival system it does not intend to represent the whole research workflow.

# 3 Research Data Management

As mentioned before, the data lifecycles are a high level presentation of processes. On the other hand data management model should be specific and detailed enough to serve as blueprint. In order to design the workflow the stakeholder, the different steps, and their dependences have to be identified for every task/scenario. While the abstract lifecycle models can serve as a guidance in practice the workflows will usually be more complex/irregular, due to context-specific constraints.

## 3.1 Scenarios

There are mainly two different scenarios for which the institutional research data management model has to be applied. The first scenario is when a new project proposal is written here we call this scenario *new project* (Fig. 3) the second scenario is when the project is already over, here (Fig. 3) we call this scenario *legacy data*.

## 3.2 Workflow Model

The research data management workflow model below shows the different scenarios from the perspective of the institute. We can identify five different phases, the pre-processing, the processing, the storage, the publishing phase and quality assurance. As illustrated in the model (Fig. 3) not all the phases are clear cut and can overlap. The quality assurance is also special, because it is a process that runs along or behind each of the workflow steps.

When a new project, in the proposal stage, approaches the institute for advice, in the pre-processing phase the data management plan is elaborated and in the ideal case the institute and its support is included into the project proposal. The pre-processing phase in our model would correspond to the activities *plan* and *acquire* in the USGS Science Data Lifecycle Model (Fig. 1). If the new project involves digitisation, this is also done in the pre-processing phase as well as the data modelling. In the processing phase we understand the whole of actual research activities related to the data. Ideally the researchers work in an integrated collaborative working space, where they get offered a series of tools for annotating, analysing, visualizing etc. run as a service by the institute. Currently this portfolio of tools is being built up combining existing open source applications as well as solutions specific to a task. It is important to realize/note that the processing phase as the whole workflow cannot be seen as a simple step or linear sequence of steps, but rather a complex, non-linear, iterative process, both within one project as well as beyond the project boundaries. The processing phase corresponds to the activities *process* and *analyse* in the USGS Science Data Lifecycle Model. The collaborative working space reflects the activities *share data and conclusion and discuss with private group* in the data lifecycle by Allan (2009). In both lifecycle models publishing activities are foreseen as well as in our proposed workflow.

When confronted with legacy data, in a first step, all the relevant data is stored in a kind of “quarantine” repository to be further processed. Then the data and the data model/structure are examined, especially with respect to the suitability of the format, existence of metadata and documentation and internal structure of the data. Based on the analysis it is decided if the data model needs to be adapted, transformed together with the estimation of the required resources of such transformation. Then the data is stored (see storage phase below) in the repositories and archived without going through the processing phase. Usually, there are only limited resources to deal with legacy data, the primary goal is to ensure a reliable deposition of the data. Thus as long as no new user/project interested in this data arises, no interaction with the data is expected in the working space, neither is an online publication.

In the storage phase the data and metadata are stored in one of the repositories and archived. Even though we say “phase” this aspect cannot be seen detached from the other processes. Providing storage is obviously an indispensable part of the whole process. Though we need to distinguish different kinds of storage. In the processing phase a lot of data is produced, oftentimes of transitional nature. We call this working data. Stable data aimed at long-term availability and/or publication is moved to the institutional or domain specific repository, which in the long run represent the main source for the datasets. At the archiving stage it is necessary to ensure long-term availability of the data even beyond a disaster scenario (Main repository is damaged through fire or similar). This involves geographically distributed replication/mirroring of the data to reliable providers of storage services, like scientific data centres. Here, we build up alliances with national providers as well as international players mainly in the context of the

EUDAT initiative. Archiving and preservation activities are only mentioned in the USGS Model and in our workflow model.

The publishing phase refers primarily to presentation, online and/or print, of the results of the project but also in line with the open access policy and subject to copyright and ethical restriction - the provision of the underlying research data.

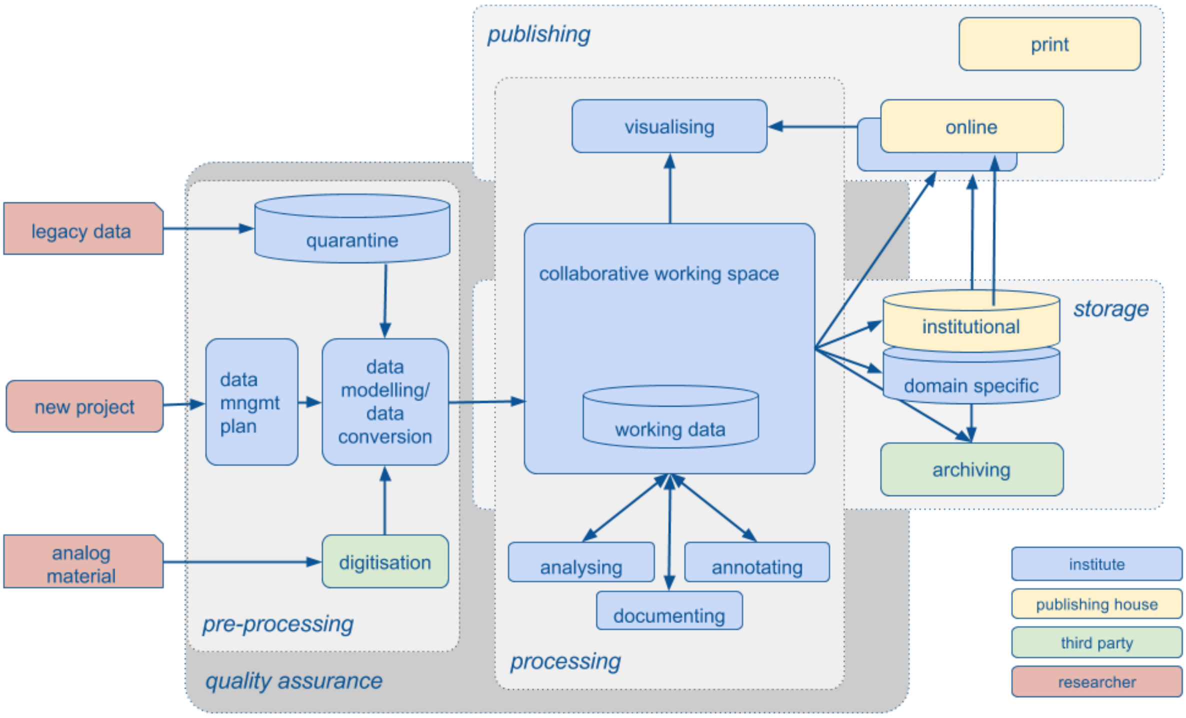


*Figure 3: Proposed institutional research data management workflow.*

### 3.3 Real Use Case - Stakeholders and the Current Status

Following stakeholders play a role in the described workflow model in the specific institutional setting of the Austrian Academy of Sciences: ACDH-OEAW[2], a newly founded research institute of the Academy that also supports researchers in the humanities as service unit; the institutional publishing house Academy Press; the institutional computing centre of the Academy; the institutional library; third-party service providers and the researchers, both within and outside the academy.

The ACDH-OEAW runs a domain specific repository for the arts and humanities, especially for language resources, the Language Resources Portal at the CLARIN Centre Vienna. It also offers a range of applications and services for processing, analysing, visualising and querying different kinds of data.

The Press has been operating the institutional repository of the Academy, epub.oeaw[3] that is designated to hold especially scientific publications, but also increasingly research data. The repository serves a double role: publication and archiving, data in the repository being replicated to Austrian National Library. So, while there is some overlap in the task description of epub.oeaw and CCV/LRP, they provide distinct features, that justify the co-existence of the two repositories.

Currently, the stakeholders are elaborating a common strategy to act as a coordinated network of providers for data-related services, with clear division of roles. In this plan, ACDH-OEAW will concentrate more on the interaction with the researchers (consulting, data modelling), development and provision of innovative tools. The Press will keep the running the repository for archiving and publishing of publications and certain types of research data. However not all kinds of resources are equally well suited for the digital asset management system underlying epub.oeaw, like for example relational databases, corpora or graph-based data. ACDH-OEAW still needs to work out together with the computing

[2] http://www.oeaw.ac.at/acdh
[3] http://epub.oeaw.ac.at/

centre a strategy for archiving for this kind of data. Furthermore, there are plans to establish in-house capacities for digitization at the institutional library that also serves as an important content provider.

### 3.4 Relation to CLARIN

Given that the considered institute runs a CLARIN Centre and is a national coordinator of CLARIN activities, many aspects of the workflow are strongly guided by the requirements expected by CLARIN-ERIC – assignment of persistent identifiers, metadata in CMDI format, OAI-PMH endpoint as a dissemination channel for the metadata harvested by the CLARIN harvester. One of the aims of the presented strategy is to make new resources automatically available via the CLARIN infrastructure.
Currently, for the resources we use 4 different CMDI-Profiles, and make the resources available in different forms, partly as raw data, partly within complex web applications that allow search and browsing through the data via different dimensions (linguistic, semantic). The access to resources and services is managed through Federated Identity.

Also with respect to the tools offered for use, there is a reciprocal relation to CLARIN, where tools from the CLARIN community are part of the portfolio (WebLicht) as well the solutions developed at the institute are made available to the whole of CLARIN community.

The plans for archiving evolve around the relation of CLARIN-ERIC to the EUDAT initiative as well as the options offered by the new CLARIN-PLUS project.

## 4 Conclusion

In this paper we presented an institutional workflow model for research data as it is implemented at the ACDH-OEAW, a newly founded research institute of the Austrian Academy of Sciences, but acts also as a service unit for researchers in the art and humanities in the institutional and national context. Starting from abstract (research) data lifecycle models, we discussed the stakeholders and scenarios for the specific institutional settings and elaborated a workflow model that caters to the specific situation of the implementing institution.
The paper shows, that the exposition of an institutional research data workflow model is important since there is no “one-size-fits-all-solution” but high level data lifecycle models are a good basis to start with. Once the workflow model is implemented, it can not only used as quality assurance itself but it can also help researchers, at first sight, in the project planning phase, when and whom to approach for advice and assistance.

## Reference

[Allan, 2009] Robert Allan. 2009. Virtual Research Environments. From portals to science gateways. Oxford: Chandos Publishing.

[Barkow et al., 2013] Ingo Barkow, William Block, Jay Greenfield, Arofan Gregory, Marcel Hebing, Larry Hoyle, Wolfgang Zenk-Möltgen. 2013. Generic longitudinal business process model. DDI Working Paper Series – Longitudinal Best Practice, No. 5. DOI: http://dx.doi.org/10.3886/DDILongitudinal2-01

[Faundeen et al., 2013] John L. Faundeen, Thomas E. Burley, Jennifer A. Carlino, David L. Govoni, Heather S. Henkel, Sally L. Holl, Vivian B. Hutchison, Elizabeth Martín, Ellyn T. Montgomery, Cassandra C. Ladino, Steven Tessler, and Lisa S. Zolly . 2013. The United States Geological Survey Science Data Lifecycle Model. U.S. Geological Survey Open-File Report 2013–1265, 4 p, http://dx.doi.org/10.3133/ofr20131265.

[Higgins, 2008] Sarah, Higgins. 2008. The DCC Curation Lifecycle Model. The International Journal of Digital Curation. Issue 1, Volume 3, 2008

[Lavoie, 2004] Brian F. Lavoie. 2004. The Open Archival Information System Reference Model: Introductory Guide. OCLC Online Computer Library Center.

# Towards depositor-controlled data ingest in digital repositories for linguistic corpora: automation of metadata processing

**Dennis Zielke**
Computer- und Medienservice
Humboldt-Universität zu Berlin
Unter den Linden 6, D-10099 Berlin
zielkede@cms.hu-berlin.de

**Daniel Jettka**
Hamburger Zentrum für Sprachkorpora
Universität Hamburg
Max-Brauer-Allee 60, 22765 Hamburg
daniel.jettka@uni-hamburg.de

## Abstract

This paper is a joint work of two projects that deal with the establishment and advancement of digital repositories for linguistic corpora at their research institutes. Although the repositories have fairly different profiles, with freely available historical written corpora on the one hand and mostly access-restricted multilingual spoken corpora on the other hand, they face similar challenges with regard to the pressure of automation and optimization of common workflows and processes because of limited financial and human resources. Against this background, the present paper focuses on the possibilities of automating validation and indexing processes as necessary steps towards supporting depositor-controlled data ingest. The technical involvement of depositors in the ingest process of research data becomes inevitable, especially when taking into account the incorporation of resources that are created over a long period of time (like reference corpora), while being updated and published in relatively short cycles.

## 1 Introduction

The present paper is based on the experiences from two projects that deal with the establishment and further development of digital repositories for linguistic corpora.

The LAUDATIO repository[1] was created in the course of a cooperative research data infrastructure project since 2011 at the Humboldt-Universität zu Berlin (hereafter HU Berlin). The repository is an open access environment for persistent[2] storage of historical text and their annotations (Zielke et al 2014). It currently contains historical corpora from various disciplines with a total of 2000 texts that contain about two million word forms. The main focus lies on German historical texts and linguistic annotations including all dialects of time periods ranging from the 9th to the 19th century. The technical repository infrastructure is based on generalizable software modules[3] such as the graphical user interface, the data exchange module between research data and the Fedora REST API. The metadata search for indexing and faceting is based on the Lucene-based technology ElasticSearch[4]. The imported corpora are stored in their original structure in a permanent and unchangeable version[5].

The repository of the Hamburger Zentrum für Sprachkorpora (HZSK)[6] is developed since 2011 with regard to the necessity to store and preserve the linguistic corpora that were created during the funding period the Collaborative Research Centre 538 "Multilinguality" (SFB 538 "Mehrsprachigkeit") at the Universität Hamburg between 1999 and 2011 (cf. Jettka/Stein 2014). The repository is based on the software triad Fedora, Islandora, and Drupal, and currently contains 19 corpora of transcribed spoken language and one dependency treebank, with additional corpora being included continuously. The preserved research data includes texts, transcripts, audio and video data, images, metadata, and other data types. The repository is connected to the CLARIN-D infrastructure

[1] http://www.laudatio-repository.org
[2] Persistent referencing for each version with EPIC-handle, https://de.dariah.eu/pid-service
[3] https://github.com/DZielke/laudatio
[4] ElasticSearch website, http://www.elasticsearch.org
[5] http://www.laudatio-repository.org/repository/technical-documentation/structure/versioning.html
[6] https://corpora.uni-hamburg.de/repository

on various levels, e.g. the central services Virtual Language Observatory (for metadata search) and the CLARIN Federated Content Search (for search directly in the content).

Although the two repositories have fairly different profiles, they have a number of commonalities, not only on the technical level (e.g. both are based on the Fedora repository system) but also with regard to the procedures that constitute the basis of the ingest and management of research data. A generalized model of the data ingest process, that takes into account necessary administrative, technical, and communicative actions of depositors and repository managers, will be presented in the following section. A specific commonality of the repositories can be found in the central role that standardized metadata plays on various stages of the data processing. Since this is fundamental to many processes that take place during the ingest of new research data, we identified the automation of metadata processing as one of the most important steps towards the optimization and reduction of the workload of repository managers. Therefore, a special focus of this paper lies on the handling of standardized metadata at various steps in the data ingest process, e.g. for the validation of datasets, modelling of data structures, creation of front-end views, indexing and faceting (i.e. search facilities).

## 2 A generalized model of the data ingest process

Before discussing the possibilities of automating sub-processes of the data ingest, we will take a brief look at a simplified model that demonstrates the necessary actions of depositors and repository managers for initiating and finalizing an ingest process. The model does not claim to be exhaustive but rather serves as a common ground for the continuing discussion.

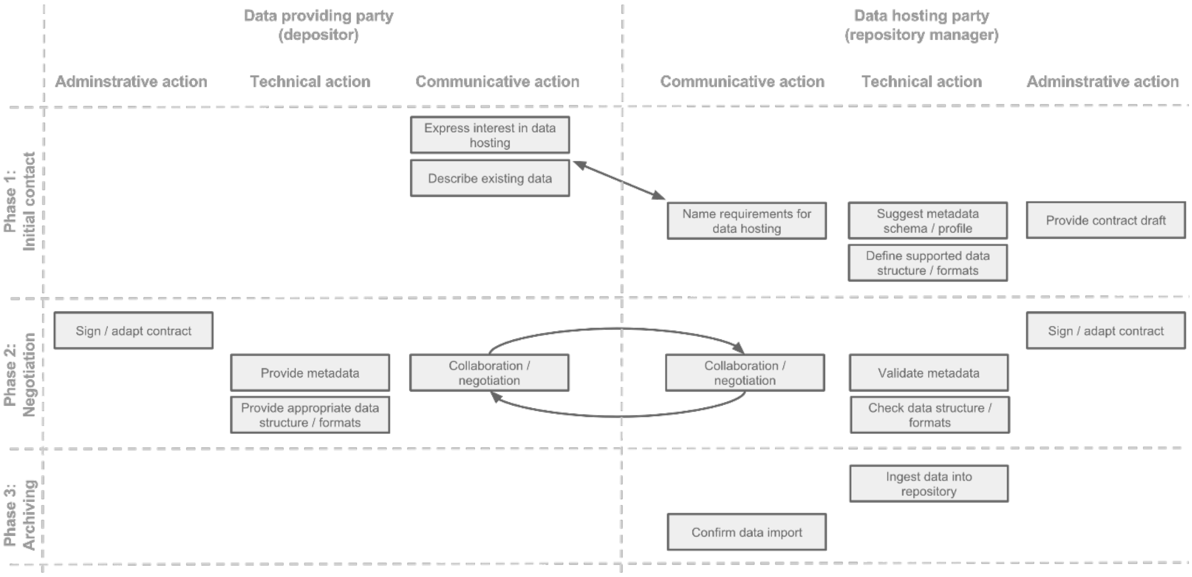

Data providing party (depositor)
Data hosting party (repository manager)
Adminstrative action
Technical action
Communicative action
Communicative action
Technical action
Adminstrative action
Phase 1: Initial contact
Express interest in data hosting
Describe existing data
Name requirements for data hosting
Suggest metadata schema / profile
Define supported data structure / formats
Provide contract draft
Phase 2: Negotiation
Sign / adapt contract
Provide metadata
Provide appropriate data structure / formats
Collaboration / negotiation
Collaboration / negotiation
Validate metadata
Check data structure / formats
Sign / adapt contract
Phase 3: Archiving
Ingest data into repository
Confirm data import

*Figure 1: Generalized model of the data ingest process*

It can be seen that both parties are intensively involved in all phases of the process. During the initial contact phase the Submission Information Package/s (SIP, cf. the OAIS model in CCSDS 2012) have to be defined, which often is a collaborative process. In addition the formal and technical requirements for submitting data to the repository have to be pointed out by the data hosting party (repository manager).

There is some potential for optimizing processes already in the initial contact phase:

- by providing as much information about data hosting requirements as possible via easily accessible webpages, a repository can reduce individual support cycles;

- by giving detailed information about existing licensing models and pointing depositors to tools like the CLARIN License Category Calculator[7], a repository can speed up the potentially time-consuming process of identifying applicable licenses.

## 3 The role of standardized metadata in the ingest process

This section deals with the role of standardized metadata, like CMDI[8] and TEI[9], which are the main frameworks used by the underlying repositories at various stages of the ingest process. Although in principle it is possible to extend existing, supported metadata schemas and profiles, or to support additional schemas in the ingest process, this normally entails necessary manual adaptations of repository mechanisms that rely on a specific metadata format (e.g. indexing, generation of frontend views). Therefore, in the remainder of this paper we will concentrate on the state of the ingest process when the depositor and repository manager have already agreed upon the use of a currently supported metadata schema and the ingest process goes into the phase of negotiation (cf. Figure 1).

### 3.1 Validation of datasets

The initial validation of incoming metadata against appropriate metadata schemes is a necessary step to ensure a minimal degree of consistency and represents the starting point for later processing. A repository may provide a graphical user interface for depositors to validate their metadata against existing schemas. Such a module can be implemented fairly easily and relocates necessary work from the repository manager to the depositor who can check and correct his metadata immediately and independently.

However, schema-based metadata validation may not be sufficient in all cases. There can be important free-text metadata fields that have been filled in by the depositor, but the associated schema might not be precise enough or the provided values, although valid to a schema, may not conform to the requirements of a repository that might pose additional (soft) constraints on individual metadata fields, e.g. again for later indexing or the creation of frontend views.

For instance, a repository may expect certain metadata values to be ISO-compliant while a metadata schema does not. This makes it necessary to perform additional validation, e.g. for ensuring the accuracy of numerical values and date formats (YY or YYYY or YYYY-MM-DD). After this step, mixed dates e.g. 2000-MM-DD and incorrect dates e.g. 31.09.2014 or unexpected dot notation in integer values, e.g. 200.000 should be reported.

Another example is the unintended use of metadata fields which come into play in later processing steps. Providing the exact same values for name and title values for a resource may result in an unintended display of resource information (e.g. “Demo corpus (Demo corpus)”), which would have to be captured before the finalization of the ingest process.

Cases like this cannot necessarily be covered completely by automatic metadata validation but often require personal inspection of the metadata by a repository manager. Apparently, the repository should provide detailed information on the semantic and pragmatic functions of metadata fields to avoid unnecessary communicative and technical processing cycles at this stage.

### 3.2 Modelling import formats and data structures

Metadata records often do not only include a more or less specific description of digital content, but also represent relationships between individual content items, i.e. structural information about datasets. Because of the rich information about the data and its relations, metadata records, at least in the cases of the LAUDATIO and HZSK repositories, serve as the basis for the creation of specific ingest formats, which in the context of the Fedora repository system[10] is Fedora Object XML (FOXML).

Because of the relatively simple structure of FOXML, the conversion can be carried out completely automatically, e.g. by XSLT transformations when dealing with XML metadata like in our cases (TEI/CMDI to FOXML). Once the conversion has been implemented, and if previous validation and

[7] https://www.clarin.eu/content/clarin-license-category-calculator
[8] http://www.clarin.eu/content/component-metadata
[9] http://www.tei-c.org/Guidelines/P5/
[10] https://wiki.duraspace.org/display/FF/Downloads

consistency checks have been performed successfully, there should not be too much work to be done by the repository manager for preparing new data at this step of the ingest process.

### 3.3 Indexing of metadata

In order to provide search facilities e.g. inclusion of a faceted search, the use of search engine technology is helpful and allows for comprehensive indexing and searching of valid metadata records. The above-mentioned repositories adopted Lucene-based approaches (ElasticSearch and Apache Solr), both of which require the initial building of search indexes, a task that can be implemented as a completely automatic processing step if valid and (semantically and pragmatically) consistent metadata is available. During the mapping previously identified relevant metadata fields are integrated into an exchange format of the search technology (e.g. TEI to JavaScript Object Notation (JSON) in the LAUDATIO repository's IndexMapping[11], or CMDI to Solr XML at the HZSK) and thereby can be exposed to performant search queries.

The process of indexing is related to the creation of frontend views (e.g. HTML pages) for browsing through digital collections and content, whether existing frameworks are used or custom solutions are implemented. By providing well-defined and controlled mechanisms for depositors to influence the indexing and/or creation of views on the basis of (parts of) their metadata[12], repositories can not only prevent potentially extensive manual adaptions of internal processing mechanisms but also further motivate the usefulness of depositor-controlled data ingest.

## 4 Conclusion

The automation of the processing of standardized metadata as a preliminary step towards depositor-controlled ingest of research data in digital repositories is subject to a number of requirements. It should for instance be accessible to depositors with various levels of experience in the different processing steps, including validation, quality/conformance assurance, conversion and indexing of metadata. In this paper we aimed at the identification of some benefits and limits of the automation of processing steps.

Besides technical aspects, the importance of providing detailed and easily accessible information for depositors describing the requirements and execution of the data ingest process/es has to be emphasized as it can help to optimize communicative processes between depositors and repository managers.

While many processes, like the derivation of data structures, creation of exchange formats, indexing, and creation of frontend views, on the basis of standardized metadata can be executed automatically to a great extent, as it has been suggested in the previous sections, there are steps like the assurance of quality and compliance of metadata with specific repository requirements that at the moment rely on personal involvement of repository managers and their direct collaboration with depositors.

## References

[CCSDS 2012] CCSDS. 2012. Reference Model for an Open Archival Information System (OAIS). *Magenta Book. Issue 2. June 2012*. URL: http://public.ccsds.org/publications/archive/650x0m2.pdf [Last access: 14.07.2015]

[Jettka/Stein 2014] Jettka, D. & Stein, D. 2014. The HZSK Repository: Implementation, Features, and Use Cases of a Repository for Spoken Language Corpora. *D-Lib Magazine*, Vol. 20, No. 9/10 (2014). DOI: doi:10.1045/september2014-jettka

[Zielke et al. 2014] Zielke, D., Schernickau, T. & Kindling, M. 2014. Open Access Research Data Repository for Corpus Linguistic Data – A Modular Approach. *Open Repositories 2014 Conference*. June 9.-13. Helsinki, Finland. URL: http://urn.fi/URN:NBN:fi-fe2014070432241 (Last visited: 30.09.2015)

[11] https://github.com/DZielke/LAUDATIO-IndexMapping

[12] In the LAUDATIO repository it is possible to provide customized JSON as a basis for index mapping, and the HZSK repository allows the free use of XSLT stylesheets for creating customized HTML views.

# Author Index

# Keyword Index