

Proceedings of the Second Workshop on

# Corpus-Based Research in the Humanities

## CRH-2

25-26 January 2018 Vienna, Austria

Editors:

Andrew U. Frank

Christine Ivanovic

Francesco Mambrini

Marco Passarotti

Caroline Sporleder

## Proposed BibTeX entries:

```
@Proceedings{crh-2,  
  title = {Proceedings of the  
    Second Workshop on      Corpus-  
Based Research      in the  
Humanities {CRH-2}},  
  year  = {2018},  
  editor = {Andrew U. Frank and  
    Christine Ivanovic and  
    Francesco Mambrini and  
    Marco Passarotti and  
    Caroline Sporleder},  
  volume = {1},  
  series = {Gerastree proceedings},  
  isbn  = {978-3-901716-43-0},  
}  
  
@InProceedings{crh2intro2018,  
  author   = {Francesco Mambrini and Marco Passarotti  
    and Caroline Sporleder},  
  title    = {Preface},  
  booktitle = {Proceedings of the Second Workshop on  
    Corpus-Based Research in the  
    Humanities {CRH-2}},  
  year     = {2018},  
  editor   = {Andrew U. Frank and Christine Ivanovic  
    and Francesco Mambrini and Marco  
    Passarotti and Caroline Sporleder},  
  volume   = {1},  
  series   = {Gerastree proceedings},  
  pages    = {I-IV},  
  isbn     = {978-3-901716-43-0},  
}
```

Copyright ©2018 by the individual authors.  
All rights reserved.

ISBN 978-3-901716-43-0  
Published by Gerastree Proceedings, GTP 1.  
Dept. of Geoinformation, TU Wien, Austria.

## Cover:

Les Fourches, seen from the Bretagne coast near Plouarzel  
(France). Photo by Andrew U. Frank.

## Preface

The second edition of the international workshop on "Corpus-based Research in the Humanities" (CRH) is held in Vienna, hosted by Academy Corpora of the Austrian Academy of Science (<https://www.oeaw.ac.at/ac/>). It follows, on a biannual basis, the edition held in Warsaw in December 2015. But the origins of the workshop go back even further, CRH being the direct descendant of the former workshop on "Annotation of Corpora for Research in the Humanities" (ACRH), which was held three times: in Heidelberg (January 2012), Lisbon (November 2012), and Sofia (December 2013).

All the previous editions of ACRH/CRH were co-located with the international workshop on "Treebanks and Linguistic Theories" (TLT). This year, for the first time, CRH is an event on its own and it spans over two days. However, both the organizers of TLT and CRH worked to keep the connection between the two workshops as tight as possible and the two events as close as possible, both in time and place. As the sixteenth edition of TLT takes place in Prague in the two days before CRH, we hope that many scholars will be able to attend both workshops in a row. We want to thank very much Jan Hajič, the co-chairs of TLT-16 and the colleagues at the Institute of Formal and Applied Linguistics in Prague for doing their best to organize TLT in those days. And we thank Erhard Hinrichs, who took care of TLT since its first edition, for supporting ACRH/CRH as the co-located event of TLT for all these years.

Although for the 2015 edition of CRH we had received a rather limited number of submissions (17), we had the feeling that the topic of CRH was not only well motivated but also promising.

During the days in Warsaw, we met Andrew Frank and Christine Ivanovic from Vienna, who gave a joint talk there and were very positively impressed by both the motivations and the results of the workshop. In light of the ever growing Digital Humanities in Vienna, they offered to organize an edition of CRH there. We thought that this was a wonderful opportunity for CRH to grow and finally become independent. We accepted the invitation gladly and we are now happy to welcome Andrew Frank and Christine Ivanovic in the team of CRH's organizers as co-editors of these proceedings.

At Andrew Frank's suggestion, we selected "Time and Space Annotation" as the special topic for the Vienna edition of CRH. We believe the theme is aptly chosen, given the interest in the topic in the Viennese institutions and the international reputation of Andrew in the field. But in particular, we

believe that the topic suits the aim of our workshop perfectly, especially in a period when geodata are becoming easier to access and increasingly dominant in many disciplines, while many fields experience what it is sometimes referred to as a "geographic turn". Spatial information is often used as linking point between different data sources, and gazetteers are adopted in the context of Linked Open Data. While time gazetteer are arguably still less mature, the interest in solutions that could use a grid of time-space coordinates for comparable Linked Open Data approaches is constantly raising. Potentially, chronological and spatial information in large corpora is a subject where research in archaeology, history, computational linguistics as well as ontologies and the semantic web can fruitfully converge.

Our hopes were fulfilled by a number of submissions, much higher than we expected. This year, we received 54 long abstracts (up to 6 pages) from scholars of 20 different countries all over the world: Austria, Brazil, China, Czech Republic, France, Germany, Greece, Hungary, India, Italy, Lithuania, Moldova, Norway, Portugal, Romania, Russia, Spain, Taiwan, UK and USA. We accepted 25 proposals, which corresponds to an acceptance rate of 46.3. The authors of the accepted abstracts were invited to submit full papers (up to 10 pages), which are collected in these proceedings. In the program, 18 proposals were presented as talks in oral sessions, while 7 made the poster session of the workshop.

Each submitted abstract was reviewed in double-blind fashion by three members of a program committee consisting of 36 scholars from 11 countries.

The program was completed by two invited talks (whose abstracts are published here), which tackle the special topic of this edition of CRH from different perspectives. The contribution by James Pustejovsky (Brandeis University, USA) focuses on semantic data modeling for temporal and spatial information from multimodal corpora, while Tara Andrews (University of Vienna) discusses a number of challenges opened by time and place annotation of historical data.

Looking at the table of contents of these proceedings, it becomes obvious that "Time and Space" was a felicitous choice as a special topic of this edition of CRH. 10 papers out of 25 deal with issues related with time and/or spacial information in corpora. 4 out of them focus on annotation questions, namely those by Ainara Estarrona and Izaskun Aldezabal (*Towards a Spatial Annotation Scheme for Basque based on ISO-Space*), by Katharina Korecky-Kröll and Lisa Buchegger (*Tagging spatial and temporal PPs with two-way prepositions in adult-child and adult-adult conversation in German in Austria*), by Dmitri Sitchinava and Boris Orekhov (*The Poetic Corpus of Russian: Where the Poems are Written*) and by Matthias Lindemann and Thierry Declerck (*Annotation and Classification of Locations in Folktales*). As it can be seen, 2 of them deal with the topic in literary or narrative texts.

Annotation is strictly linked both with its exploitation and with tools for automatic processing of data.

As for the former, the paper by Marie Mikulová, Eduard Bejček and Jarmila Panevová (*What Can We Find Out about Time and Space in ForFun Database?*) makes use of time and space annotation for investigating some linguistic phenomena of Czech, while those by Maria Moritz (*Time Proximity as a Means to Align Spelling Variants in old English Bibles: A Case Study*) and by Jean-Baptiste Camps (*Manuscripts in Time and Space: Experiments in Scriptometrics on an Old French Corpus*) are good examples of the exploitation of such annotation in the philological area. The paper by Venumadhav Kattagoni and Navjyoti Singh (*Towards an unsupervised learning method to generate international political event data using spatio-temporal annotations*) makes use of metadata from time and space annotation for generating political events.

As for the latter, the papers by Adrien Barbaresi (*Towards a toolkit for toponym analysis in historical texts*) and by Delphine Bernhard, Pierre Magistry, Anne-Laure Ligozat and Sophie Rosset (*Resources and Methods for the Automatic Recognition of Place Names in Alsatian*) present practical applications of tools and methods for automatic place annotation.

After time and place annotation, the main topic of this edition of CRH (like for the previous ones) is linguistic annotation of (mostly historical) corpora, covering quite diverse issues, ranging from annotation of poetry or legal texts to questions of normalization and dialect.

Given the wide variety of approaches and perspectives represented in the proceedings, we think that the area dealing with the use of empirical evidence provided by corpora for research in the Humanities is lively and diverse. Such diversity can be at the same time a pro and a con.

On one side, it allows to join different competences and research objectives on common issues. On the other hand, it runs the risk of missing a distinctive identity, which is essential to move from being just an empirical methodology to becoming a clear-cut research field. In this respect, the core issue is understanding what we mean with "the Humanities", at least in the CRH context. A tentative answer can come from looking at the papers published in these proceedings, which mostly deal with peculiar kinds of textual data stored in corpora, deviating from "regular" collections of linguistic empirical evidence, which want to include (supposedly) representative selections of modern languages. The corpora concerned in CRH papers feature texts in ancient/dead languages or diachronic varieties of modern ones, they feature either prose or poetry literary texts, and they open different kinds of philological questions.

All this implies and involves a wide variety of new end-users both of the corpora themselves and of the results of the research work coming from their use. Users will no longer be only linguists and NLP professionals, but philologists, historical linguists, classicists and scholars in literature.

The dialogue between such actors is sometimes not straightforward. Still today, the Humanities suffer an unfortunate separation between the so-called "Traditional Humanities" and "Digital Humanities". Most likely, CRH would be considered a "Digital Humanities" event. But such separation is today simply meaningless. In a sense, all the Humanities are now (at least partly) digital and there is no research in the Humanities that is not (at least partly) traditional. These two sides of the same coin must collaborate, as the heritage of centuries of research in the Humanities can now be optimally exploited thanks to new technologies, methodologies and resources, among which are corpora. CRH wants to support and encourage such coming together of different research paradigms and, ultimately, render the distinction between "Digital" and "Traditional" Humanities superfluous and old-fashioned.

We hope you will enjoy the workshop and the proceedings. We wish to thank all authors who submitted papers, the members of the program committee, James Pustejovsky and Tara Andrews, the local organizers and in particular Hanno Biber, Andrew Frank and Christine Ivanovic, who made the Austrian edition of CRH possible.

The CRH Co-Chairs

Francesco Mambrini (Deutsches Archäologisches Institut, Berlin, Germany)

Marco Passarotti (Università Cattolica del Sacro Cuore, Milan, Italy)

Caroline Sporleder (University of Göttingen, Germany)

# Program Committee

## **Chairs:**

Francesco Mambrini (Deutsches Archäologisches Institut, Berlin, Germany)  
Marco Passarotti (Università Cattolica del Sacro Cuore, Milan, Italy)  
Caroline Sporleder (University of Göttingen, Germany)

## **Members:**

John A. Bateman (Germany)	Uwe Springmann (Germany)
Gerhard Budin (Austria)	Martin Thiering (Germany)
Giuseppe Celano (Germany)	Sara Tonelli (Italy)
Arianna Ciula (UK)	Martin Wynne (UK)
Giovanni Colavizza (Switzerland)	Amir Zeldes (USA)
Marco Coniglio (Germany)	
Maud Ehrmann (Switzerland)	
Andrew U. Frank (Austria)	
Emiliano Giovannetti (Italy)	
Stefan Th. Gries (USA)	
Dag Haug (Norway)	
Leif Isaksen (UK)	
Christine Ivanovic (Austria)	
Mike Kestemont (Belgium)	
Puneet Kishor (Germany)	
Dimitrios Kokkinakis (Sweden)	
Sandra Kübler (USA)	
Werner Kuhn (USA)	
Piroska Lendvai (Germany)	
Eleonora Litta (Italy)	
Yudong Liu (USA)	
Melanie Malzahn (Austria)	
Roland Meyer (Germany)	
Willard McCarty (UK)	
John Nerbonne (The Netherlands)	
Julianne Nyhan (UK)	
Michael Piotrowski (Switzerland)	
Geoffrey Rockwell (Canada)	
Matteo Romanello (Germany)	
Rainer Simon (Austria)	
Neel Smith (USA)	

## Organising Committee

### **Chairs:**

Francesco Mambrini (Deutsches Archäologisches Institut, Berlin, Germany)

Marco Passarotti (Università Cattolica del Sacro Cuore, Milan, Italy)

Caroline Sporleder (University of Göttingen, Germany)

### **Local Committee:**

Hanno Biber

Andreas Dittrich

Andrew U. Frank

Katharina Godler

Christine Ivanovic



# Contents

<b>Wait - When, Where? Difficult Answers to Simple Questions About Historical Time and Place</b>	
Tara Andrews	1
<b>Temporal and Spatial Data Modeling for Multimodal Corpora</b>	
James Pustejovsky	3
<b>Evaluating Part-of-Speech and Morphological Tagging for Humanities' Interpretation</b>	
Benedikt Adelmann, Melanie Andresen, Wolfgang Menzel and Heike Zinsmeister	5
<b>An Event Factuality Annotation Proposal for Basque</b>	
Begoña Altuna, María Jesús Aranzabe and Arantza Díaz de Ilarraza	15
<b>Placenames analysis in historical texts: tools, risks and side effects</b>	
Adrien Barbaresi	25
<b>Resources and Methods for the Automatic Recognition of Place Names in Alsatian</b>	
Delphine Bernhard, Pierre Magistry, Anne-Laure Ligozat and Sophie Rosset	35
<b>TEI and Bavarian dialect resources in Austria: updates from the DBÖ and WBÖ</b>	
Jack Bowers and Philipp Stöckle	45
<b>Manuscripts in Time and Space: Experiments in Scriptometrics on an Old French Corpus</b>	
Jean-Baptiste Camps	55
<b>Stemmatology: an R package for the computer-assisted analysis of textual traditions</b>	
Jean-Baptiste Camps and Florian Cafiero	65

<b>Towards a Spatial Annotation Scheme for Basque based on ISO-Space</b> Ainara Estarrona and Izaskun Aldezabal	75
<b>LitText: Realizing the "All Methods Applied to All Texts" Motto: Exploring a Corpus of Literary Text With SPARQL</b> Andrew U. Frank and Christine Ivanovic	85
<b>Incorporating Hittite into PROIEL: a pilot project</b> Guglielmo Inglese, Maria Molina and Hanne Eckhoff	95
<b>Towards an unsupervised learning method to generate international political event data with spatio-temporal annotations</b> VenuMadhav Kattagoni and Navjyoti Singh	105
<b>Tagging spatial and temporal PPs with two-way prepositions in adult-child and adult-adult conversation in German in Austria</b> Katharina Korecky-Kröll and Lisa Buchegger	113
<b>Annotation and Classification of Locations in Folktales</b> Matthias Lindemann, Stefan Grünewald and Thierry Declerck	123
<b>What Can We Find Out about Time and Space in the ForFun Database?</b> Marie Mikulová, Eduard Bejček and Jarmila Panevová	133
<b>On the Impact of Time Proximity on the Alignment of Spelling Variants in Old English Bibles: A Case Study</b> Maria Moritz	143
<b>A Toolkit for lemmatising, analysing, and visualising Middle English Data</b> Michael Percillier	153
<b>Word-level and higher level annotation of the Sardinian Medieval Corpus</b> Nicoletta Puddu and Achim Stein	161
<b>Marking Poetic Time: Building and Annotating a Hindi-Urdu Poetry Corpus for Computational Humanities Research</b> A. Sean Pue and Scott J. Nelson	171
<b>Who <i>came riding</i> first? Le chevalier or the knight? A multiple corpus analysis investigating historical language contact</b> Yela Schauwecker and Carola Trips	181

<b>The Opacity of Modal Verbs in German: An 'Optimal' Answer to a Difficult Question</b> Elisabeth Scherr	191
<b>The Poetic Corpus of Russian: Where the Poems are Written</b> Dmitri Sitchinava and Boris Orekhov	201
<b>The Use of Language Corpora to Process Particles in a Monolingual Dictionary</b> Barbora Štěpánková	207
<b>The European Union case law corpus (EUCLCORP) – a multilingual parallel and comparative corpus of EU court judgments</b> Aleksandar Trklja and Karen McAuliffe	217
<b>NORMO: An Automatic Normalization Tool for Middle Hungarian</b> Noémi Vadász and Eszter Simon	227
<b>A Study on Appraisal Resources in Argumentative Essays by non-English Major Postgraduates</b> Hongli Wang and Ying Chen	237



# Wait - When, Where? Difficult Answers to Simple Questions About Historical Time and Place

Tara Andrews  
Institut für Geschichte, Universität Wien, Austria  
E-mail: [tara.andrews@univie.ac.at](mailto:tara.andrews@univie.ac.at)

## Abstract

The study of history can be pithily summarised as an attempt to answer the ‘five W’ questions: who? what? when? where? and why? While the ‘what’ and the ‘why’ are disproportionately the focus of monographs, this talk will focus on two of the questions that are usually considered more straightforward—the When and the Where—with reference to current work on an annotated critical edition of the Armenian-language Chronicle of Matthew of Edessa, written in stages between c. 1102 and 1131, and covering the period from 952–1129.

The most well-known issue with historical time is that of a wide variation in the respective levels of precision and accuracy. While it is typical for database software to enforce the specification of a particular date or even a particular second in time, this kind of precision is extremely uncommon in historical texts. Chronicles will usually date an event to a particular year, sometimes to a particular month or day within that year, occasionally even to a particular hour. The author might alternatively use a dating that is not only vague but relative, with formulations such as ‘In this period’ that refer back to an event just described. In a similar vein, the text (especially when it has reached us in several versions) might give an ambiguous dating, or it may be known that the author’s given date is simply wrong. Although there have been several proposed or implemented means of handling precision of dates, such as the `notBefore`, `notAfter` and `confidence` attributes in TEI XML (TEI Consortium 2017), the visualisations of these via tools such as Topotime (Grossner and Meeks 2014), proposals for expression of dates as probability curves (Stokes 2015), and suggestions of standards for relative dating in CIDOC-CRM (Hiebel, Doerr, and Eide 2017), there has been less attention to how to represent the internal consistency or judged accuracy of dates expressed in text.

The issue of space presents deeper challenges yet. Here again the almost universal standard for specification of place, GIS, assumes—quite reasonably for most present-day applications—that geography can be treated as a set of specific points, connected into paths or polygons where appropriate. Above and beyond the issues of precision and accuracy of specification, we must contend with place

names that vary across language and across period as well as borders that are poorly defined in some periods and that shift in others—issues that are also of active interest to historians (Gregory and Hardie 2011; Seydi and Romanov 2017). At this point, however, the parallel to similar challenges for annotation of time begins to break down, as historians must contend not only with a single real-world concept of geography, but also with ideas of geography that can vary across cultural and ethnic groups, and even allegorical geography. The ‘Garden of Eden’ or the ‘River Gihon’ are toponyms that would easily be picked up by named entity recognition tools, and quite a few medieval people might happily have attempted to place them on a map. How do we represent these places in our annotations?

We can see, therefore, that the sheer complexity of the ideas of ‘time’ and ‘space’ require much more comprehensive digital models, and much more fine-grained support, than are currently available in the typical relational or NoSQL database, or commercial GIS systems. A fuller solution for the expression of time would allow us to re-orient a text around its internal timeline, compare this internal timeline to one or more externally-derived suggestions for the historical sequence of events described by the text, and allow for the comparison of these dating arguments. A fuller solution for the expression of place would allow for the creation of maps according not only to physical real-world geography, but also according to the conceptual geography of people who lived in the past—a combination of places that were experienced by those people, real places that were nevertheless beyond the bounds of their spatial awareness, and the allegorical places that may well have seemed, to them, to be in the same category.

## References

- Gregory, Ian N., and Andrew Hardie. 2011. “Visual GISting: Bringing Together Corpus Linguistics and Geographical Information Systems.” *Literary and Linguistic Computing* 26 (3):297–314. <https://doi.org/10.1093/lilc/fqr022>.
- Grossner, Karl, and Elijah Meeks. 2014. “Topotime: Representing Historical Temporality.” *Proceedings of DH2014*, Lusanne.
- Hiebel, Gerald, Martin Doerr, and Øyvind Eide. 2017. “CRMgeo: A Spatiotemporal Extension of CIDOC-CRM.” *International Journal on Digital Libraries* 18 (4):271–79. <https://doi.org/10.1007/s00799-016-0192-4>.
- Seydi, Masoumeh, and Maxim Romanov. 2017. “Modeling Regions from Premodern Geographical Hierarchical Data.” In . Rome. [https://www.academia.edu/32362252/Modeling\\_Regions\\_from\\_Premodern\\_Geographical\\_Hierarchical\\_Data](https://www.academia.edu/32362252/Modeling_Regions_from_Premodern_Geographical_Hierarchical_Data).
- Stokes, Peter Anthony. 2015. “The Problem of Digital Dating: A Model for Uncertainty in Medieval Documents.” In . Sydney. [http://dh2015.org/abstracts/xml/STOKES\\_Peter\\_Anthony\\_The\\_Problem\\_of\\_Digital\\_Datin/STOKES\\_Peter\\_Anthony\\_The\\_Problem\\_of\\_Digital\\_Dating\\_A\\_M.html](http://dh2015.org/abstracts/xml/STOKES_Peter_Anthony_The_Problem_of_Digital_Datin/STOKES_Peter_Anthony_The_Problem_of_Digital_Dating_A_M.html).
- TEI Consortium, ed. 2017. “Guidelines for Electronic Text Encoding and Interchange. Version 2.8.0.” <http://www.tei-c.org/p5/>.

# Temporal and Spatial Data Modeling for Multimodal Corpora

James Pustejovsky  
Department of Computer Science, Brandeis University, USA  
E-mail: [jamesp@brandeis.edu](mailto:jamesp@brandeis.edu)

## Abstract

This talk is about how to develop the most appropriate semantic model for a corpus. Specifically, I discuss the challenges encountered when identifying the conceptual inventory used for modeling the meaning of a specific language domain for an associated corpus. I address the importance of lexical semantic modeling in creating a coherent abstraction of a linguistic phenomenon. I will use two use cases that demonstrate the interplay between semantic analysis and corpus data when working towards defining a data model, namely the domains of temporal and spatial semantic modeling.

Drawing on experience from the creation, implementation, and adoption of TimeML and ISO-Space, I trace the design and development of both these specification languages, focusing on the demands imposed by specific linguistic corpora, and the ontological needs presented by actual language data. Such concerns are overlapping but not necessarily identical to theoretical problems encountered in these fields. The goal of lexical data modeling for corpora is to provide a coherent semantic analysis associated with the lexicon of the language, such that logical inferences and lexical semantic relations are derivable and automatically computable when performing analytics over the collection.





# Evaluating Part-of-Speech and Morphological Tagging for Humanities’ Interpretation

Benedikt Adelman<sup>\*</sup>, Melanie Andresen<sup>+</sup>,  
Wolfgang Menzel<sup>\*</sup> and Heike Zinsmeister<sup>+</sup>

<sup>\*</sup>Department of Informatics

<sup>+</sup>Institute for German Language  
and Literature

Universität Hamburg

E-mail:

{adelmann, menzel}@informatik.uni-hamburg.de  
{melanie.andresen, heike.zinsmeister}@uni-hamburg.de

## Abstract

Many digital humanities projects analyze non-standard texts and project settings require the researchers to rely on state-of-the-art analyzers. In this paper we report on the evaluation of part-of-speech and morphological tagging of German texts from different domains in the context of such a project. Our evaluation of overall tagging performance and features especially relevant for the project (here: tense) show a good performance of state-of-the-art tools and the potential of an ensemble approach.

## 1 Introduction

Many digital humanities (DH) projects analyze texts in non-standard language, but project settings require the researchers to rely on state-of-the-art analyzers instead of developing their own specialized tools, for instance because they lack appropriate training data, the language type under investigation is inherently heterogeneous (such as literary texts), or the text type is undetermined at the beginning of the research project and only evolves over time (as is the case with abductive methodologies). In this paper we report on the evaluation of part-of-speech (POS) and morphological tagging in different domains of German: academic language, modern and historic literary texts. This is important given that errors on these annotation levels will propagate to or at least influence downstream analyses. The evaluation is part of the DH project herma<sup>1</sup> that explores the potential of automated annotation for empirical research in the humanities and social sciences, particularly in use

---

<sup>1</sup> <https://www.herma.uni-hamburg.de/en.html>, all URLs in this paper were last accessed on December 15, 2017.

cases of literary studies, cultural anthropology and nursing science. Evaluations of the overall performance include many features that are of little interest to the humanist’s interpretation. We therefore also report results for a feature that can be translated into information directly relevant to the humanist scholar, namely tense.

In the next section, we introduce related work on the evaluation of morphological tagging. In the following sections we outline the evaluation study by presenting the taggers we use (3), the test data (4), and our results (5). We sketch our plans to improve the results with an active learning approach in section 6 and discuss our study’s implications for a DH setting in the final section.

## 2 Related Work

Morphological tagging of out-of-domain data has been evaluated with respect to diachronic variation [6] and text-genre variation (e. g. [16]). Most recently, an approach of character-level neural morphological tagging was evaluated in a transfer learning scheme, successfully predicting morphological taggings for high-resource languages and low-resource languages together [5].

Eger et al. (2016) [7] evaluate state-of-the-art taggers for German and Latin, using TGermaCorp [14] as out-of-domain data for German, which comprises literary texts. They report accuracies up to 0.90 for MarMoT (see below) when trained on the TIGER corpus and tested on TGermaCorp. In terms of morphology, they evaluate the ‘case’ feature, which they consider the most difficult one. From a DH practitioner’s point of view, however, case as such is of little interest for immediate analysis as it only becomes relevant in downstream applications, which is why we focus on the ‘tense’ feature instead. Tense is relevant, for example, to the analysis of temporal relations in narrative texts. Our tense feature is morphologically realized tense only (present and preterite), thus it does not cover compound tenses such as perfect, but is nonetheless needed to correctly analyze them (see [4] and [17] for this task).

## 3 POS and Morphological Taggers

In this study we evaluate morphological taggers for German that rely on different architectures and algorithms. In this way, we expect them to make different types of mistakes and hope to maximize the potential of an ensemble approach.

**MarMoT**<sup>2</sup> [15] is based on Conditional Random Fields. Morphological features are treated as independent variables. We use the pre-trained, first-order model available online, which has been trained on a subset of the TIGER 2 corpus [2] comprising 736,613 tokens (for the split see [10]).<sup>3</sup> TIGER is manually annotated with

---

<sup>2</sup> <http://cistern.cis.lmu.de/marmot/>

<sup>3</sup> A shell script that splits the TIGER corpus accordingly can be found here: <ftp://ftp.ims.uni-stuttgart.de/pub/corpora/TigerPreprocessing.tar.gz>

POS and morphological features from the STTS tagset [19].

**RFTagger**<sup>4</sup> [20] splits tags into part-of-speech and morphological attributes. Attribute values are determined using decision trees, while HMMs are employed to represent sequential information. We trained a model on the same TIGER split used for MarMoT above.

**HunPos**<sup>5</sup> [11] is a reimplementation and extension of the HMM trigram tagger TnT [3]. It treats POS tags and morphological features as an atomic unit. We trained it using default settings on the same training data used for the above taggers. For tagging, we used default settings as well.

**JWCDG**<sup>6</sup> [1] is a Java reimplementation of the weighted constraint dependency parser (WCDG). It consists of weighted hand-written constraints for morphology and dependency trees developed on the basis of the Hamburg Dependency Treebank [9]. Possible feature values come from a full-form lexicon containing suffix-based heuristics for unknown words. JWCDG has its own morphological feature representation, which we converted into the STTS format.<sup>7</sup> Instead of determining feature values directly, JWCDG merely establishes dependency relations which violate constraints (over these features) as little as possible. A subsequent feature unification along the established relations is required to achieve unambiguous values. In some cases, reduction to a unique value is impossible; we use the pseudo-value ‘\*’ in such cases.

**Ensemble:** In addition, we combine the individual taggers to an ensemble approach by majority vote. In the case of a draw, we accept the answer of a random tagger. For the morphological features, we only include votes by taggers that voted for the POS tag adopted by the ensemble.

## 4 Test Data

In the interdisciplinary project hermA we process texts with the ultimate goal of supporting text analysis in literary studies, cultural anthropology and nursing science. Therefore, we deal with many text types that differ from ‘standard’ language (i. e. newspaper language used for training) in several dimensions. In the case of our current test data the relevant dimensions are genre and time. The test data were provided by our project partners. In addition, we evaluated on two different newspaper texts.

---

<sup>4</sup> <http://www.cis.uni-muenchen.de/~schmid/tools/RFTagger/>

<sup>5</sup> <https://github.com/mivoq/hunpos>

<sup>6</sup> The CDG Team (1997–2015): <https://gitlab.com/nats/jwcdg>

<sup>7</sup> There is an unambiguous correspondence between them. JWCDG distinguishes subjunctive 1 and subjunctive 2. We map both to ‘subjunctive’ (subj), but distinguish them using the ‘tense’ feature (present for subjunctive 1 and past for subjunctive 2, in line with the indicative forms they are derived from).

text	tokens	sentences	mean sentence length	median sentence length	standard deviation	type- token ratio
Lit2009	1,518	114	13.32	12.0	8.66	0.46
Lit1850	1,647	82	20.10	18.5	11.56	0.46
Aca2009	1,790	75	24.00	22.0	18.42	0.44
News1999	1,735	93	18.67	18.0	10.64	0.48
TIGER	36,225	2,000	18.11	16.0	11.91	0.47

Table 1: Descriptive measures for the test texts (tokens incl. punctuation, TIGER data reduced to 120 sentences for TTR calculation)

- Modern literature (Lit2009): novel *Corpus Delicti: Ein Prozess* by German author Juli Zeh, published in Frankfurt/Main in 2009.
- Non-contemporary literature (Lit1850): *Eine Frauenfahrt um die Welt* (‘A woman’s journey around the world’) by Austrian author Ida Pfeiffer (1850).<sup>8</sup>
- Modern health science (Aca2009): *Stand, Möglichkeiten und Grenzen der Telemedizin in Deutschland* (‘Telemedicine in Germany: status, chances and limits’) by Rüdiger Klar and Ernst Pelikan, published in Bundesgesundheitsblatt (‘Federal Health Gazette’) in 2009.
- Newspaper (News1999): morphologically annotated text from the TüBa-D/Z treebank 9.1 as an almost in-domain text for the purpose of comparison: It is newspaper text, but from a different newspaper than the training data. The articles were published in 1999.
- Newspaper (TIGER): the first 2,000 sentences from the test set of the TIGER split by [10], i. e. in-domain data very similar to the training data, published between 1992 and 1997.<sup>9</sup>

The text excerpts we include comprise between 1500 and 1800 tokens each. Sentence segmentation was performed with NLTK’s implementation of PUNKT<sup>10</sup> [13] and manually corrected in very few cases. Table 1 shows some descriptive measures for the test texts. We can see that the newspaper data from News1999 and TIGER have a very similar sentence length. The sentences in both the historic literary and the academic text are considerably longer, which is a challenge for NLP tools. The modern literary text on the other hand has much shorter sentences. An inspection of the text shows that this is mainly due to an extensive use of elliptic sentences, which is also challenging for automatic (and manual) annotation.

<sup>8</sup> Available at Deutsches Textarchiv: [http://www.deutschestextarchiv.de/pfeiffer\\_frauenfahrt01\\_1850/6](http://www.deutschestextarchiv.de/pfeiffer_frauenfahrt01_1850/6).

<sup>9</sup> We did not use the whole test set as JWCDG’s annotation process is rather time-consuming.

<sup>10</sup> <http://www.nltk.org/api/nltk.tokenize.html>

text	pos only	morph only	pos+morph
Lit2009	0.974	0.910	0.931
Lit1850	0.965	0.853	0.891
Aca2009	0.950	0.867	0.887
mean	0.963	0.877	0.903

Table 2: Inter-annotator agreement measured in Fleiss’ Kappa

For the gold standard annotation we applied preprocessing and subsequent manual correction. In a training phase, three annotators with a linguistic background annotated five sentences of each text from scratch and discussed the results thoroughly. Afterwards, they independently corrected the rest of the texts pre-annotated by MarMoT.<sup>11</sup> Finally, a fourth annotation expert adjudicated on mismatches. For POS tags only, the annotators reached a very good inter-annotator agreement score of 0.963 (Fleiss’  $\kappa$  [8], see table 2). For morphological annotations, the score is 0.877.<sup>12</sup> We get the lowest morphology scores for the historic text, indicating specific morphological constellations the annotators were unfamiliar with, e. g. the grammatical gender of the word *Sauersop* (‘soursop’), which is not used in Modern Standard German.

## 5 Results

Our study focuses on whether the performance of state-of-the-art tools is sufficiently good to allow for the results to be used in DH studies. Therefore we first provide an overall evaluation of the accuracy of POS and of inflectional morphology tags on known and on unknown (i. e. ‘out-of-vocabulary’) tokens (5.1). In particular, we compare tagger performance over the different text types and evaluate tense as a specific feature relevant to humanists’ interpretation (5.2).

### 5.1 Overall tagging results

In table 3 we present an overview of the tagger accuracies for each of the five texts. As expected, we get the best results for the in-domain data from the TIGER corpus. However, the results for POS alone are very similar across the texts, ranging from 0.96 to 0.97. The differences become more pronounced when we include morphology (POS+MORPH): All texts get much lower scores than the TIGER data, even the almost in-domain News1999. For the task of determining POS of out-of-vocabulary (OOV) words, the scores of the two literary texts are especially low. Opaque morphology and deviations in syntax are possible explanations of

<sup>11</sup> Note that this gives MarMoT a slight advantage in the evaluation as it is likely that the annotators did not correct all errors and might simply agree with the tagger in doubtful cases.

<sup>12</sup> For this calculation, we regarded all morphological features as one unit. Consequently, with a deviation in one morphological tag the whole token is considered a mismatch.

this fact. In the historical text, for instance, many of the OOV words deviate from modern spelling (e. g. *getheilt* instead of modern *geteilt* ‘divided’).

For evaluating the accuracy of the morphological features we focus on those tokens that are assigned a correct POS ( $\text{POS}_{\text{true}}$ ). The morphology is mostly correct on the TIGER data, while morphological analysis is still challenging on the out-of-domain texts (including News1999).

When comparing the performance of the taggers, MarMoT achieves the highest accuracies in the vast majority of cases. However, we have to bear in mind that our gold standard is a manual correction of MarMoT’s output and hence the data is probably slightly biased. Therefore, we additionally marked in italics the best results excluding MarMoT. In most cases, the results of the ensemble approach are the best or second-best after MarMoT, which is encouraging as it shows that the taggers have different weaknesses that can be balanced even in a simple majority vote ensemble. The only rule-based system, JWCDG, generally achieves results slightly below the best result, but frequently outperforms HunPos and the RFTagger.

One would expect the statistical taggers modeling POS and morphological features in a compositional way (MarMoT, RFTagger) to outperform HunPos, which treats them as an atomic unit. Surprisingly, HunPos surpasses RFTagger instead by up to four percentage points for POS (e. g. Lit2009) and up to two percentage points for POS+MORPH. However, the relation is inverse for the morphology of words the taggers assign correct POS tags (MORPH of  $\text{POS}_{\text{true}}$ ): Here, RFTagger’s accuracies are up to two percentage points better than those of HunPos. We interpret this finding as HunPos being relatively good at narrowing the set of possible tags (which is vast when treating POS plus morphology as atomic) to a set of likely candidates (that share the same correct POS tag), but RFTagger being better at choosing the correct candidate due to the availability of more useful information.

The ensemble is very robust in both determining POS and morphology. Among all taggers except MarMoT, it yields the best accuracies (with one exception: POS of Lit2009, where both JWCDG and MarMoT are better by one percentage point). We thus consider combining taggers into an ensemble an advisable approach.

## 5.2 Task-specific evaluation: Tense

For scholars from the humanities, not all POS and their morphological features are equally relevant. We thus also report a task-specific evaluation that focuses on specific features. One feature that allows conclusions about e. g. narrative patterns in literature is tense (see e. g. [4]). Table 4 shows the performance of the taggers with respect to this feature (regardless of whether the POS tag is correct or not).<sup>13</sup>

MarMoT achieves the best results for News1999 and TIGER, the newspaper texts, with the ensemble’s results being the second best. For the out-of-domain texts, the ensemble is superior. In two cases, JWCDG performs as well as the en-

<sup>13</sup> Note that the relation pres:past varies considerably: Lit2009: 137:10, Lit1850: 60:128, Aca2009: 94:10, News1999: 85:44, TIGER: 4898:2242.

Text	Tagger	POS+ MORPH	POS	POS <sub>oov</sub>	MORPH of POS <sub>true</sub>	MORPH of POS <sub>oov, true</sub>
<b>Lit2009</b>	MarMoT	<b>0.89</b>	<b>0.97</b>	<b>0.88</b>	<b>0.95</b>	<b>0.93</b>
	RFTagger	0.80	0.91	0.79	0.92	0.89
	HunPos	0.82	0.95	<i>0.83</i>	0.91	0.86
	JWCDG	0.87	<b>0.97</b>	0.81	0.94	<i>0.90</i>
	mean	0.84	0.95	0.83	0.93	0.90
	ensemble	<i>0.88</i>	0.96	—	<b>0.95</b>	—
<b>Lit1850</b>	MarMoT	<b>0.86</b>	<b>0.96</b>	<b>0.88</b>	<b>0.93</b>	<b>0.89</b>
	RFTagger	0.78	0.90	0.77	0.91	<b>0.89</b>
	HunPos	0.79	0.94	0.80	0.89	0.82
	JWCDG	0.82	0.94	<i>0.82</i>	0.91	0.83
	mean	0.81	0.94	0.82	0.91	0.86
	ensemble	<b>0.86</b>	<b>0.96</b>	—	<b>0.93</b>	—
<b>Aca2009</b>	MarMoT	<b>0.86</b>	<b>0.96</b>	<b>0.93</b>	<b>0.94</b>	<b>0.90</b>
	RFTagger	0.80	0.94	<i>0.90</i>	0.91	0.88
	HunPos	0.80	0.95	<i>0.90</i>	0.90	<i>0.89</i>
	JWCDG	0.82	0.94	0.73	0.91	0.77
	mean	0.82	0.95	0.86	0.92	0.86
	ensemble	<i>0.85</i>	<b>0.96</b>	—	<i>0.93</i>	—
<b>News1999</b>	MarMoT	0.80	0.95	<b>0.92</b>	0.91	<b>0.87</b>
	RFTagger	0.75	0.93	<i>0.89</i>	0.88	<i>0.82</i>
	HunPos	0.75	0.94	0.84	0.87	0.80
	JWCDG	0.79	<b>0.96</b>	0.86	0.89	0.64
	mean	0.77	0.94	0.88	0.89	0.78
	ensemble	<b>0.86</b>	<b>0.96</b>	—	<b>0.94</b>	—
<b>TIGER</b>	MarMoT	<b>0.98</b>	<b>0.99</b>	<b>0.90</b>	<b>0.99</b>	<b>1.00</b>
	RFTagger	0.82	0.94	0.86	0.92	<i>0.87</i>
	HunPos	0.83	0.96	<i>0.88</i>	0.91	0.83
	JWCDG	0.80	0.95	0.83	0.89	0.68
	mean	0.86	0.96	0.87	0.93	0.84
	ensemble	<i>0.90</i>	<i>0.97</i>	—	<i>0.96</i>	—

Table 3: Detailed tagging results for the test texts: atomic tag (POS+MORPH), POS only, POS of out-of-vocabulary tokens (POS<sub>oov</sub>), morphology of correct parts of speech (MORPH of POS<sub>true</sub>), morphology of correct out-of-vocabulary part of speech (MORPH of POS<sub>oov, true</sub>). Bold indicates best values per text and column. Italics indicate best non-MarMoT values per text and column. Note that for the ensemble, dividing into in-vocabulary and out-of-vocabulary tokens is not possible as JWCDG’s vocabulary differs from that of the training set used for the other taggers.

	Lit2009		Lit1850		Aca2009		News1999		TIGER		mean
	past	pres	past	pres	past	pres	past	pres	past	pres	
MarMoT	<b>1.00</b>	0.97	0.98	0.87	0.80	0.94	<b>0.97</b>	0.93	<b>1.00</b>	<b>1.00</b>	0.95
RFTagger	0.82	0.96	0.97	0.86	0.80	<b>0.95</b>	0.95	<b>0.94</b>	0.96	0.94	0.91
HunPos	0.95	0.96	0.97	0.90	0.80	0.92	0.95	0.92	0.96	0.94	0.93
JWCDG	0.89	<b>0.98</b>	0.89	0.79	0.75	<b>0.95</b>	0.60	0.85	0.89	0.93	0.85
mean	0.92	0.97	0.95	0.86	0.79	0.94	0.87	0.91	0.95	0.95	0.91
ensemble	<b>1.00</b>	<b>0.98</b>	<b>0.99</b>	<b>0.95</b>	<b>0.89</b>	<b>0.95</b>	0.96	0.93	0.99	0.96	<b>0.96</b>

Table 4: F-scores for the correct detection of tense features. Bold indicates best values per column. Italics indicate best non-MarMoT values per column.

semble. The ensemble’s average F-score is best, too. In a nutshell, the ensemble can robustly determine morphological tense with high precision and recall.

A closer look at the errors made by the taggers reveals certain patterns: The ambiguity between finite 1st/3rd person plural present and infinitive form leads to errors, e. g. in the case of the tokens *treffen* ‘(we/they/to) meet’ and *stattfinden* ‘(they/to) take place’. This type of error is not specific to text type, but results from a systematic word form ambiguity in German. Sometimes there is also ambiguity between finite verb and participle (possibly used as an adjective), as in the case of the token *bedeckt* ‘cover/covered’.

In summary, the results of the state-of-the-art taggers on different textual domains are quite good and can be considered a reasonable basis, for instance, for a distant reading approach.

## 6 Future Work

In order to further improve our annotations within a reasonable amount of time, we currently implement the active learning approach proposed by [18], based on [12]. The approach centers around two ideas: First, if we have several taggers producing different errors, an algorithm can learn when to trust which of the taggers and give a more precise prediction than a simple majority vote. We trained an unsupervised algorithm on the full version of text Lit2009 (58,805 tokens) and achieved an accuracy of 0.97 for POS only on the test set, which is superior to the majority vote. Second, the accuracy can be boosted further by having humans annotate a few instances doubtful to the algorithm. In the future, we will further explore the potential of this approach.

## 7 Conclusion

Our analyses show that off-the-shelf morphological taggers for German offer reasonably good tagging performance even when applied to out-of-domain data. An



ensemble approach based on majority vote resulted in an average accuracy of 0.87 across text types (0.96 for POS only and 0.94 for morphological features of correctly POS-tagged tokens). A detailed analysis of tense features that are relevant for text interpretation in DH shows that the ensemble yields very robust results (F-score: 0.96) and outperforms other taggers. As a next step we will experiment with an active learning approach and explore the optimization potential of the statistical taggers.

## 8 Acknowledgements

This work has been funded by the ‘Landesforschungsförderung Hamburg’ in the context of the *hermA* project (LFF-FV 35). We would like to thank Lea Röseler, Henny Sluyter-Gäthje and Sarah Wennefehr for contributing to the manual annotation and Piklu Gupta for checking our English. All remaining errors are our own.

## References

- [1] Niels Beuck, Arne Köhn, and Wolfgang Menzel. Predictive incremental parsing and its evaluation. In Kim Gerdes, Eva Hajičová, and Leo Wanner, editors, *Computational Dependency Theory*, pages 186–206. IOS press, 2013.
- [2] Sabine Brants, Stefanie Dipper, Peter Eisenberg, Silvia Hansen-Schirra, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith, and Hans Uszkoreit. TIGER: Linguistic interpretation of a German corpus. *Research on language and computation*, 2(4):597–620, 2004.
- [3] Thorsten Brants. TnT – A Statistical Part-of-Speech Tagger. In *Proceedings of ANLP 2000*, Seattle, Washington, 2000.
- [4] Thomas Bögel, Jannik Strötgen, and Michael Gertz. Computational Narratology: Extracting Tense Clusters from Narrative Texts. In *Proceedings of LREC 2014*, pages 950–955, Reykjavik, Iceland, 2014.
- [5] Ryan Cotterell and Georg Heigold. Cross-lingual, Character-Level Neural Morphological Tagging. *arXiv preprint arXiv:1708.09157*, 2017.
- [6] Stefanie Dipper. Morphological and Part-of-Speech Tagging of Historical Language Data: A Comparison. *JLCL*, 26(2):25–37, 2011.
- [7] Steffen Eger, Rüdiger Gleim, and Alexander Mehler. Lemmatization and Morphological Tagging in German and Latin: A Comparison and a Survey of the State-of-the-art. In *Proceedings of LREC 2016*, pages 1507–1513, Portorož, Slovenia, 2016.
- [8] Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971.

- [9] Kilian Foth, Arne Köhn, Niels Beuck, and Wolfgang Menzel. Because Size Does Matter: The Hamburg Dependency Treebank. In *Proceedings of LREC 2014*, pages 2326–2333, Reykjavik, Iceland, 2014.
- [10] Alexander Fraser, Helmut Schmid, Richard Farkas, Renjing Wang, and Hinrich Schütze. Knowledge Sources for Constituent Parsing of German, a Morphologically Rich and Less-Configurational Language. *Computational Linguistics*, 39(1):57–85, 2013.
- [11] Péter Halácsy, András Kornai, and Csaba Oravecz. HunPos: an open source trigram tagger. In *Proceedings of ACL 2007 on interactive poster and demonstration sessions*, pages 209–212, Prague, Czech Republic, 2007.
- [12] Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. Learning Whom to Trust with MACE. In *Proceedings of NAACL-HLT 2013*, pages 1120–1130, Atlanta, Georgia, 2013.
- [13] Tibor Kiss and Jan Strunk. Unsupervised Multilingual Sentence Boundary Detection. *Computational Linguistics*, 32(4):485–525, 2006.
- [14] Andy Luecking, Armin Hoenen, and Alexander Mehler. TGermaCorp – A (Digital) Humanities Resource for (Computational) Linguistics. In *Proceedings of LREC 2016*, Paris, France, 2016.
- [15] Thomas Müller, Helmut Schmid, and Hinrich Schütze. Efficient higher-order CRFs for morphological tagging. In *Proceedings of EMNLP 2013*, pages 322–332, Seattle, Washington, 2013.
- [16] Thomas Müller and Hinrich Schütze. Robust Morphological Tagging with Word Representations. In *Proceedings of NAACL-HLT 2015*, pages 526–536, Denver, Colorado, 2015.
- [17] Anita Ramm, Sharid Loáiciga, Annemarie Friedrich, and Alexander Fraser. Annotating tense, mood and voice for English, French and German. *Proceedings of ACL 2017, System Demonstrations*, pages 1–6, 2017.
- [18] Ines Rehbein and Josef Ruppenhofer. Detecting annotation noise in automatically labelled data. In *Proceedings of ACL (Volume 1: Long Papers)*, pages 1160–1170, Vancouver, Canada, 2017.
- [19] Anne Schiller, Simone Teufel, Christine Stöckert, and Christine Thielen. Guidelines für das Tagging deutscher Textcorpora mit STTS. Technical report, Institut für maschinelle Sprachverarbeitung/Seminar für Sprachwissenschaft, Stuttgart/Tübingen, 1999.
- [20] Helmut Schmid and Florian Laws. Estimation of Conditional Probabilities With Decision Trees and an Application to Fine-Grained POS Tagging. In *Proceedings of COLING 2008*, pages 777–784, Manchester, UK, 2008.

# An Event Factuality Annotation Proposal for Basque

Begoña Altuna      María Jesús Aranzabe

Arantza Díaz de Ilarraza

University of the Basque Country

E-mail: begona.altuna;maxux.aranzabe;a.diazdeillaraza@ehu.eus

## Abstract

Factuality information gives evidence on whether the events in texts have happened. This information can be relevant in natural language processing tasks such as timeline generation as it helps discriminating the events that are relevant to a certain timeline. We analysed some factuality annotation schemes and proposed a new scheme that aims at concise and easy annotation. We worked on a Basque corpus for the creation of the factuality annotation scheme as an additional layer to temporal information and we evaluated our annotation decisions through an inter-annotator agreement experiment.

## 1 Introduction

Temporal information plays a crucial role in the structuring of the information in text since it allows placing the events along a temporal axis, commonly known as a “timeline”. The automatic creation of those is our final goal. For this, the events and the time points and intervals in text have to be identified, as well as identifying whether those events have happened, as only events that have happened are to be displayed in the timeline. That is what is called “factuality”, that is to say “whether events mentioned in text correspond to real situations in the world or, instead, to situations of uncertain status” [9].

In this paper we present a proposal for factuality annotation for events and a manual annotation effort to evaluate the annotation decisions. We aimed at a scheme as simplified as possible, but without compromising too much information. For that we have examined previous factuality schemes and we have built a proposal that has been proved in a corpus of news documents in Basque.

## 2 Event factuality and its classifications

Event factuality is described in [8] as “the level of information expressing the factual nature of eventualities mentioned in text”, that is, whether events correspond

to a fact in the world, a possibility or a situation that does not hold. For example, *prompted* in example (1) is a fact in the world that has happened.

- (1) President Donald Trump’s move **prompted** international criticism.

In order to classify events according to the factuality they express, many factuality classifications have been defined. After analysing those, a basic division can be made: i) facts, situations that hold in the world, ii) counterfactuals, situations that do not hold in the world and iii) a wide spectrum of uncertain or undefined values of factuality. A summary of the different factuality feature proposals is given in the following lines and Table 1<sup>1</sup>.

First, a bi-dimensional deterministic scale of factuality degrees depending on their certainty and their polarity is offered in [8] for FactBank, a corpus that contains factuality. From the combination of these two features, 8 values for factuality were defined: factual, counterfactual, probable, not probable, possible, not possible, certain but unknown output and unknown or uncommitted.

On its part, SIBILA [11] is an annotation scheme for temporal information that was explicitly focused on event factuality, what they called *factivity*. Although they admitted factivity was closely related to tense, polarity and modality, they were aware of the association not being automatic. As a consequence, they disagree with [8], who claims for a deterministic model to assign factuality values.

More recent works on factuality annotation include the approaches of [10] and [5]. The first [10] analysed event factuality and sentiments for the extraction and interpretation of perspectives expressed in news texts, in order to divide the information into positive and negative views on the actual or future world. For factuality, they got inspiration on the FactBank annotation scheme, although some conceptualisation changes were done; namely making a clear distinction between past and present events and future events, as these will always convey a certain amount of uncertainty. Factuality was described as “the level of information expressing the commitment of relevant sources towards the factual nature of events mentioned in discourse”. They proposed a four-value factuality classification which was built on three axes: polarity, certainty and temporality. Temporality was added as their corpus did not have previous temporality annotation.

The factuality annotation scheme in [5] is included in the NewsReader project framework<sup>2</sup>. The factuality annotation was done on a temporal information annotation, similar to TimeML, and was inspired on [10]. They followed the aforementioned when they proposed factuality values that were determined by time, polarity and certainty. Nevertheless, they were aware of some special cases, such as the the hypothetical events in conditionals and the general statements that are not anchored in time for they are ever-present situations. These special cases were explicitly annotated by means of a dedicated attribute.

---

<sup>1</sup>The features between parentheses respond to indirectly considered features.

<sup>2</sup><http://www.newsreader-project.eu/>

Table 1: Factuality features in different factuality annotation schemes

Features	FactBank [8]	SIBILA [11]	van Son et al. [10]	NewsReader [5]
Polarity	✓	(✓)	✓	✓
Certainty	✓	(✓)	✓	✓
Temporality	(✓)	(✓)	✓	✓
Special Cases				✓
Factuality	Factual Probable Possible Counterfactual Not probable Not possible Certain but unknown output Unknown or uncommitted	Yes No Programmed_future Negated_future Possible Indefinite	Fact Counterfact Possibility (uncertain) Possibility (future)	Factual Counterfactual Non factual

### 3 Event factuality in Basque

We aim to create a factuality annotation scheme that will assign factuality values to the events defined in Basque. As being previously done for other languages, Basque has a corpus annotated with temporal information, EusTimeBank, and the factuality annotation has been conducted on this previously annotated corpus. In this section we present the temporal mark-up scheme for Basque, EusTimeML [3], the EusTimeBank corpus, our proposal for factuality annotation and the manual annotation effort we conducted in order to achieve a robust factuality mark-up scheme and an annotated corpus.

However, although our working language is Basque and that we have built our factuality annotation scheme on top of a temporal information annotation scheme, we aim at offering a universal factuality scheme, as, whereas the expressions of factuality vary among languages, the factuality information remains the same.

#### 3.1 EusTimeML and EusTimeBank

As mentioned before, temporal information comprehends information about the events and the time points and intervals, as well as the relations that are created among those. In order to normalise and make that information machine readable, EusTimeML, a mark-up language for temporal information inspired in TimeML [6] has been developed for Basque.

EusTimeML is a mark-up scheme that offers XML tags for events, time expressions and temporal relation signals as well as tags for temporal, subordination and aspectual relations. An example of an annotation using EusTimeML can be seen in Figure 1. In this example the event *fakturatu zituen* (“turned over”) and the time expression *iaz* (“last year”) which refers to 2016 are displayed and their main attributes represented. The temporal relation between those is also represented as an inclusion relation, that is to say: the turning over happened in 2016.

Nonetheless, the mere temporal information is not enough when developing more complex tools. In our case, in order to build timelines, we considered adding

Figure 1: Temporal information in *laz 1.167 milioi euro fakturatu zituen* (“Last year 1,167 million euros were turned over”)

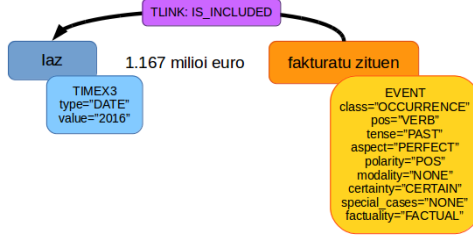


Table 2: Proposed factuality annotation scheme for Basque

FACTUALITY RELATED ATTRIBUTES		FACTUALITY VALUE
CERTAINTY	SPECIAL CASES	
Certain	Cond. Condition	Factual
Uncertain	Cond. Main clause	Counterfactual
Underspecified	Generic Statement	Non factual
	None	No factuality value
		Underspecified

factuality information to the temporal information in EusTimeML, since we reckoned that only factual events should appear in historical chronologies, for example. As a consequence, we have developed a mark-up scheme for factuality annotation in Basque (Table 2) and we have integrated it in the EusTimeML guidelines.

To conduct our analysis and experimentation on factuality, we used a section of the EusTimeBank corpus, a continuously growing economy news corpus in Basque. Nowadays, it contains 75 documents manually annotated following EusTimeML. It has been employed in temporal information annotation tasks such as guideline validation [1, 2]. It has also been used for the analysis of the negation in Basque [4], which is a crucial step towards factuality annotation. Finally, some of its documents have formed the training and evaluation sets for *bTime* [7], a tool for temporal information annotation in Basque. For this experiment, a section of 15 documents (3,463 tokens) has been manually annotated. Document length was 13 sentences in average and they contained an average of 231 tokens.

### 3.2 Factuality annotation proposal for Basque

After analysing the state of the art proposals for factuality annotation, we opted for a simple scheme in order to ease the burden of manual annotation. However, we did not want to sacrifice much information. Thus, we defined the scheme shown in Table 2. *Certainty*, as well as *polarity* and temporality (verb *tense* and *aspect*), are widely considered factuality features as they convey the majority of factual information, while identifying *special cases* adds relevant information to factuality resolution. That is the reason for adopting it from [5].

As can be seen in the table, we represent factuality through five factuality val-

ues: **FACTUAL** for events that have happened, **COUNTERFACTUAL** for events that have not happened in the past, **NON\_ FACTUAL** for future events, **UNDERSPECIFIED** for those events of which the factuality value cannot be assessed and **NO\_FACTUALITY\_VALUE** for the events that do not express any specific event. These values are conditioned by the values of the factuality related attributes.

First, *certainty* expresses the commitment of the source with the information expressed. We have considered that, unless there is an explicit uncertainty marker or it is impossible to give a certainty value, we will consider the events certain (2). The uncertainty particle *ote* in example (3) marks the uncommitment of the utterer for the certainty of *galdu* (“to loose”).

- (2) Boeing-ek 11.000 milioi dolar **lortu ditu** akordioetan.  
Boeing.ERG 11,000 million dollar **obtained has** agreement.PL.INE.  
‘Boeing **has obtained** 11,000 million dollars in agreements.’
- (3) Hegazkin-merkatuaren kontrola galdu **ote** zuen  
Airplane-market.GEN control.ABS loose **UNCERT.PART** AUX  
eztabaida piztu zen.  
discussion light AUX.  
‘Discussion on **whether** (it) had lost control over the aeroplane market was started.’

In what concerns the *special cases*, we wanted to emphasize the effects of conditionals and generic statements. For example, when using the hypothetical tense like in “If only I had come...”, although the verb has a positive polarity, humans know that the utterer has not come. In (4) *ematen badu* (“if it gives”) in the protasis is marked as **CONDITIONAL\_CONDITION** while *bilatuko du* (“will look for”) in the apodosis is marked as **CONDITIONAL\_MAIN**. The specific mark for generic statements express that those events do not refer to a specific event in a specific time and place. Such is the case of *da* (“is”) in (5).

- (4) Bilaketak fruiturik *ematen ez badu*, BEAk jarraitzeko  
Search.ERG results.PART *bring* no AUX, ANR.ERG continue.FIN  
dirua **bilatuko du**.  
money.ABS **look.for** AUX.  
‘If the search *brings* no results, the ANR **will look for** money to continue.’
- (5) Airbus A320a korridore bakarreko hegazkina **da**.  
Airbus A320.ABS aisle single.REL aeroplane **is**.  
‘The Airbus A320 **is** a single aisle aeroplane.’

As we conceived factuality as an additional layer for the temporal information annotation, we took into account previous annotation when deciding the factuality values. This is the case of *polarity*—whether the events appear affirmed (6) or negated (7)—as this feature plays a crucial role when defining the factuality value of an event.

- (6) Boeing-ek 11.000 milioi dolar **lortu** **ditu** akordioetan.  
Boeing.ERG 11,000 million dollar **obtained has** agreements.INE.  
'Boeing **has obtained** 11,000 million dollars in agreements.'
- (7) Hegazkin erraldoiak ez **du** Malasian **lur hartuko**.  
Aeroplane giant *no* **AUX** Malaysia.INE **earth take.FUT**.  
'The giant aeroplane **will not land** in Malaysia.'

In the case of temporality, one may notice that verbal events convey much more information about factuality as many of them have *aspect* and *tense* features. In fact, aspect in Basque expresses whether the verb is perfect or refers to a future action. Basque tense system, on its part, has three main values: past, present and hypothetical. In example (8) the tensed verb *erabaki dute* ("they have decided") is a factual event as there is no negation element nor any uncertainty marker and the aspect and tense and aspect suggest the event has already happened. In example (9), instead, *bidaliko dizkie* ("He/she will send (them to them)") has future aspect and present tense and, thus it is a non-factual event that has not happened yet.

- (8) Txinako Herri Errepublikako agintariak Boeing 787  
China.REL People Republic.REL leaders.ERG Boeing 787  
Dreamlinerra erostea **erabaki dute**.  
Dreamliner.ABS buy **decided have**.  
'People's Republic of China leaders **have decided** to buy the Boeing 787 Dreamliner.'
- (9) Dreamlinerrak sei aerolineari **bidaliko dizkie**.  
Dreamliners.ABS six airline.DAT **send.FUT AUX**.  
'(He/she) will send the Dreamliners to six airlines.'

Nonetheless, we wanted to provide all the events with a factuality value, so the factuality annotation of non-verbal events was conditioned by the factuality features of the verbal event that accompanies them. In complex event structures in which a noun bears the semantics of the event and the verb adds the grammar information, both elements get the same factuality values as they refer to a single event (10). In this case *baimena eman zaio* ("has been given permission") is factual and *parte hartzeko* ("to take part") is an underspecified event as one cannot say whether Boeing has taken part or will take part. For some other non-verbal events instead, we have had to rely more on the context. In the case of example (11), we know *aukeraketa-prozesu* ("selection process") is a fact as it is anchored in a certain date, 1998.

- (10) Boeing-i **baimena eman zaio** leihaketetan **parte hartzeko**.  
Boeing.DAT **permission give have** biddings.INE **part take**  
'Boeing **has been given permission** to **take part** in biddings.'



Table 3: Annotation of *lehen hegaldia* (13) according to the different schemes

EusTimeML	FactBank [8]	SIBILA [11]	van Son et al. [10]	NewsReader [5]
Positive Certain No special case	Positive	Positive	Positive Uncertain No special case	Positive Certain Future
Non factual	Probable	Programmed future	Possibility- future	Non factual

- (11) 1998ko **beste aukeraketa-prozesu batean** Boeing-en leihakidea  
1998.REL **other selection-process** **one**.INE Boeing.GEN rival  
izan zen Lockheed Martin.  
be AUX Lockheed Martin.  
‘Lockheed Martin was Boeing’s rival **in another selection process** in 1998.’

Finally, it should be highlighted that the annotators should rely on the semantics of the events and the world-knowledge to give the appropriate factuality value. *Onartzen du* (“admits”) in example (12) conditions the factuality value of *zutela* (“had”) as the admission of the fact expresses the commitment of the utterer with the truth value of the event. That is to say, we know *zutela* is a counterfactual event as it is in the past tense, it is negated, it is not one of the aforementioned special cases and the utterer considers it is certain.

- (12) *Onartzen du ez zutela* nahikoa gaitasun tekniko.  
*Admit AUX no have.PAST.COMP* enough skill technical.  
‘(He/she) *admits* they **did** not **have** enough technical skills.’

### 3.3 Adequacy and adaptability of the factuality scheme

In Table 3 we compare our annotation of factuality of the event *lehen hegaldia* (“first flight”) in example (13) to the annotation following other schemes.

- (13) **Lehen hegaldia** martxoaren hasierarako programatu da.  
**First flight** March.GEN beginning.ADL.REL programmed is.  
‘**The first flight** has been programmed for the beginning of March’

As can be seen from the table, some schemes are more descriptive than others. This might be caused by the fact that some annotation efforts like SIBILA are strongly integrated in a more comprehensive event annotation scheme. The main difference appears regarding to the certainty. In [11], [5] and EusTimeML the explicit presence of “programmed” justifies the absolute certainty of the event, whereas in [8] and [10], all future events convey certain amounts of uncertainty.

As mentioned before, we consider our scheme is suitable for the factuality annotation in other languages. Examples (14) and (15) and Table 4 represent the annotation of an English and a Spanish event (in bold) according to our guidelines.

Table 4: Factuality annotation of the events in examples (14) and (15).

Says/dice	Earn/ganan
Positive	Positive
Present tense	Present tense
Imperfect aspect	Imperfect aspect
Certain	Certain
No special cases	No special cases
Factual	Factual

- (14) Shell **says** male staff working for the company on average **earn** 22% more than women in the UK.
- (15) Shell **dice** que sus trabajadores masculinos **ganan** de media un 22% más en el Reino Unido.  
(Shell **says** that its male workers **earn** on average a 22% more in the United Kingdom).

### 3.4 Manual annotation

Two annotators took part in this experiment. They were asked to fully annotate the events so as to use the EusTimeML information to determine the factuality value. They were also asked to use world knowledge to resolve factuality. In total 734 events (out of 787 or 818) were annotated by both annotators and the factuality referring attributes in the agreed ones were analysed.

Table 5 shows the accuracy and  $\kappa$  values for the attributes that convey factuality information. As one can see, accuracy is rather high for most of the attributes and  $\kappa$  shows also a high agreement. The lower  $\kappa$  values are a consequence of a large quantity of certain categories. In fact, some values, such as the *certain* or *factual* values for certainty and factuality are very frequent in our corpus since news narratives tend to represent facts and this conditions the  $\kappa$  values.

Table 5: Inter-annotator agreement results for factuality annotation

	<b>Polarity</b>	<b>Certainty</b>	<b>Special Cases</b>	<b>Factuality</b>
<b>Accuracy</b>	0.98	0.89	0.95	0.77
$\kappa$	0.68	0.24	0.29	0.53

Analysing the disagreement has given us better knowledge about factuality annotation. Most of the mistakes were due to too loose definitions of the guidelines and were corrected in a guideline discussion session. In addition, we expect that i) redefining the `UNCERTAIN` and `UNDERSPECIFIED` values for certainty, ii) defining the boundaries of the generic statement and iii) better analysing the focus of the negation will help us define more accurate guidelines.

- (16) 20 milioi dolar arteko laguntza **emateko** prest dago.  
 20 million dollar until.REL help **give** ready is.  
 ‘(It) is ready to **give** up to 20 million dollar help.’

To illustrate this, *emateko* (“to give”) in example (16) has been assigned **UNCERTAIN** and **UNDERSPECIFIED** by the annotators. It is stated in the guidelines that the events that express an aim will condition the certainty value of the subordinated event (**UNCERTAIN**). Nonetheless, *prest dago* (“is ready”) is not a clear volition expression and was wrongly annotated by one of the annotators.

## 4 Conclusions

In this paper we have presented an event factuality annotation proposal. We aimed to reflect the factuality information concisely, while we wanted to create an easy-to-employ scheme to make the annotation effort easier. We have also evaluated our annotation decisions through an inter-annotator agreement experiment.

We attempted to offer a comprehensive factuality annotation scheme built as an additional layer to temporal information annotation. We have also compared our annotation scheme to other factuality annotation schemes, so as to highlight the differences between them. Finally we have also proved that our scheme is suitable for other languages (English and Spanish) even though it was modelled taking Basque as the basis for the analysis.

In order to evaluate the adequacy of our annotation decisions, two annotators have annotated a set of 15 documents. The results are satisfactory, although there is still room for improvement. From our first analysis, one can say that our corpus contained many events the factuality of which was easy to identify—news text usually contain big amounts of facts, real past events. As a consequence, inter-annotator agreement was high. The reanalysis and further discussion of the disagreement will give us a better insight of the factuality annotation.

## Acknowledgements

This research is funded by the Basque Government PRE\_2016\_2\_294 grant and by the TIN2016-77820-C3-1-R project of the Ministry of Economy and Competitiveness (Spain).

## References

- [1] Begoña Altuna, María Jesús Aranzabe, and Arantza Díaz de Ilarraza. Euskarazko denbora-egiturak. Azterketa eta etiketatze-esperimentua. *Linguamática*, 6(2):13–24, Dezembro 2014.

- [2] Begoña Altuna, María Jesús Aranzabe, and Arantza Díaz de Ilarraza. Adapting TimeML to Basque: Event Annotation. In *Lecture Notes in Computer Science (LNCS)*. Springer, 2016.
- [3] Begoña Altuna, María Jesús Aranzabe, and Arantza Díaz de Ilarraza. Euskarazko denbora-informazioaren tratamendu automatikoa TimeMLren eta HeidelTimeren bidez. *Ekaia*, (30):153–165, 2016.
- [4] Begoña Altuna, María Jesús Aranzabe, and Arantza Díaz de Ilarraza. Euskarazko ezeztapenaren tratamendu automatikorako azterketa. In Iñaki Alegria, Ainhoa Latatu, Miren Josu Ormaetxebarria, and Patxi Salaberri, editors, *II. IkerGazte, Nazioarteko Ikerketa Euskaraz: Giza Zientziak eta Arteak*, pages 127–134, Bilbao, Euskal Herria, 2017. Udako Euskal Unibertsitatea (UEU).
- [5] Anne-Lyse Minard, Alessandro Marchetti, and Manuela Speranza. Event factuality in italian: Annotation of news stories from the ita-timebank. In *Proceedings of CLiC-it 2014, First Italian Conference on Computational Linguistic*, 2014.
- [6] James Pustejovsky, José M. Castaño, Robert Ingria, Roser Saurí, Robert J. Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R. Radev. TimeML: Robust Specification of Event and Temporal Expressions in Text. *New directions in question answering*, 3:28–34, 2003.
- [7] Haritz Salaberri Izko. *Rol semantikoen etiketatzeak testuetako espazio-denbora informazioaren prozesamenduan daukan eraginaz*. PhD thesis, Universidad del País Vasco/Euskal Herriko Unibertsitatea, Donostia, Euskal Herria, 2017.
- [8] Roser Saurí. *A Factuality Profiler for Eventualities in Text*. PhD thesis, Waltham, MA, USA, 2008. AAI3304029.
- [9] Roser Saurí, Olga Batiukova, and James Pustejovsky. Annotating Events in Spanish. TimeML Annotation Guidelines. Technical report, Technical Report Version TempEval-2010., Barcelona Media-Innovation Center, 2009.
- [10] Chantal van Son, Marieke van Erp, Antske Fokkens, and Piek Vossen. Hope and Fear: Interpreting Perspectives by Integrating Sentiment and Event Factuality. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC 2014)*, pages 26–31, Reykjavik, Iceland, May 2014.
- [11] Dina Wonsever, Aiala Rosá, Marisa Malcuori, Guillermo Moncecchi, and Alan Descoins. Event Annotation Schemes and Event Recognition in Spanish Texts. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 7182 of *Lecture Notes in Computer Science*, pages 206–218. Springer Berlin Heidelberg, 2012.

# Placenames analysis in historical texts: tools, risks and side effects

Adrien Barbaresi

Academy Corpora  
Austrian Academy of Sciences  
E-mail: [adrien.barbaresi@oeaw.ac.at](mailto:adrien.barbaresi@oeaw.ac.at)

## Abstract

This article presents an approach combining linguistic analysis, geographic information retrieval and visualization in order to go from toponym extraction in historical texts to projection on customizable maps. The toolkit is released under an open source license, it features bootstrapping options, geocoding and disambiguation algorithms, as well as cartographic processing. The software setting is designed to be adaptable to various historical contexts, it can be extended by further automatically processed or user-curated gazetteers, used directly on texts or plugged-in on a larger processing pipeline. I provide an example of the issues raised by generic extraction and show the benefits of integrated knowledge-based approach, data cleaning and filtering.

## 1 Introduction

In Western tradition, a current of reflexion whose origin can be dated back to the 1960s has provided the theoretical foundations of the spatial turn, whose epitome is the concept of space as emergent rather than existing a priori, and composed of relations rather than structures. As a consequence, both the definition and the importance of space have been re-evaluated throughout the humanities. More recently, researchers have suggested the crossing of research objects between disciplines and the enforcement of the “spatial turn” in practice through specific methods of analysis. Even so, corpus linguistics and geographical information systems have traditionally had very little to do with each other, although both approaches can benefit from each other [13].

Distant reading practitioners employ computational techniques to mine the texts for significant patterns and make statements about them [29]. From the point of view of computational linguistics, toponyms are a particular kind of out-of-vocabulary tokens. On lexical level, they are a potential error source for natural language processing tools. On phrasal level, they are supposed to be identified by part-of-speech taggers as named entities or eventually by more fine-grained

named-entity recognition tools as placenames. The processing chains usually stop at this point, they do not provide visualizations in the geographical sense, even if the toponyms can be linked to meta-information such as type and georeference and although progresses in fulltext geocoding are tightly linked to progresses in mapping systems, mostly thanks to a technology-driven evolution [17]. On the other hand, publicly available geocoding solutions do not usually come with interfaces to linguistic methods such as disambiguation and/or annotation layers. Finally, existing cartographic software solutions are not typically built for the visualization of digital text collections.

This article summarizes issues related to historical texts and describes an effort to conveniently go from texts to maps by integrating several key steps in a modular software package: data curation and preparation, processing of linguistic corpora, geocoding, and projection on maps. The use of a toolkit creates a common ground for hypothesis testing and visualization, while at the same time being compatible with other software in terms of formats and software environment. I provide an example of the issues raised by generic extraction and show the benefits of integrated knowledge-based approach, data cleaning and filtering.

## 2 Previous work

Among the tendencies in geographic information retrieval and geocoding [20], the extraction and normalization of named places, itineraries, or qualitative spatial relations, as well as the extraction of locative expressions are particularly relevant to study text collections. In the field of information retrieval, named entity recognition defines a set of text mining techniques designed to discover named entities, connections and the types of relations between them. The particular task of finding placenames in texts is commonly named placenames extraction or toponym resolution. It involves first the detection of words and phrases that may potentially be proper nouns and second their classification as geographic references [21]. A further step, geocoding, resides in disambiguating and adding geographical coordinates to a placename. Geocoding mostly relies on gazetteers, i.e. geospatial dictionaries of geographic names, mostly names, locations, and metadata such as typological information, variants or dates [15]. Toponym resolution as well as named-entity recognition can use machine learning methods [18], however these are generally not ideal when tackling data not present in the training set, so that knowledge-based methods using additional fine-grained registers, for example from Wikidata [28], have already been used with encouraging results.

Especially for historical corpora, researchers face a lack of general-purpose tooling. In order to produce cartographic visualizations, both the capacity to adapt to different contexts [3] and the necessity to complement existing resources with a precise historical gazetteer [9] have been highlighted. Such historical gazetteers exist, but their development is challenging [26] even for texts as late as 20th century Europe [22]. Existing toolboxes, such as AATOS [27], mostly feature candidate

extraction and ranking as well as entity linking. Heidelberg [24] does implement a comparable series of operations but it is currently tied to a series of engineering decisions which do not make its use on historical corpora straightforward. My approach is more light-weight, modular and adaptable, with a similar scope as CORE [19] but with an overall greater focus on usability, texts as input, integration of registers, and map export as images.

## 3 Tooling

### 3.1 Requirements

In order to process linguistically annotated text, it is useful to be able to start from either raw text or common formats for part-of speech tags and named entities recognition. The toolkit is pluggable to existing NLP solutions or usable directly on text, although morpho-syntactical analyses are appropriate in order to narrow down the search to relevant tokens, such as phrase heads found by surface parsing [4]. Gazetteers can be curated in a semi-supervised way to account for historical differences. Knowledge-based techniques are a way to tailor the geoparsing to historical contexts. Nevertheless, bootstrapping geographical data can save a significant amount of time. The generic gazetteer GeoNames [12] and structured data from Wikipedia and Wikidata are widely known to the research community. Wikipedia’s API can be used to navigate in categories and to retrieve coordinates, including for historical places or areas. Current information is usually compiled from the GeoNames database, which also often includes historical variants. Additional lexical cues like stoplists or linguistic information such as suffixes or derivation ought to be configurable, as tools trained on modern texts do not necessarily tag historical morpho-syntactic patterns as needed. To provide support for manual annotation, an additional layer can be convenient as a geocoding bypass for targeted user lists which can operate on token or lemma level (using either linguistic processing or regular expressions and wildcards).

### 3.2 Concrete approach

The toolbox used for the experiments below is currently being developed [5] with historical texts in mind. It has already been used so far to map different text collections ranging from the 17th to the 20th century [6].

There is no commonly adopted standard for gazetteers, they have to be combined. Consequently, my approach allows for additional input, special sorting and prioritized merging, for example to put historical variants in the foreground. Second, it includes helpers to bootstrap geographical data, as knowledge-based methods using fine-grained data improve the results [28]. So far, import filters for GeoNames and structured data from Wikipedia and Wikidata are implemented, with a particular emphasis on data cleaning. Third, an additional layer allows to bypass geocoding for targeted, easily extensible user lists which can operate on

token or lemma level (using either linguistic processing or regular expressions and wildcards).

To spare resources, the extraction is performed by a sliding window capturing single tokens as well as multi-word expressions. Two different types of disambiguation methods [10] are included so far in the toolbox: map-based and knowledge-based. It has been shown that an acceptable precision can be reached by including meta-information [23], which consists here in distance (based on a calculation relative to a contextual setting), type and importance of the entries (as known from information extracted from GeoNames or Wikipedia), as well as immediate context (e.g. the expected range and the last country seen). The process can be controlled by parameters set by the user, such as distance calculations, filter level or size of the search area.

Additionally, the toolbox integrates its own visualization component<sup>1</sup> which makes use of the Python module *matplotlib* and its extension *cartopy*. It is profitable to allow for adaptability of projection and design and to leave it open to the user to refine the map, in a particular emphasis on the concept of visualization.

The toolkit is bundled as Python package, currently one of the most frequently used programming languages in academia,<sup>2</sup> it is available under an open-source license.<sup>3</sup> The release includes the code, especially the functions dedicated to geographic information retrieval, which form the bases of previous studies. It is meant to ensure replicability and extendability in an open science perspective and can hopefully respond to a growing demand in this field.

### 3.3 Contextual settings

The streamlined process from text to map involves a series of decisions as well as a critical reading of texts and maps. As user-definable settings make results vary, experiments can lead to diverging realizations. In fact, the extraction and visualization settings have a significant influence. In order to make them easily configurable, they are all accessible in a settings file. First and foremost, the filtering level affects both the construction of gazetteers prior to geoparsing and the toponym recognition phase in itself. Its purpose is to allow for a tighter or looser control on the data, with either restricted options or opportunistic search. Second, the minimum length of tokens to consider as valid toponyms, which is a function of the frequency, can be ignored or determined in advanced. Third, the disambiguation phase can be controlled by map-based parameters, notably the reference point for distance calculations and the countries in the vicinity, which help identifying the most probable candidates. Last, the cartographic processing in itself can be configured (window size and labels). Altogether, the settings allow for an opportune handling of historical texts. The process can adapt to different texts and contexts and it can evolve to

---

<sup>1</sup>The software used in previous experiments (TileMill) is no more under active development and needed to be installed and used separately.

<sup>2</sup><https://spectrum.ieee.org/computing/software/the-2017-top-programming-languages>

<sup>3</sup><https://github.com/adbar/geokelone>



reflect historical empires or regions for example, both during geoparsing (account for and disambiguate among historical names) and mapping (display historical or canonical names).

## 4 Risks and side effects

### 4.1 Examples

In order to better assess the impact of filtering and complementary registers, I present and discuss two different comparisons on close reading and on distant reading levels. Specially curated gazetteers are used, while current geographical information is used as a fallback, entries corresponding to European countries are retrieved and preprocessed.



Figure 1: No filter, standard GeoNames setting



Figure 2: Cleaned data with meta-information

First, I test the coverage and the options at close level on a simple historical example. The sentence to be analyzed is from the late 19th century and features a series of proper nouns so that the experimental setting has an effect both on both form and content.<sup>4</sup> The standard fallback gazetteer, GeoNames, is known to be prone to coverage and data quality issues [2]. Figure 1 displays an unfiltered view using raw text and GeoNames as only gazetteer. Only one point out of five is placed correctly while two other are wrongly considered to be placenames, and one place name is missing. The most prominent error concerns the token *Berlin*, which in GeoNames corresponds to a settlement in Northern Germany without inhabitants. The capital

<sup>4</sup>Taken from *Der Stechlin* by Theodor Fontane: “*Ich sage Ihnen, Hauptmann, das waren Preußens beste Tage, als da bei Potsdam herum die ‘russische Kirche’ und das ‘russische Haus’ gebaut wurden, und als es immer hin und her ging zwischen Berlin und Petersburg.*”

city of Germany is indeed never present as a single token in the dataset but systematically in the form of city quarters such as *Berlin-Alexanderplatz*. Figure 2 shows the impact of filtering (both knowledge-based and POS-based filtering lead here to the removal of false positives) and external resources (proper geocoding with a historical gazetteer), which lead to correct results when used in combination. This simple example illustrates how quality control and text analysis can benefit from the projection of the results on a map.



Figure 3: Minimum filtering



Figure 4: Maximum filtering

Next, the impact of filtering methods on distant reading experiments is shown. Karl Kraus (1874-1936) founded his own journal, *Die Fackel* (“The Torch”), in 1899 and published it until his death. This complex and unique work resists summary description in its whole and in detail, it has been used as a basis for distant reading experiments using placenames and collocations as entry points to provide an additional, synthetic overview of the collection [7]. The present experiments use the same text base from the digital edition of the work [8], the texts have been manually corrected as well as manually annotated with respect to the names of persons and institutions, so that most proper nouns which are not placenames can be excluded from the study. Figure 3 displays the results with a minimum filtering on a map showing most of continental Europe. Clusters can be found everywhere, not all of them being either intuitively explainable or justified with respect to the texts. In fact, the map tells more about the gazetteers used for geoparsing as about the work in itself. Current boundaries are retraceable, and numerous false positives come from plurilingual countries such as Switzerland or Belgium which are then overrepresented on the map. Figure 4 consists of a similar map featuring the results of maximum filtering level both during the construction of resources and

during the extraction process. The map is more easily readable and depicts an accurate centering on Vienna and its surroundings. The overall Westward tropism of the mapped locations seems to coincide with the texts. This map is thus well-suited for further analyses.

## 4.2 Discussion

GeoNames, arguably the most commonly used gazetteer, has to be put under scrutiny, as the entries and their classification are subject to numerous problems, mostly unevenly distributed data and sparse metadata, which impact both detection and disambiguation of placenames [1]. Nevertheless, this resource is still valuable mostly because of its coverage of language variants and thus potentially historical variants.

The status of placenames that are to be found and projected on the map also ought to be discussed. There are consubstantial ambiguities on linguistic level that complicate the search [25]: the referent ambiguity (one name used for more than one location) and the referent class ambiguity (placenames used as organization or person names) are commonly addressed by disambiguation processes, whereas reference ambiguity (more than one name for the same location) has to be dealt with during the compilation of geographical databases. In general, successful detection and disambiguation relies on a smart interplay of resources and tools at different levels. Last, the case of either imprecise, vague or vernacular names [16] is a prominently linguistic issue which can at least be addressed by manual curation and should in any case be attended to.

Concerning the maps themselves, the consensus in the research community has evolved towards a relativity in construction and uses of maps, as there is neither a ground truth nor a cartographic truth. Although the maps seem immediately interpretable, they are not an objective outcome but a construct resulting from a series of interventions. "Selection, omission, simplification, classification, the creation of hierarchies, and 'symbolization' – are all inherently rhetorical" [14]. As such, cartography is not the realization of static maps, but rather the description of emergent structures, and there is no single or best map.

Finally, the object of scientific inquiry does not simply reside in linking text to space, it is tightly linked to the interpretation of texts and maps. Even if the methodology conveys a feeling of scientific objectivity, the validity of mental and computerized operations described here should always be examined with respect to their relevance. Geospatial analysis and spatial representation may indeed be deficient or inadequate. The anthropological significance of toponyms has been emphasized by testimonies gathered on the field [11], but the symbolic role and the expressive power of placenames do not necessarily coincide with Western instrumental science and cartography, in that particular case the world geodetic system and the chosen map projection.

## 5 Conclusion

This article introduced theoretical and practical instruments combining philological knowledge, geographic information retrieval and visualization, in order to streamline the steps needed to go from texts to maps. Examples of the issues raised by generic extraction have been discussed, they show the advantages of a methodology centered on historical texts and subsequent data cleaning and filtering. Being able to go through all the operation in one shot is ideal to assess the risks, to spot problems in methodology or datasets, and hopefully to mitigate the side effects.

The maps play an ambiguous role in distant reading, since they have to be flexible enough to adapt to new contexts and analyses, while remaining exact and in this regard trustworthy. The information they contain and reveal cannot always be verified on a point-per-point basis, yet it can be the starting ground of scientific reasoning. In fact, text visualizations are the substrate of interpretable representations which do not follow data but rather confront them by putting them in perspective. The difference between data wrangling and research in digital humanities resides precisely in the number and diversity of conceptual and technical filters which are repeatedly applied, consciously or sometimes unknowingly. The chosen approach and its inevitable imperfections have to be brought to light, documented and criticized.

In a linguistic perspective, the tools allow for the systematization of research as well as for a critical approach to the extraction and the very concept of placenames. As quantitative and qualitative analysis can go hand in hand, digital literary studies are not mere numeric accounts. They are first and foremost a discovery process. The use of filtering, the customized gazeteers and maps, in short the human interventions as well as the technical competence to do so recreate the hermeneutic circle of the philological tradition.

## References

- [1] Elise Acheson, Stefano De Sabbata, and Ross S. Purves. A quantitative analysis of global gazeteers: Patterns of coverage for common feature types. *Computers, Environment and Urban Systems*, 64:309–320, 2017.
- [2] Dirk Ahlers. Assessment of the Accuracy of Geonames Gazetteer Data. In *Proceedings of the 7th Workshop on GIR*, pages 74–81. ACM, 2013.
- [3] Beatrice Alex, Kate Byrne, Claire Grover, and Richard Tobin. Adapting the Edinburgh geoparser for historical georeferencing. *International Journal of Humanities and Arts Computing*, 9(1):15–35, 2015.
- [4] Adrien Barbaresi. A one-pass valency-oriented chunker for German. In *Proceedings of the 6th Language & Technology Conference*, pages 157–161, 2013.

- [5] Adrien Barbaresi. Towards a Toolbox to Map Historical Text Collections. In *Proceedings of the 11th Workshop on Geographic Information Retrieval, GIR'17*. ACM, 2017.
- [6] Adrien Barbaresi. A constellation and a rhizome: two studies on toponyms in literary texts. In Bubenhofer Noah and Kupietz Marc, editors, *Visual Linguistics*. Heidelberg University Publishing, Heidelberg, 2018. To appear.
- [7] Adrien Barbaresi. Toponyms as Entry Points into a Digital Edition: Mapping Die Fackel. *Open Information Science*, 2018. To appear.
- [8] Hanno Biber, Evelyn Breiteneder, Heinrich Kabas, and Karlheinz Mörth. Austrian Academy Corpus AAC–FACKEL. Die Fackel. Herausgeber: Karl Kraus, Wien 1899-1936, Edition No 1, Online Version: <http://www.aac.ac.at/fackel>.
- [9] Lars Borin, Dana Dannélls, and Leif-Jöran Olsson. Geographic visualization of place names in Swedish literary texts. *Literary and Linguistic Computing*, 29(3):400–404, 2014.
- [10] Davide Buscaldi. Approaches to disambiguating toponyms. *Sigspatial Special*, 3(2):16–19, 2011.
- [11] S. Feld. Waterfalls of song: An acoustemology of place resounding in Bosavi, Papua New Guinea. In S. Feld and K.H. Basso, editors, *Senses of place*, pages 91–135. School of American Research Press, 1996.
- [12] Unxos GmbH. Geonames, 2017. <http://www.geonames.org>.
- [13] Ian N. Gregory and Andrew Hardie. Visual GISTing: Bringing together corpus linguistics and Geographical Information Systems. *Literary and Linguistic Computing*, 26(3):297–314, 2011.
- [14] John Brian Harley. Deconstructing the map. *Cartographica: The international journal for geographic information and geovisualization*, 26(2):1–20, 1989.
- [15] Linda Hill. Core elements of digital gazetteers: placenames, categories, and footprints. *Research and advanced technology for digital libraries*, pages 280–290, 2000.
- [16] Christopher B. Jones, Ross S. Purves, Paul D. Clough, and Hideo Joho. Modelling vague places with knowledge from the Web. *International Journal of Geographical Information Science*, 22(10):1045–1065, 2008.
- [17] Marko Juvan. From spatial turn to GIS-mapping of literary cultures. *European Review*, 23(1):81–96, 2015.

- [18] Jochen L. Leidner and Michael D. Lieberman. Detecting Geographical References in the Form of Place Names and Associated Spatial Natural Language. *SIGSPATIAL Special*, 3(2):5–11, 2011.
- [19] Eetu Mäkelä, Thea Lindquist, and Eero Hyvönen. CORE – a Contextual Reader Based on Linked Data. *Digital Humanities 2016*, pages 267–269, 2016.
- [20] Fernando Melo and Bruno Martins. Automated Geocoding of Textual Documents: A Survey of Current Approaches. *Transactions in GIS*, 21(1):3–38, 2017.
- [21] Damien Nouvel, Maud Ehrmann, and Sophie Rosset. *Les entités nommées pour le traitement automatique des langues*. ISTE editions, 2015.
- [22] Paolo Plini, Sabina Di Franco, and Rosamaria Salvatori. One name one place? Dealing with toponyms in WWI. *GeoJournal*, pages 1–13, 2016.
- [23] Bruno Pouliquen et al. Geocoding multilingual texts: Recognition, disambiguation and visualisation. In *Proceedings of LREC*, pages 53–58. ELRA, 2006.
- [24] Ludwig Richter, Johanna Geiß, Andreas Spitz, and Michael Gertz. HeidelbergPlace: An Extensible Framework for Geoparsing. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 85–90, 2017.
- [25] David A. Smith and Gideon S. Mann. Bootstrapping toponym classifiers. In *Proceedings of the HLT-NAACL 2003 workshop on Analysis of Geographic References*, pages 45–49. Association for Computational Linguistics, 2003.
- [26] Humphrey Southall, Ruth Mostern, and Merrick Lex Berman. On historical gazetteers. *International Journal of Humanities and Arts Computing*, 5(2):127–145, 2011.
- [27] Minna Tamper, Petri Leskinen, Esko Ikkala, Arttu Oksanen, Eetu Mäkelä, Erkki Heino, Jouni Tuominen, Mikko Koho, and Eero Hyvönen. AATOS – a Configurable Tool for Automatic Annotation. In *International Conference on Language, Data and Knowledge*, pages 276–289. Springer, 2017.
- [28] Denny Vrandečić and Markus Krötzsch. Wikidata: a Free Collaborative Knowledge Base. *Communications of the ACM*, 57(10):78–85, 2014.
- [29] Clifford E. Wulfman. The Plot of the Plot: Graphs and Visualizations. *The Journal of Modern Periodical Studies*, 5(1):94–109, 2014.

# Resources and Methods for the Automatic Recognition of Place Names in Alsatian

Delphine Bernhard<sup>1</sup>, Pierre Magistry<sup>2</sup>,  
Anne-Laure Ligozat<sup>3</sup> and Sophie Rosset<sup>2</sup>

<sup>1</sup>LiLPa - EA 1339, Université de Strasbourg, France

<sup>2</sup>LIMSI, CNRS, Université Paris-Saclay, F-91405 Orsay, France

<sup>3</sup>LIMSI, CNRS, ENSIIE, Université Paris-Saclay, F-91405 Orsay France

E-mail: dbernhard@unistra.fr,

{annlor,magistry,sophie.rosset}@limsi.fr

## Abstract

This article describes annotated resources (corpus, lexicons) for the automatic recognition of place names in the Alsatian dialects. The two main issues are related to their non-standardized orthography, leading to spelling variants, and the scarcity of available resources. We also present automatic methods using recurrent neural networks for the identification of place names that take these aspects into account.

## 1 Introduction

The detection of real-world entities is important for many text understanding applications. The main purpose of named entity recognition (NER) is to identify such pieces of information as names of persons, places and organisations. Detecting named entities in text relies on resources, mainly lexicons, whatever the approach used to develop the system (rule-based or statistical).

In this article, we present resources and methods for the automatic identification of place names in the High German Alsatian dialects, spoken in North-Eastern France. The Alsatian dialects are an heritage of the linguistic changes brought in the region by the Alemanni and the Franks as early as the 6th century [12]. The geographic region now called Alsace started to be progressively integrated into France during the 17th century. However, this did not have a real impact on everyday language use, especially for the middle and lower classes. Major changes to this situation occurred only after World War II, and in particular in the last third of the 20th century, when French also became the language of communication in all everyday activities. This period is also characterized by the gradual decline of the Alsatian dialects.

The Alsatian dialects have always been used mainly orally, with either French or German being used as the languages of choice for writing, depending on the time period. The earliest writings in Alsatian can be traced back to the second half of the 17th century [20], but the real beginnings of Alsatian literature are attributed to the comedy in verse published by Jean-Georges-Daniel Arnold in 1816 (*Der Pfingstmontag*). Since then, there has been an ongoing –albeit not very numerous– literary and cultural production with a focus on two main text genres : theater plays (mostly comedies) and poetry [21, 22, 23, 24]. Other genres are also represented: poetic prose, songs, nursery rhymes, tales, translations and adaptations of works in other languages. In addition to this literary production, there have also been efforts to provide linguistic descriptions in the form of dictionaries, glossaries and grammars. However, it should be noted that texts in prose are rarer, with the exception of a few authors such as Marie Hart: either French or German are used in this case.

While French is clearly dominant nowadays in the public space, place names are one of the cases where Alsatian is still present in everyday life, even to non dialect speakers. An increasing number of villages and cities choose to have French-Alsatian bilingual town entrance and street signs.<sup>1</sup> Each location in Alsace has two names: an “official” name and a “popular” name [16]. Official names are written and stem from the German language for 95% of them, though often with spelling deviations with respect to the norm [16]. Popular names are oral and often correspond to the dialectal pronunciation of the official name. There may however exist form differences, e.g. the popular name of the village *Schwindratzheim* is *Schwingelse* [16]. Moreover, the pronunciation may differ depending on the dialectal variant in use, e.g. *Wissembourg* is locally pronounced *Waisseburch*, but *Wisseburi* or *Wissaburg* elsewhere in the region [16].

Repositories of place names in Alsatian are very scarce and have to be collected, improved and categorized according to a well-defined typology. Various typologies of named entities exist and they differ by their semantic coverage or categorical representations [7]. The MUC-6 project [9] was the first to propose a definition of NEs as proper nouns that refer to specific semantic classes: Person, Location, Organisation, etc. New typologies were proposed based on this first definition, aiming at more fine grained classes [5, 18]. More recently, a new typology was defined within the QUAERO project [10] with the objective of being more general than Sekine’s typology [18] while having wide semantic coverage. In this work, we used the location type definition as defined in QUAERO’s typology. As shown in Figure 1 (left-hand side), the different location categories are: administrative locations with geopolitical definitions (cities, countries, states...), physical locations (geonyms, hydronyms, astronyms), toponyms (streets, squares...), facilities (train stations, universities...) and addresses (physical or electronic).

The main contributions of this article are as follows (1) we present lexicons of Alsatian place names which have been manually categorized according to the

---

<sup>1</sup>See for instance [https://commons.wikimedia.org/wiki/File:Mulhouse\\_entr%C3%A9e\\_agglom%C3%A9ration.JPG](https://commons.wikimedia.org/wiki/File:Mulhouse_entr%C3%A9e_agglom%C3%A9ration.JPG)



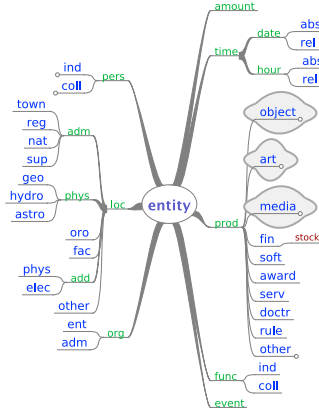


Figure 1: Quaero Typology [17].

QUAERO named entity types (Section 2.1) ; (2) we describe the first corpus manually annotated with place names for the Alsatian dialects (Section 2.2) ; (3) we propose and experiment with methods to automatically identify locations in Alsatian corpora (Section 3).

## 2 Description of the Resources

### 2.1 Lexicons

The lexicons we collected consist of bilingual lists of place names in French and Alsatian, manually categorized with respect to the QUAERO location types. The sources used for the collection include printed and online material:

- Alsadico: a printed French-Alsatian dictionary [14];
- Elsàsser: a website which includes, among others, a list of Alsatian place names with their phonetic transcription and translation into French [13];
- WikiAls: the Alemannic Wikipedia ([als.wikipedia.org](http://als.wikipedia.org)). Wikipedia pages corresponding to locations were collected thanks to specific categories (e.g. [[Kategorie:Ort (Unterelsass)]] and the dialect tag of the page;
- WikiFr: French Wikipedia ([fr.wikipedia.org](http://fr.wikipedia.org)). Relevant pages were identified thanks to their category. Pages are written in French but often contain the Alsatian place names in the text, for example “La ville est appelée *Mil-hüsa* en alsacien” (The city is called Milhüsa in Alsatian). These instances were identified thanks to regular expressions detecting the presence of the expression “(en) [[alsacien]]” in the context.

Printed material, which was available only on paper, has been re-typed to obtain digital resources, while online material has been collected automatically and manually corrected afterwards. It should be noted that both Alsadico and Elsässer are copyrighted and hence cannot be freely redistributed.<sup>2</sup> We nevertheless used these resources in order to perform some linguistic analyses and assess the quality and coverage of the Wikipedia-derived datasets. The categorization into the QUAERO entity types has been performed by a first annotator and then reviewed and corrected by a second annotator. Table 1 shows an extract from the lexicon. The translation into French makes it possible to retrieve spelling variants.

Alsatian	French	Type	Sources
Algelse	Algolsheim	<loc.adm.town>	Elsässer
Àlgelse	Algolsheim	<loc.adm.town>	AlsaDico
Algolse	Algolsheim	<loc.adm.town>	WikiAls
Elsass	Alsace	<loc.adm.reg>	WikiFr
Elsäss	Alsace	<loc.adm.reg>	WikiAls

Table 1: Extract from the lexicon.

Table 2 details the contents of the lexicons. The overlap between the lexicons is rather low : on average, an Alsatian spelling variant is found in 1.18 lexicons : 2,484 spellings are found in one lexicon only, 376 in two, 66 in three and only 3 are present in all four lexicons (*Kurzehüse* ; *Molse* ; *Sundhüse*).

Type	AlsaDico	Elsässer	WikiAls	WikiFr	Corpus
loc.fac – Facility	0	9	2	1	2
loc.phys.astro – Astronym	0	0	0	0	1
loc.phys.geo – Geonym	0	55	16	6	60
loc.phys.hydro – Hydronym	0	22	22	1	9
loc.adm.nat – Country	0	0	10	0	50
loc.adm.reg – Region	3	1	89	1	119
loc.adm.sup – Supranational	0	0	0	0	3
loc.adm.town – City	1,184	1,038	891	94	154
loc.oro – Odonym	0	0	0	1	3
Total	1,187	1,125	1,030	104	401

Table 2: Types of locations in the resources. For the corpus, the figures correspond to the number of occurrences.

## 2.2 Corpus

The corpus is composed of two main sources of Alsatian documents: Wikipedia articles from the Alemannic Wikipedia and chronicles from an information magazine

<sup>2</sup>The non-copyrighted lexicons are scheduled to be released on the project’s website <http://restaure.unistra.fr/> under a CC BY-SA license.

published by the Haut-Rhin department (southern Alsace) General Council. The Wikipedia articles are written in different Alemannic geolinguistic variants found in the Alsace region, while the chronicles are written in Low Alemannic from the southern part of the region. In addition, two more specific genres were used for the annotator training phase: one excerpt from a theater play and some recipes. The two main sources meet the requirements set within our project whose overall goal is to provide resources and tools for language learning and documentation: (i) freely redistributable,<sup>3</sup> (ii) contemporary forms of the Alsatian dialects, (iii) texts in prose with a wide potential audience.<sup>4</sup> Texts in prose were favoured over poems or theatre plays because they are easier material both for the development of NLP tools and for language learners. Texts from other genres and time periods will be considered in the future.

The annotated corpus contains 21 documents and 12,570 tokens (including punctuation). Location annotation was performed following a BIO (Begin-Inside-Out) notation: the first token of a location is annotated with a B-LOC tag, the following ones with I-LOC, and non location tokens with an O tag.

Overall, 401 occurrences of place names have been annotated in the corpus, corresponding to 207 different place names. As can be seen in Table 2, location types found in the corpus are much more varied than those found in the lexicons. While names of cities are largely predominant in the lexicons, the corpus also contains many geonyms, countries and regions. Only 40 tokens were annotated with a I-LOC tag: most location entities are expressed by a single token.

### 2.3 Spelling Variation in the Resources

Spelling variation is an important issue in the Alsatian dialects, since there is no widely acknowledged and standardised orthography. In the lexicons, a location name has on average 1.92 Alsatian graphical variants. The maximum number of spelling variants is 14, for the town Mulhouse (see Figure 2). In the corpus, a location name has on average 1.38 Alsatian graphical variants. Interestingly, the maximum number of spelling variants is also found for the town Mulhouse, with six different spellings (see Figure 2).



Figure 2: Spelling variants for the town “Mulhouse”

<sup>3</sup>We secured the authorization of the publisher for the chronicles.  
<sup>4</sup>The Alemannic Wikipedia is freely available on the Web, while the Haut-Rhin information magazine is distributed in all homes in the department.

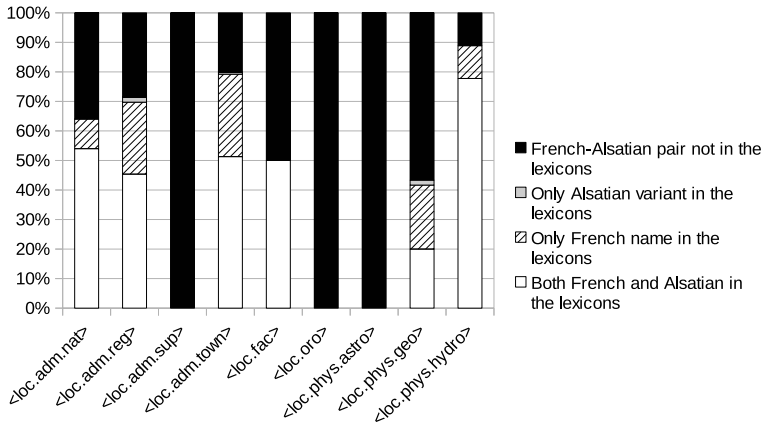


Figure 3: Overlap between the corpus and the lexicons.

We also measured the overlap between the corpus and the lexicons for each entity type (see Figure 3). Overall, from the 401 occurrences in the corpus, 180 French-Alsatian pairs are found in the lexicons, 91 French location names are found in the lexicon but not with the same Alsatian spelling variant, 4 Alsatian spelling variants are found in the lexicons but not with the same French location name and 126 French-Alsatian pairs are totally absent from the lexicons (neither the French location name, nor the Alsatian spelling variant is found in the lexicons). As could be expected from the composition of the lexicons, the coverage is highly dependent on the entity type: supranational entities, odonyms and astronyms are completely absent from the lexicons.

To sum up, there are two main issues: (i) the low coverage of the lexicons with respect to the place names found in the corpus and (ii) the large amount of spelling variants found in the corpus.

### 3 Location Recognition and Linking

Since the location lexicons lack coverage, we chose to use word embeddings to inform our location recognition and linking methods. Word embeddings are semantic representations of words in a low dimension vector space, learnt in an unsupervised way from large corpora.

We first tested a baseline system for location named entity recognition. The system performs named entity recognition for locations and is based on a bi-directional LSTM (Long Short Term Memory, which is a type of recurrent neural network) and word representations built with fastText [1]. fastText is a library for learning word representations using character  $n$ -grams of words, which improves

the representation of rare words. The architecture of this system is the biLSTM-CRF as proposed in [15]. This model predicts a sequence of tags for any sequence of observations (in our case, a sentence) based on multiple sources of information. To select a tag for a word, it combines contextual information from surrounding words and characters (modelled as vectors) and tagging decisions to be made about the other tokens of the sentence. The model uses the BIO tagged corpus. Results are given in Table 3.

System	Recall	Precision	F-measure
LSTM (1doc vs. all)	.53	.60	.56
LSTM (80/20)	.61	.79	.69

Table 3: Named entity recognition for locations

We used cross-validation on documents to evaluate our system under two conditions:

- first, the corpus of 21 annotated documents was divided into 21 train/test sets, with one document being the test corpus and all the others the training corpus (1 doc vs. all condition).
- we also tested cross-validation on sentences (80/20 condition): 80% of the corpus was used to train the model, and 20% to test it.

To build the word embeddings, we rely only on the WikiAls data. The missing vectors for the words from our annotated documents which are absent from the WikiAls corpus are generated in a second step using the method included in fast-Text to avoid having Out-of-Vocabulary tokens. This vector space is then used for the initialization of the embedding layer of the biLSTM-CRF. This system achieves a .56 F-measure with the 1 doc vs. all condition and .69 F-measure with the 80/20 condition, which is a little easier since parts of the train and test corpora come from the same documents.

The second system performs named entity linking: when given a new location name recognized by the first system, it links it to an existing location cluster by selecting the most similar locations in the embeddings. We tested the following method: for each location in the annotated corpus, we selected the 10 most similar words in the embeddings, and checked if these similar words were associated to the same lexicon entry. For example, for the location “Stroßburg” in the text, one of the most similar word is “Stroßburi”, and both are Alsatian variants for “Strasbourg” in the lexicon. The goal is to assess whether, if the text location was absent from the lexicon, it could still be linked to the correct lexicon entry. For the 207 distinct location entities from the annotated corpus, 61 can be linked to the correct entry. Graphical variants are found among the most similar words, such as “Milhüüse” for “Milhüüsa”, “Schwitz” for “Schwiz”, “Tännchel” for “Taennchel” or “Sawere” for “Zàwere”.

## 4 Related Work

Most NER systems rely on information on the words to be annotated: forms of the word, presence in a specific lexicon, part-of-speech tags etc. All this information is obviously affected by surface form variations [19, 2]. Spelling variation is indeed a well-documented issue for the identification of place names in historical texts [3], given that the standardization of orthography is often quite recent.

Concerning specifically location detection in historical data, Borin et al. [2] proposed a knowledge- and rule-based approach which aims specifically at handling the variation in 19th century Swedish literature. Their best system reaches an F-measure of 86.4%. For the Arabic language, which also presents high variation, a knowledge and rule-based approach is described in [19]. The presented system reached an F-measure of 85.9%. All these works consider only one class for the location type. The QUAERO typology, on which our work is based, was used on French old press data [8] and on Swiss old press data in French [6] which also contains a lot of variation. In the latter, the authors compared various systems and for the location type the results ranged between 48% and 69% of F-measure depending on the system tested, which is similar to what we obtained in this work. Most current models for NER consider it as a supervised sequential classification problem where each sentence is a sequence [4, 11, 15]. In order to categorize words, the model can rely on orthographic information, captured by character-based representations, and distributional information, captured by word embeddings. Recently a method to represent such information which includes character level information was proposed [1]. This approach is often considered as being robust against variation. Our hypothesis was that this model may be useful for our purpose.

## 5 Conclusion and Perspectives

We have described the first corpus manually annotated with place names for the Alsatian dialects as well as lexicons collected from different sources. The automatic methods proposed to identify place names face two main issues: low coverage of the lexicons and spelling variants.

The corpora and lexicons correspond to contemporary Alsatian and the texts in the corpus belong to two particular text genres (encyclopaedia and chronicle). The next step will be to assess whether the methods and resources can be applied to older texts and different text genres, in particular theatre plays and tales.

## Acknowledgements

We would like to thank Clément Dorffer, Elisa Feuerstein, Mario Luis Figueroa Miranda and Gwendoline Hollner for their participation in the collection and annotation of the datasets. This work was supported by the French “Agence Nationale de la Recherche” (ANR) (project no.: ANR-14-CE24-0003).

## References

- [1] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016.
- [2] Lars Borin, Dimitrios Kokkinakis, and Leif-Jöran Olsson. Naming the Past: Named Entity and Animacy Recognition in 19th Century Swedish Literature. In *Proceedings of LaTeCH 2007*, pages 1–8, Prague, Czech Republic, June 2007.
- [3] James O. Butler, Christopher E. Donaldson, Joanna E. Taylor, and Ian N. Gregory. Alts, Abbreviations, and AKAs: historical onomastic variation and automated named entity recognition. *Journal of Map & Geography Libraries*, 13(1):58–81, 2017.
- [4] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537, 2011.
- [5] George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. The Automatic Content Extraction (ACE) Program, Tasks, Data, and Evaluation. In *Proceedings of LREC-2004*, Lisbon, Portugal, May 2004.
- [6] Maud Ehrmann, Giovanni Colavizza, Yannick Rochat, and Frédéric Kaplan. Diachronic Evaluation of NER Systems on Old Newspapers. In *Proc. of KONVENS 2016*, pages 97–107, Bochum, Germany, 2016.
- [7] Maud Ehrmann, Damien Nouvel, and Sophie Rosset. Named Entity Resources - Overview and Outlook. In *Proceedings of LREC 2016*, may 2016.
- [8] Olivier Galibert, Sophie Rosset, Cyril Grouin, Pierre Zweigenbaum, and Ludovic Quintard. Extended Named Entities Annotation on OCRred Documents: From Corpus Constitution to Evaluation Campaign. In *Proceedings of LREC’12*, Istanbul, Turkey, may 2012.
- [9] Ralph Grishman and Beth Sundheim. Design of the muc-6 evaluation. In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*, 1995.
- [10] Cyril Grouin, Sophie Rosset, Pierre Zweigenbaum, Karèn Fort, Olivier Galibert, and Ludovic Quintard. Proposal for an Extension of Traditional Named Entities: From Guidelines to Evaluation, an Overview. In *Proceedings of LAW-V, ACL*, pages 92–100, Portland, OR, June 2011.
- [11] James Hammerton. Named entity recognition with long short-term memory. In *Proceedings of HLT-NAACL 2003-Volume 4*, pages 172–175, 2003.

- [12] Dominique Huck. *Une histoire des langues de l'Alsace*. la Nuée bleue, Strasbourg, 2015.
- [13] Marc Hug. Toponymes d'Alsace. Online, <http://elsasser.free.fr/NomCommu/ecrantot.html>, 2007.
- [14] Edmond Jung. *L'alsadico : 22 000 mots et expressions français-alsacien*. La Nuée bleue, Strasbourg, 2006.
- [15] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural Architectures for Named Entity Recognition. In *Proceedings of NAACL-HLT*, pages 260–270, June 2016.
- [16] Michel Paul Urban. *La grande encyclopédie des lieux d'Alsace*. La Nuée Bleue, Strasbourg, 2nd edition, 2010.
- [17] Sophie Rosset, Cyril Grouin, and Pierre Zweigenbaum. Entités nommées structurées : guide d'annotation Quaero. Technical report, 2011. Notes et Documents LIMSI N° : 2011-04.
- [18] Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata. Extended Named Entity Hierarchy. In *Proceedings of LREC 2002, May 29-31, 2002, Las Palmas, Canary Islands, Spain*, 2002.
- [19] Khaled Shaalan and Hafsa Raza. NERA: Named entity recognition for Arabic. *Journal of the Association for Information Science and Technology*, 60(8):1652–1663, 2009.
- [20] Auguste Wackenheim. *Tome 1. Du XVIIe au XIXe siècle – Les anonymes, les précurseurs, les fondateurs*. La littérature dialectale alsacienne: une anthologie illustrée. Prat-éditions, Paris, 1993.
- [21] Auguste Wackenheim. *Tome 2. L'âge d'or du XIXe siècle: la fin de l'Empire, la Restauration, le Second Empire*. La littérature dialectale alsacienne: une anthologie illustrée. Prat-éditions, Paris, 1994.
- [22] Auguste Wackenheim. *Tome 3. La période allemande, 1870-1918*. La littérature dialectale alsacienne: une anthologie illustrée. Prat-éditions, Paris, 1997.
- [23] Auguste Wackenheim. *Tome 4. D'une guerre mondiale à l'autre: 1918 - 1945*. La littérature dialectale alsacienne: une anthologie illustrée. Prat-éditions, Paris, 1999.
- [24] Auguste Wackenheim. *Tome 5. De 1945 à la fin du XXe siècle*. La littérature dialectale alsacienne: une anthologie illustrée. Prat-éditions, Paris, 2003.



# TEI and Bavarian dialect resources in Austria: updates from the DBÖ and WBÖ

Jack Bowers and Philipp Stöckle

Austrian Centre for Digital Humanities

Austrian Academy of Sciences

E-mail: [jack.bowers@oeaw.ac.at](mailto:jack.bowers@oeaw.ac.at)

## Abstract

In our paper, we present a large historical database of Bavarian dialects (from the *Dictionary of Bavarian Dialects in Austria*) and its conversion from hand-written paper slips via TUSTEP into TEI-XML while elaborating on the topics discussed by Bowers [2] with regards to enhancement of its contents. While the original purpose of the digitalization was to facilitate the writing of dictionary articles, our current TEI database will be used as a corpus from which the materials are being gathered to both write print dictionary articles as well as serving as a basis for a web-based lexicographic information system. Herein we trace the different steps that have already been taken to create our current digital database from a legacy data collection, discuss the challenges we are still facing, and describe the approaches we are taking and considering to address such challenges.

## 1 Introduction: A short history of the WBÖ

The *Dictionary of Bavarian Dialects in Austria* (Wörterbuch der bairischen Mundarten in Österreich ‘WBÖ’) is a long-term project, whose main goal is the comprehensive lexicographic documentation of the manifold Bavarian base dialects in Austria and South Tyrol. Shortly after its initiation in 1911, the language data in this collection was obtained either indirectly with the help of so-called collectors (“*Sammler*”) on the basis of questionnaires (“*Fragebücher*”) sent out by mail, or directly during field explorations (“*Kundfahrten*”), and was further complemented with excerpts from specialized literature. All data were written down on paper slips and collected in the main catalog (“*Hauptkatalog*”), which contains approximately 3.6 million entries. To date, five volumes of the WBÖ have been published, covering the entries *A-Ezzes*.<sup>1</sup>

---

<sup>1</sup> For more information on the history of the WBÖ cf. Geyer [4] and Reiffenstein [5].

With the main purpose of facilitating and accelerating the process of writing dictionary articles, the hand-written paper slips were entered manually into a TUSTEP system in the 1990's (Barabas et al. [1]) and, subsequently, converted into TEI-XML.

After the relocation of the WBÖ to the department 'Variation and Change of German in Austria' at the Austrian Academy of Sciences (ÖAW) – Austrian Centre for Digital Humanities (ACDH) in December 2016, a new team is working on a revised and modernized conception of the dictionary, which will include a continuation of writing dictionary articles as well as the creation of a web based research platform.

## 2 Database & Content Description

The TUSTEP database system<sup>2</sup> played a major, and beneficial role in the evolution of the DBÖ (Datenbank der bairischen Mundarten in Österreich) project contents. However, this system was reliant on an antiquated and complex database structure which required its own software and less than trivial programming language to search and extract the data. Additionally, results of these searches can often be inexplicably incomplete or inconsistent. Given that TUSTEP is self contained and can only be accessed internally (using the system's native programming language), such errors cannot easily be investigated or resolved in a system-independent manner. Moreover, the system and previous practices were carried out prior to the widespread availability of Unicode leaving the data in serious need of modernization in order to properly represent and make full use of its linguistic contents. Thus given the renewed need to more efficiently access, reuse and preserve this data, as well as to bring it more into line with contemporary principles for best practice in language markup, it was necessary to extract the data out of the increasingly obsolete system and convert it into a format that will both help ensure that moving forward, we were able to meet these needs.

Therefore in order to achieve this, the data was converted to TEI (TEI Consortium [6]) which is widely accepted in the digital lexicographic community as the de facto standard for the encoding of both retro-digitized and born digital dictionaries. As we describe below, the TEI has the capacity to encode the entirety of the legacy existing dataset and all its various data fields.

---

<sup>2</sup> TUSTEP is a set of word processing programs, a tool for scientific processing of text data (<http://www.tustep.uni-tuebingen.de/> and the Handbuch TUSTEP 2017). It was used by the DBÖ team as a database because of its macro capabilities.

**Conversion:** Over a period of a year the database was converted in stages using a series of transformation processes using the XSLT language in which certain aspects of the data structure were addressed sequentially. Between each stage of transformation, both the effects of the transformations and the remaining contents of the data were thoroughly checked semi-manually which allowed us to encounter and investigate and log the remaining flaws in the content needing to be addressed in future stages of the transformation.

**Improvement to the Data Structure:** The conversion process did not only involve the interchange from one data format to the other, nor were the benefits limited to issues related to data access issues endemic to TUSTEP. The improvements were achieved and permitted by: the correction of human errors; enhancements in the data structural efficiency inherent to TEI-XML markup vocabulary (in contrast to TUSTEP); technological advancements since the first digitization in the 1990's and the refinement of certain flaws in the content structure from the original project guidelines.

**Human Error:** Because of the particularities in the TUSTEP data structure and labelling, the size of the database, the duration of the project and the large number of different individuals who worked on the process of digitizing the entries from the notecards into the original database, there was a large degree of irregularity due to idiosyncratic practices as well as simple typing errors. Of the 510 data field tags present in the initial export from TUSTEP, 197 of them were found to be due to human error (either by way of typos or non-adherence to the project guidelines). However, whereas the sheer number of these incorrect tags is large, with a few exceptions, the vast majority of them had less than 10 instances.

**TEI-XML improvement to TUSTEP-Inherent Structural Flaws:** In addition to the structural errors identified and corrected due to human error, the contents adhering to the project guidelines comprised of 313 unique field tags occurring in hundreds of thousands of entries. The sole reason for this extremely high number of different tags was due to the nature of the TUSTEP database structure, which is a flat sequence of unique fields and that has no means of pointing between or expression relations between different specific instances of different fields outside of the name of the field itself. One of the foremost benefits of using TEI, as given that it is an XML vocabulary, it can readily solve this issue by making use of attributes, which can be used for labelling and/or pointing and nested data structure to reduce the excess data complexity necessitated by TUSTEP.

**Numbering:** In a TUSTEP entry the data field tags are simply a single string of uppercase letters (and possibly digits) encased in asterisks, e.g.

“\*HL\*” is the “Hauptlemma”; (“*headword*”) tag. Numerous different fields often could occur more than once, for example an entry could have up to ten dialect forms, and even though the content was the same in nature, and there is no reason a user would ever specifically want to search for a specific numbered instance of the category, in TUSTEP they were required to have unique tags, e.g. \*LT1\*, \*LT2\*, \*LT3\*, etc.

```

===
*LT1* Zügeln
*LT2* ziglen
*LT3* zigln
*****

```

Example 1: Numbering tag labels in TUSTEP

Thus, what was in TUSTEP \*LT1\*...\*LT10\* (all distinct tags), is in TEI represented as <form type="lautung"> and each unique number be expressed using the number attribute: @n.

```

<form type="lautung" n="1">
  <pron notation="tustep">Zügeln</pron>
</form>
<form type="lautung" n="2">
  <pron notation="tustep">ziglen</pron>
</form>
<form type="lautung" n="3">
  <pron notation="tustep">zigln</pron>
</form>

```

Example 2: TEI version of entry with multiple dialect forms

**Nesting:** Where in TUSTEP, there is a complementary or supplementary category that modifies or adds to a field above (e.g. *translations of example sentences, comments, location, miscellaneous notes, references*, among others) these categories need to specify the tag they pertain to within their tag name as well, e.g. “Bedutung Kontext 1” (“*meaning usage context 1*”) would be \*BD/KT1\*.

```

===
*KT1* in Auszug [m.sg4] g.eijn
*BD/KT1* in den Auszug gehen
*****

```

Example 3: TUSTEP entry with complimentary field \*BD/KT1\*

In TEI these relations are encoded as nested elements with the complimentary content of the main field nested within the latter. Given the fact that the sub-ordinate relationships between nested elements and their parent are defined as part of the fundamental data model of XML, the TEI conversions of these contents do not need to maintain and further reference to the target element. In the TEI version, the translations of a usage example is encoded in the definition element <def> and labelled with the language attribute the value of which is the ISO 639-2 code for High German “de”.

```
<cit type="kontext" n="1">
  <quote>in Auszu|g [m,sg4] g;ei;n</quote>
  <def xml:lang="de">in den Auszug gehen</def>
</cit>
```

Example 4: TEI translation of usage context example

**Pointers:** Many fields such as: *meaning*, *usage context*, *context-specific sense*, *references*, *notes*, *etymology*, can apply to one or all of the fields given in the entry, for each specific variant, a unique field tag was required, so there existed tags such as: \*BD/LT2/LT3\*. Any one form could have (1..n) meanings as well which would be represented along the same lines: e.g. \*BD2/LT2/LT3\*<sup>3</sup>.

```
====
*LT1* Hamperl [D,m]
*LT2* Hamperl [D,n]
*BD/LT1/LT2* allzu nachgiebiger Mensch; Siemannldumme
Person; Blödling
*BD2/LT1/LT2* dumme Person; Blödling
*****
```

Example 5: TUSTEP entry with complex referential tags

In certain contexts it is not appropriate to nest in TEI and instead it is better to use pointers to express relations between content. In such cases, this was done in TEI using a pointer attribute-value combination with the

---

<sup>3</sup> The initial export had 123 variants of the “Bedeutung” (“*meaning*”) field and TUSTEP does not allow for partial string searches of tag content, which means that to search for specific content within all of those variants, one would have to specify each one, or just run a string search of the desired contents without specifying the fields resulting in a large number of false positives and greatly increasing run-time.

@corresp making use of the TEI prefix definition (<prefixDef><sup>4</sup>) scheme to point to the specific corresponding content in a predefined structure within an entry. Note also in the example below that in the TEI version the content in brackets from TUSTEP which is the grammatical information ([D,m] ‘diminutive, masculine’ & [D, n] ‘diminutive, neuter’) for each form is moved from the line with the form itself in the TUSTEP to its own element block <gramGrp><gram>.

```
<form type="lautung" n="1">
  <pron notation="tustep">Hamperl</pron>
  <gramGrp><gram>[D,m]</gram></gramGrp>
</form>
<form type="lautung" n="2">
  <pron notation="tustep">Hamperl</pron>
  <gramGrp><gram>[D,n]</gram></gramGrp>
</form>
<sense n="1" corresp="this:LT1 this:LT2">
  <def xml:lang="de">allzu nachgiebiger Mensch;
    Siemannldumme Person; Blödling</def>
</sense>
<sense n="2" corresp="this:LT1 this:LT2">
  <def xml:lang="de">dumme Person; Blödling</def>
</sense>
```

Example 6: Pointing to non-adjacent contents in TEI

**Attribution:** Additionally, certain feilds distinguished the editorial responsibility of its contents e.g. \*ANMO\* “anmerkung original” (“*comment by the original editor*”) and \*ANMB\* “anmerkung bearbeiter” (“*comment by the editor*”). It was common and possible to have combinations of many of these complex tags as well e.g. \*VRWO/BD/LT1\* (“*reference - original editor - meaning form one*”). In TEI this tag feature was converted to the responsibility attribute (@resp).

```
<sense corresp="this:LT1">
  <def xml:lang="de">abfärben; die Farbe abgeben,
    verlieren</def>
</sense>
```

---

<sup>4</sup> <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-prefixDef.html> [accessed on 17th December 2017]

<note type="anmerkung" resp="O" corresp="this:BD">(z.B.:  
die Wand färbt ab)</note>

Example 7: TEI @resp

**Conclusion on Conversion:** Thus, among the most significant achievements of the conversion of the database is the reduction of the number of data field tags from 510 to 37. This, in combination with the use of a BaseX database system using the XQuery language (which are both open source and have extensive online user resources) the conversion to TEI allows us to greatly improve: the process of searching and manipulating the contents of the data are greatly simplified with a greatly improved level of granularity over those of the TUSTEP internal database search system; and our ability to accurately document the contents of our database for new users. Additionally, the TEI community is constantly growing and more and more projects are adopting conversions to and from it thus its use helps in reaching a wider audience. Given the structural improvements to the data, and the fact that the XML markup language and the TEI guidelines are open source, systematically documented, our data now has a much higher degree of long-term sustainability as well as compatibility with potential partner projects.

**Remaining Issues and Ongoing Work in the DBÖ:** As described above, to this point, the contents of DBÖ have been greatly improved in structure, consistency, accessibility have been greatly improved in the conversion. However, due to the legacy data structure, a number of significant issues which inhibit the quality of the resource remain. Some of the most notable of which are as follows.

The transcription notation of both the headwords and similar forms, as well the phonetic forms do not correspond to any standard and in many cases they are not entirely human readable and/or are complex to search for directly, often requiring the use of regular expressions. These are in the process of being normalized. Such changes will allow us to both maintain the linguistically significant morphological segmentation information while allowing users to search and retrieve the contents. The example below shows a complex compound headword what was previously just the form in <orth type="orig"> which will be enhanced as follows:

```
<form type="hauptlemma">
  <orth type="orig">(Amts-pflicht)for-halt</orth>
  <orth type="parsed">
    <seg>
      <seg>Amts</seg><seg>pflicht</seg>
    </seg>
  </orth>
</form>
```

```

</seg>
<seg>for</seg><seg>halt</seg>
</orth>
<orth type="normalized">Amtpflichtforhalt</orth>
</form>

```

#### Example 8: Normalized and segmented headwords in TEI

The phonetic dialect forms will be converted from a TUSTEP interpretation of the original Teuthonista script and characters to fully Unicode Teuthonista: for example, what is currently: “d-es” will be converted to: “dēs”.

Loanwords (which are in fact dialect forms), are expressed in a different category from the rest of the forms. These will be converted and the loanword information will be included in <eytm type="loanword">. Several categories containing multiple distinct fields of information have not been entirely decomposed. There are several thousand entries with no headword, many dialect forms are not directly accessible as many entries have only an example in contextual usage within which the dialect form is not explicitly tagged. Many areas of the data structure are not in line with various international standards for language markup. Finally because of the lack of consistency in both the form-related contents, sense related contents, and the nature of the questionnaires used to elicit the data, there remains a significant gap in the means in which users can search for both semasiological (form-based) and onomasiological (concept-based) contents. To alleviate this we are working on creating a normalized inventory of semantic labels.

### 3 Resuming WBÖ publication & Creation of an Online DB

The planned output of the future WBÖ work is twofold: On the one hand, the WBÖ staff will continue to write “classical” dictionary articles, which will, however, appear in a revised and modernized form. These revisions include a more standardized structure of the articles, a modernized layout, more condensed and generalized information about pronunciation, etymology and geography. For each ‘Hauptlemma’ which will enter the dictionary as a headword, the semantic information is categorized, as well as phonetic variants, geographic distribution and more information essential to the dictionary articles.

Another goal for this project is to create a comprehensive online lexicographic information system, i.e. a (re)search tool for professional linguists and the general public, where users will be able to perform queries



regarding different aspects of the database, both linguistic, (i.e. lemma, sense, etc.), metalinguistic (i.e. geographic location/region) as well as legacy materials such as scans of the original paper slips or scans of questionnaires. This lexicographic information system will be integrated into the SFB “Deutsch in Österreich” (“German in Austria”) research platform (DiÖ [3]), thus providing a multi-perspective approach to language variation in Austria.

Moreover, the articles will be accessible via the online lexicographic information system, which makes them directly linkable with the different types of information stored in the database. In addition to semasiological research which is characteristic for dictionaries (i.e. different meanings connected with one lemma), it will allow users also to perform onomasiological queries, i.e. different linguistic forms connected with the same semantic concept (such as ‘*Fasching*’, ‘*Fastnacht*’ and ‘*Fasnacht*’ meaning carnival).

**Interface of Classical Dictionaries and Digital Humanities:** On the back-end, the data structure that will be used will also be TEI, though, it will involve the creation of a much more complex set of entry templates in order to accommodate the various different data fields common in dictionary articles. While the use of the TEI dictionary module is of course well established in accommodation of both retro-digitized and born digital dictionary content, this usage will represent a rather novel usage of the standard in a two ways.

First, while it is of course common for print dictionaries to be retro-digitized, it is less common (perhaps even unprecedented) for the print dictionary to be compiled, generated and edited first in TEI. Second, given the inherent complexity of the contents of a dialect dictionaries, this usage of the TEI as a digital template (or templates) for the creation of such information provides an opportunity to balance out a number of issues that often are in conflict in such projects. Such issues include: the structural and content demands of print dictionaries and those of the digital data structure (i.e. *best practice in TEI markup*); editorially, the potential conflict in the usability for non-experts tools needed to edit and create articles in TEI directly versus those used in traditional practice (e.g. *basic word processing*).

## 4 Conclusion

In our paper we present a large historical database of Bavarian dialects (DBÖ) and give an example-based overview of the TEI structure, contents and remaining issues pertaining to the revised TEI-XML dataset. Additionally we introduce the plans already underway for a revived print

version of the WBÖ and the creation of an online publicly searchable version of the database which will both be structured and edited within task-specific TEI templates. Finally, we discuss the challenges we are still facing, and the approaches we are taking and considering in order to address such challenges. Our project offers potential insights for the use of the TEI vocabulary for such tasks.

## References

- [1] Barabas, B., Hareter-Kroiss, C., Hofstetter, B., Mayer, L., Piringer, B. and Schwaiger, S. (2010) Digitalisierung handschriftlicher Mundartbelege. Herausforderungen einer Datenbank. In *Germanistische Linguistik* 199-201, pp. 47–64.
- [2] Bowers, J. (2017) *TEI conversion of Bavarian dialect lexical resources: insights, observations, challenges, next steps*. Presented at the COST-ENeL Meeting, Budapest.
- [3] DiÖ (2017) Task Cluster E: Collaborative Online Research Platform. In *DiÖ-Online*. Available at: <https://dioe.at/en/article-details/> [accessed on 14<sup>th</sup> October 2017]
- [4] Geyer, I. (2004) Arbeitsbericht zum Wörterbuch der bairischen Mundarten in Österreich. In: Gaisbauer, S. and H. Scheuringer eds. *Linzerschnitten. Beiträge zur 8. Bayerisch-österreichischen Dialektologentagung, zugleich 3. Arbeitstagung zu Sprache und Dialekt in Oberösterreich, in Linz, September 2001*, pp. 583–588, Linz.
- [5] Reiffenstein, I. (2004) Die Geschichte des “Wörterbuchs der bairischen Mundarten in Österreich” (WBÖ). Wörter und Sachen im Lichte der Kulturgeschichte. In: Hausner, I. and Wiesinger, P. eds. *Deutsche Wortforschung als Kulturgeschichte. Beiträge zum Symposium “90 Jahre Wörterbuchkanzlei” der Österreichischen Akademie der Wissenschaften, Wien, 25-27. September 2003*, pp. 1–13, Wien.
- [6] TEI Consortium, eds. (2016) *TEI P5: Guidelines for Electronic Text Encoding and Interchange. [Version 3.1.0]*. [Last updated on 15th December 2016]. TEI Consortium. Available at: <http://www.tei-c.org/Guidelines/P5/> [accessed on 13<sup>th</sup> February 2017].

# Manuscripts in Time and Space: Experiments in Scriptometrics on an Old French Corpus\*

Jean-Baptiste Camps

Centre Jean-Mabillon

École nationale des chartes | Paris Sciences & Lettres

E-mail: jbcamps@hotmail.com

## Abstract

Witnesses of medieval literary texts, preserved in manuscript, are layered objects, being almost exclusively copies of copies. This results in multiple and hard to distinguish linguistic strata – the author’s *scripta* interacting with the *scriptae* of the various scribes – in a context where literary written language is already a dialectal hybrid. Moreover, no single linguistic phenomenon allows to distinguish between different *scriptae*, and only the combination of multiple characteristics is likely to be significant [9] – but which ones? The most common approach is to search for these features in a set of previously selected texts, that are supposed to be representative of a given *scripta*. This can induce a circularity, in which texts are used to select features that in turn characterise them as belonging to a linguistic area. To counter this issue, this paper offers an unsupervised and corpus-based approach, in which clustering methods are applied to an Old French corpus to identify main divisions and groups. Ultimately, scriptometric profiles are built for each of them.

## 1 Introduction

Study on the diatopic variation of medieval French texts rests on the distinction proposed by Remacle [20] between *scripta*, written language (*Schriftsprache*), and dialect, spoken language, the latter mostly inaccessible to us. Based on his study of Walloon, this distinction was put forward as a mean to reconcile the difference he observed between the very characterized modern dialect and the medieval written texts from the area, presumably less marked by local traits. In the medieval *scripta*, he argued, the distinctive traits inherited from spoken Walloon would be present only by mistake or ignorance. Consequently, he formulated the apparently self-contradictory hypothesis that “1. the *scripta* was the result of a local development, 2. the *scripta* was a common language whose essential elements were found

---

\*A digital appendix to this paper is available on Zenodo, DOI: 10.5281/zenodo.1117924. My gratitude, for discussion on this subject over the years, goes to Frédéric Duval, Martin D. Gleßgen, Hans Goebel and Achim Stein. I also thank the anonymous reviewers for their insightful advice.

in most spoken dialects of the *langue d'oïl*" (my translation). This distinction is now commonly accepted though sometimes criticised because it sets in stone our inability to ever gain insights into the reality of medieval dialects [4]. For the scholar who wants to date and localise the *scripta* of medieval texts, this implies that he will face a language that was never spoken as such and the very building blocks of which might be made of elements taken from various dialectal areas, maybe even a *koinè*, in which truly local traits are only marginal [7, p. 40].

The exact reality of this notion of *scripta* is still debated, but, as a working definition, we will take it as the written language, practised by a restricted number of literates, around scriptural centres (e.g. chancelleries), and supposedly conceived to allow for a broader comprehension than oral dialects, but still containing traits that can be geographically assigned to a specific area. The possible connexion between the main modern dialectal areas (as delimited by modern dialectologists) and the geographical hold of medieval documentary *scriptae* can be estimated due to the fact that administrative documents (charters, for instance) are usually dated (time and place date). It seems confirmed by Goebel's work [12].

The case is even more complex for literary witnesses<sup>1</sup>. While documentary texts (charters, wills, ...) are practical documents, often of only local interest, most literary texts were made to be able to circulate through different linguistic areas, written by the more knowledgeable amongst the population, and influenced by the written codes of Latin [7, p. 41]. Sociolinguistics played a part, as well as factors related to production of books, such as the implantation of workshops. Variation in prestige between dialects led to difference in behaviour among writers, up to the point where some *scriptae* were judged distinctive of a genre, and its features imitated, like Western dialects or Picard for epic texts [1]. Two scribes working in the same workshop but from different origin might produce a text with different features. As such, localising the *scripta* of a witness does not mean as much finding its place of origin as identifying the linguistic inclinations of its writers [26]. But the major difficulty is of another nature yet: literary witnesses are layered objects, in which the language of the author interacts with each scribe's, up to the point where it is a very delicate task to assign any trait to a given layer, especially since any layer might already have included an alternation of forms or mixed forms [20].

As a consequence, it is very hard for dialectologists to determine isoglosses, or more precisely isographs [17, p. 166], that could clearly separate different *scriptae*. In fact, it is likely that no single trait can be used to define a *scripta* [9, p. 315]: most isographs are shared among several – usually neighbouring – regions [14, p. 65]. Even for the rare isographs that would be very distinctive, the information they provide is blurred by the hybrid nature of *scriptae* or the stratification of textual

---

<sup>1</sup>I define *witness* as a given instance of a text, as preserved in a particular document (usually, a manuscript) that is accessible to us. See Duval [6] for an account on the meaning of the terms *text* and *witness* ("texte" and "témoin") in (neo-lachamannian) textual criticism. It allows me to distinguish between the more abstract work (e.g. the story of Roland and the battle at Roncevaux) and its expression in particular texts (i.e. the *Chanson de Roland* or the *Cân Roland*), attested in witnesses (e.g. *O*), preserved in documents (the ms. Digby 23).

witnesses. As a consequence, only a combination of traits, individually common with other *scriptae*, in a given relative frequency, makes the distinction possible. This has led to an emphasis put on quantification, and eventually on statistical multivariate analysis [9, p. 317]. This approach is named “dialectometry” since Séguy [22], or, better in our case, “scriptometry”. It is defined by Goebel [10, p. 60-61] as an alliance between linguistic geography and clustering, and it shares some similarities with, for instance, stylometry and other historical text analysis fields. More generally, it can be defined as *the measure of scriptologic features*. As an exploratory approach, its goal is to reveal underlying structures that escape close reading analysis and are supposed to be more important than the superficial structures visible in the traditional maps of linguistic atlases [10, 11].

The dialectometric work of Dees or Goebel have been mostly founded on the listing of lexical, phonological or morpho-syntactical traits (“taxation” [10]), and the analysis of the resulting data. The atlases produced by Dees’ team [3, 5] so include a series of maps that each present a quantified opposition between two groups of forms, and can be used [11, 3] as a matrix for computational analysis, both to study the underlying structures of dialectal variation or to locate a new text by confrontation with the already localised ones or to cartography similarities between regions and map dialectal areas [12, 3, 4, 5].

The work of Dees and his Amsterdam School and, after him, of Goebel and the Salzburg School, have given the rise to a more systematic and objective way to study medieval *scriptae* (for an historical synthesis, see Volker [27, chap. 2, p. 9-79]). Yet, an issue of circularity might still exist, since previous analyses usually based themselves on the localisation assigned to witnesses to identify linguistic areas and scriptological features. I would like to suggest a less supervised approach to the scriptometric analysis of the witnesses of a specific Old French epic genre, the *chansons de geste*. My aim will be to identify main divisions in the corpus and to create profiles for each of them, and to verify both customary separations between *scriptae* and the belonging of each individual witness to one of them.

## 2 Corpus and Method

In order to limit biases caused by stylistic, thematic or generic variations, this study will be limited to a single genre, the *chansons de geste*. Previous exploratory analyses, not shown here, on a multi-generic corpus of 299 texts, did confirm that generic differences interacted with linguistic boundaries and created too much noise. Authorship related biases are hard to avoid, but might be counteracted by the very graphic variation observed in the witnesses, a problem in the stylometric analysis of medieval vernacular texts. The corpus of *chansons* used here is composed of 50 witnesses (see app. A), with 1 104 296 tokens (geometric mean, 12 016, median, 11 490; min., 387; max., 217 942). The tokens are distributed between 52 202 forms (long-tail distribution, with 25 811 hapaxes; geom. mean of 2,57 occurrences, median, 2; 3<sup>rd</sup> quartile, 4). Editions were chosen for their use of a base

witness (“copy-text”) – the emphasis here being on the witnesses and not on the original text – as well as for their availability in digital form. The selection of witnesses was done empirically to have the largest corpus with a representativity of several putative regions of origin. Yet, its heterogeneity is a limitation<sup>2</sup>.

Variation in editorial practice regarding the allographs **i/j** or **u/v** and their transcription led me to map all of them on **i** and **u**. More generally, to avoid interferences with paleographic variation and perform on the graphematic level, all allographs (including “capitals”) were normalized and all abbreviations expanded. The latter might be problematic, as it makes the process dependent on the choices of the editors, and can induce a bias, given that the norm is to use the majority unabbreviated form for expansion, inducing a distortion favorable to this majority form as compared to the coexisting alternative ones [18, p. 33].

It is to be noted that the exclusion of allographic variation is an important simplification of the reality of textual witnesses, done both for contextual (the unavailability of consistent information) and theoretical reasons, based on the assumption that the variation in use of variant letter forms is more dependent on scribe’s idiosyncrasies or script variation (*textualis*, *cursiva*, etc.), sometimes termed “scribal mode” [15, 16]. In the terminology offered by McIntosh for his “scribal profiles”, this means we will restrict ourselves to the “linguistic” by opposition to the “graphic” components [15], that is “graphematic”, opposed to “allographic” in the terminology retained here [25]. Yet, given the interest of this latter kind of variation for dating and localising witnesses or identifying scribes, I have undertaken elsewhere to build a corpus of allographic transcriptions and analyse them using similar techniques<sup>3</sup>. Another dimension of these witnesses that we will not take into account concerns the alterations to the content of the text during its transmission (variants), that is the way in which the behaviour of the scribe alters the text of his model to result in a new copy, that we could term the “diasystemic” component, after Segre’s definition [21].

If previous scriptometric works were based on the “taxation” of a defined list of features, I chose to use a bag-of-words approach on the graphic forms of the texts, in order to avoid inducing *a priori* the features of the profiles. The main drawback is that occurrences of an identical phenomenon (e.g. graphs of a given diphthong) will be divided between all the forms that attest it. It will also prevent any syntactic feature to be taken into account and will limit the analysis to graphic or morphologic features. On the other hand, more limited habits, on the particular graph of a given lemma, will be fully accounted for. Lexical variation, important for the localisation of texts through the identification of regional words [7, p. 93],

---

<sup>2</sup>I intend to work, in the coming years, on the constitution of a corpus as exhaustive as possible of epic witnesses (transcriptions, critical editions, manuscript descriptions). The first few texts, encoded in TEI XML, are available on Github [8]. The data, in csv, used for this paper, are available with scripts to reproduce analysis, on the Zenodo repository.

<sup>3</sup>More details can be found in [2, chap. 2], including unsupervised clustering and allographic profiles (sect. 2.4), with a digital appendix giving access to the datasets and analysis procedures. An updated version of the corpus is available in [8].

will also be analysed this way, even if it makes the analysis highly dependent on content-based variation. For this last reason, the database will be constituted of word rather than n-grams frequencies.

To limit content-based biases (and issues related to the non-Gaussian distribution of word-frequencies), only the most frequent words (MFW) are retained for analysis, an approach common in stylometry as well. Proper names were removed. This selection also leads to focusing the analysis on the dominant linguistic stratum (scribal or otherwise). Since no precise guidelines exist on the number of MFW to retain, robustness of the results will be checked with different levels of selection.

To cluster the witnesses, hierarchical clustering was retained, a common analysis in scriptometrics [10, 12]. We do not yet possess guidelines on the effectiveness of various linkage criteria or distance measures in this field. Experimenting with a variety of those, to retain the one that would seem the best to me, though a heuristic approach advocated by Goebel [10, p. 85], would induce a validation bias. As a consequence, I retained Ward's method, because it relies on the barycentre of the data clouds and allows for the constitution of balanced and coherent clusters, often referred to as *types*, as it minimises intra-cluster variation and maximises inter-cluster variation [24]. It is usually claimed that only squared euclidean distance is correct to use with Ward's linkage, because it relies on computations in euclidean space. Yet, recent research by Strauss and von Maltitz [24] seems to demonstrate that it can be generalised to use with Manhattan distance, and that this metrics outperforms euclidean in what regards the classification of (indo-european) languages, a statement that agrees with previous research in computational phonology applied to the clustering of (Dutch) dialects [19], or with the supposed greater efficiency of Manhattan distance with highly dimensional data.

### 3 Results

Results were mostly stable with between 600 and 3000 MFW, as well as the agglomerative coefficient (between 0.83 and 0.8). The main divisions (fig. 1) are consistent with scriptological knowledge<sup>4</sup>. The first opposes supposedly Anglo-Norman witnesses to Continental ones. Inside the Anglo-Norman group, a division opposes older (XII or XIII<sup>1/2</sup>) to more recent (XIII-XIV) witnesses, arranged in an imperfect chronological order. The orientation is in itself interesting as it seems to confirm the hypothesis that later Anglo-Norman texts, written in a fossilising linguistic context, were more subject to continental norm. The diachronic division of the Anglo-Norman group might also reveal the weakness of diatopic variation in this *scripta*, in a country where "*Normannica lingua, que adventitia est, univoca*

---

<sup>4</sup>Following preliminary experiments, a few too short (<2000 words) witnesses were removed, because their inclusion tended to slightly twist the analysis. Nonetheless, their placement was consistent with the rest of the clustering: Aspreme\_C was placed in the Anglo-Norman cluster, among witnesses from the middle of the XIII<sup>th</sup> century, at an intermediary position between witnesses of earlier or later texts, just on the left of MacaireA12B, whose placement was also consistent with chronology; the CharroiSch\_fragm was in the Southern Lorraine group, with CharroiSch\_D and PriseCordD; Fier\_V was in the Lorraine/Burgundy group. See the online appendix.

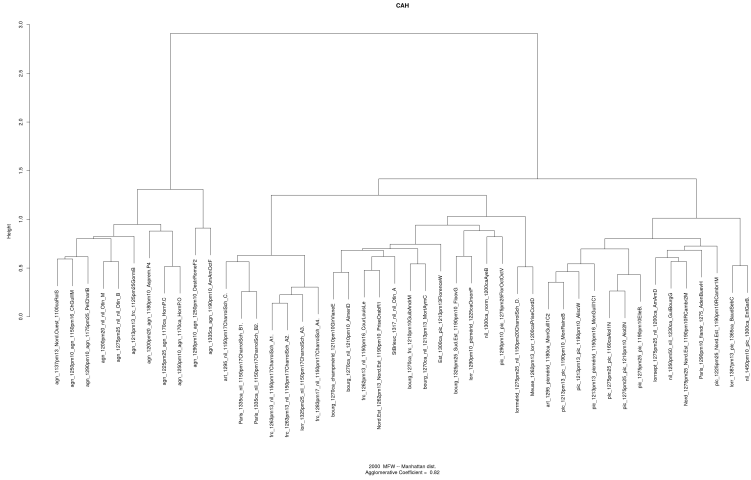


Figure 1: Hierarchical clustering of the *Geste* corpus (Ward’s method, Manhattan dist., 2000 MFW, relative freq.)

*maneant penes cunctos*” (Ranulf Higden, *Polycronicon*, lib. I, cap. 59). The second division, considerably lower, creates a separation within the continental groups, namely dividing Picard witnesses of Picard texts from the rest.

The third division isolates mostly Central witnesses, but might also be due to authorial attraction between copies of the same text, that are even distributed between witnesses of the *A*, *B* and *C* versions (not the *D*). This might nonetheless have a linguistic sense, since *A1* and *A2* (and probably *A4*), for instance, are known to come from the same workshop [26, p. 434-436], as well as *B1* and *B2*.

Inside the group containing the rest of the Continental witnesses, which are mostly Eastern (or Lotharingian), divisions are weaker. Nonetheless, three subgroups can be individuated: witnesses from southern Lorraine (right), Burgundy (left), and Lorraine (centre). Many of the apparent exceptions can be explained and concern witnesses whose origin is subject to debate or need rectification. A subgroup of witnesses from Northern Lorraine or North-East appeared in the centre of this subgroup on some of the analyses (AmAmD, GuiBourG, RCambr), but are here grouped with Picard witnesses, maybe because one of them (RCambr) is a Northern copy of a text from the North-East.

Once groups are constituted, linguistic profiles for each of them can be built, at different levels, by estimating which features are the most characteristic with the values-test described by Lebart, Morineau et Piron [13, p. 181-184]<sup>5</sup>, giving us an

<sup>5</sup>The values-test is done by comparing  $\bar{X}_k$ , the mean of variable  $X$  in category  $k$  to the overall mean  $\bar{X}$ , while taking into account the variance  $s_k(X)$  of this variable inside the class:  $t_k(X) = \frac{\bar{X}_k - \bar{X}}{s_k(X)}$ .



	v.test	mean in cat.	overall mean	sd in cat.	overall sd		v.test	mean in cat.	overall mean	sd in cat.	overall sd
Group 1 ( <i>Anglo-Norman</i> )						Group 4 ( <i>Picard</i> )					
pur	5.8438	0.0067	0.0018	0.0026	0.0032	ains	5.6322	0.0016	0.0005	0.0003	0.0007
sunt	5.7222	0.0058	0.0016	0.0024	0.0028	tous	5.4891	0.0021	0.0006	0.0006	0.0010
ad	5.6188	0.0120	0.0031	0.0056	0.0060	passes	5.2743	0.0002	0.0000	0.0001	0.0001
mei	5.5343	0.0019	0.0005	0.0010	0.0010	chou	5.2216	0.0009	0.0002	0.0006	0.0005
sur	5.5101	0.0044	0.0012	0.0021	0.0022	trestous	5.0875	0.0003	0.0001	0.0001	0.0002
lur	5.4663	0.0040	0.0010	0.0021	0.0021	tout	5.0120	0.0043	0.0015	0.0010	0.0020
tut	5.4522	0.0045	0.0012	0.0023	0.0023	sarrasins	4.9654	0.0004	0.0001	0.0003	0.0002
al	5.3361	0.0072	0.0022	0.0034	0.0036	sains	4.9536	0.0004	0.0001	0.0002	0.0002
e	5.3131	0.0357	0.0108	0.0127	0.0179	toutes	4.9496	0.0004	0.0001	0.0001	0.0002
sun	5.2683	0.0070	0.0018	0.0041	0.0037	commanda	4.9074	0.0001	0.0000	0.0001	0.0001
seit	5.2186	0.0020	0.0006	0.0012	0.0011	cha	4.9023	0.0006	0.0001	0.0004	0.0003
dunt	5.1968	0.0018	0.0005	0.0011	0.0010	mieus	4.8405	0.0004	0.0001	0.0003	0.0002
od	5.1781	0.0033	0.0009	0.0019	0.0017	ochis	4.7118	0.0002	0.0000	0.0002	0.0001
si	5.1214	0.0186	0.0136	0.0030	0.0037	no	4.6579	0.0005	0.0002	0.0004	0.0003
mun	5.0508	0.0018	0.0005	0.0012	0.0010	lieu	4.6264	0.0002	0.0001	0.0002	0.0001
funt	5.0045	0.0008	0.0002	0.0006	0.0005	uausist	4.6239	0.0002	0.0000	0.0001	0.0001
reis	4.9249	0.0046	0.0012	0.0033	0.0026	espiel	4.6180	0.0004	0.0001	0.0003	0.0002
seignurs	4.9082	0.0009	0.0002	0.0006	0.0005	laisa	4.6063	0.0001	0.0000	0.0001	0.0001
rei	4.8912	0.0038	0.0010	0.0027	0.0022	dolans	4.5675	0.0003	0.0001	0.0002	0.0002
a	-4.8186	0.0246	0.0328	0.0050	0.0065	chi	4.5667	0.0009	0.0003	0.0006	0.0005
droit	-4.8320	0.0001	0.0009	0.0002	0.0006	toute	4.5588	0.0009	0.0004	0.0002	0.0004
qui	-4.8793	0.0037	0.0101	0.0032	0.0050	cief	4.4868	0.0007	0.0002	0.0005	0.0004
mon	-4.9032	0.0003	0.0023	0.0006	0.0015	ainc	4.4662	0.0008	0.0002	0.0005	0.0004
et	-4.9212	0.0093	0.0352	0.0195	0.0201	mais	4.4656	0.0052	0.0023	0.0014	0.0023
sont	-4.9557	0.0003	0.0028	0.0009	0.0019	ceual	4.4543	0.0006	0.0002	0.0005	0.0003

Table 1: Scriptometric profiles for the Anglo-Norman (left) and Picard groups (right, without the Northern Lorraine subgroup), giving the 25 most characteristic forms (in positive or negative), rounded to 4 decimals

insight as to how clusters were constituted. To do so, the `catdes` function of the `FactoMineR` package by Francois Husson will be used.

The profiles for Anglo-Norman (table 1) shows known features of this *scripta*, like “the replacement of Standard Medieval French (SMF) *o* or *ou* in all positions by *u*”, “the retention of *ei* where *SMF* develops *oi*”, and “the retention of dentals in 12<sup>th</sup>-century texts”[23, p. 45-46]. Some are not usually cited: the use of *e* (not *et*), for instance, or *al* (not *au*). The Picard group is also distinctively characterized by its palatalizations, its possessive of 1st and 2nd pers. pl. without *-s* at the singular regime case or nominative plural (*no*, *vo*), the use of *tout/tous* (not *tuit*) at the masc. pl. nom., as well as the feminine *toutes*, or the finales in *-s* instead of *-z*.

## 4 Further research

For the future of this research, an important aspect is the constitution of a corpus more homogeneous in terms of editorial practice. The extension of the corpus, by the addition of new witnesses, would make possible more focused analyses, with, for instance, more restricted chronological limits. The study of the relevance, both from a mathematical and philological point of view, of other metrics, is also a lead for future improvements. It has been shown here, that, though interesting results on the grouping of the witnesses of literary texts can be obtained, their stratified nature remains an obstacle, causing some witnesses to switch groups according to either the presumed *scripta* of their scribe, or the language of the author of the original text. Finding a more satisfying way to account for this phenomenon would be paramount to the scriptometric study of the tradition of medieval literary texts.

## References

- [1] Bennett, Philip E., 2003, “Le Normand, le picard et les koïnés littéraires de l’épopée aux XII<sup>e</sup> et XIII<sup>e</sup> siècles”, *Bien Dire et Bien Aprandre*, 21, p. 43-56.
- [2] Camps, Jean-Baptiste, 2016, *La ‘Chanson d’Otinél’: édition complète du corpus manuscrit et prolégomènes à l’édition critique*, dir. Dominique Boutet, thèse de doct., Paris-Sorbonne, DOI: 10.5281/zenodo.1116736.
- [3] Dees, Anthonij, Van Reenen, Pieter and De Vries, Johan A., 1980, *Atlas des formes et des constructions des chartes françaises du XIII<sup>e</sup> siècle*, Tübingen, DOI: 10.1515/9783111328980.
- [4] Dees, Anthonij, 1985, “Dialectes et scriptae à l’époque de l’ancien français”, *Revue de Linguistique Romane*, 49-193, p. 87–117.
- [5] Dees, Anthonij, Dekker, Marcel, Huber, Onno and Van Reenen-Stein, Karin, 1987, *Atlas des formes linguistiques des textes littéraires de l’ancien français*, Tübingen, DOI: 10.1515/9783110935493.
- [6] Duval, Frédéric, 2017, “Pour des éditions numériques critiques”, *Médiévales*, to be published.
- [7] Duval, Frédéric, 2009, *Le français médiéval*, Turnhout.
- [8] *Geste: un corpus de chansons de geste*, ed. Jean-Baptiste Camps, 2016-..., Paris, <http://github.com/Jean-Baptiste-Camps/Geste>.
- [9] Goebel, Hans, 1995, “Les scriptae françaises III. Normandie”, *Les différentes langues romanes et leurs régions d’implantation du Moyen Âge à la Renaissance*, ed. Günter Holtus, et al., Berlin, New York.
- [10] Goebel, Hans, 2003, “Regards dialectométriques sur les données de l’Atlas linguistique de la France’ (ALF): Relations quantitatives et structures de profondeur”, *Estudis Romànics*, 25, p. 59-120.
- [11] Goebel, Hans, 2008, “Sur le changement macrolinguistique survenu entre 1300 et 1900 dans le domaine d’oïl: une étude diachronique d’inspiration dialectométrique”, *Dialectologia*, 1.
- [12] Goebel, Hans, 2011, “L’aménagement scripturaire du Domaine d’Oïl médiéval à la lumière des calculs de localisation d’Anthonij Dees effectués en 1983: une étude d’inspiration scriptométrique”, *Medioevo romanzo*, Seminario 2011: Il problema della scripta, Venezia, <http://www.medioevoromanzo.it/modules/content/index.php?id=14>.
- [13] Lebart, Ludovic, Morineau, Alain and Piron, Marie, 1995, *Statistique exploratoire multidimensionnelle*, Paris.

- [14] Lusignan, Serge, 2004, *La langue des rois au Moyen Âge: le français en France et en Angleterre*, Paris.
- [15] McIntosh, Angus, 1975, "Scribal profiles from Middle English texts", *Neuphilologische Mitteilungen*, 76, p. 218-235.
- [16] McIntosh, Angus, 1974, "Towards an inventory of Middle English scribes", *Neuphilologische Mitteilungen*, 75, p. 602-624.
- [17] Monfrin, Jacques, 2001, "Le mode de tradition des actes écrits et les études de dialectologie", *Études de philologie romane*, Geneva, p. 145–173.
- [18] Morin, Yves-Charles, 2007, "Histoire du corpus d'Amsterdam: le Traitement des données dialectales", *Le Nouveau Corpus d'Amsterdam: actes de l'atelier de Lauterbad, 23-26 février 2006*, ed. Pierre Kunstmann and Achim Stein, Stuttgart, p. 9-27.
- [19] Nerbonne, John, and Heeringa Wilbert, 1997, "Measuring dialect distance phonetically", *Workshop on Computational Phonology, Special Interest Group of the ACL*, p. 11–18.
- [20] Remacle, Louis, 1948, *Le Problème de l'ancien wallon*, Liège, URL: <http://books.openedition.org/pulg/338>.
- [21] Segre, Cesare, 1976, "Critique textuelle, théorie des ensembles et diastème", *Bulletin de la classe des lettres et des sciences morales et politiques de l'Académie royale de Belgique*, 62, p. 279-92.
- [22] Séguy, Jean, 1973, "La dialectométrie dans l'Atlas linguistique de la Gascogne", *Revue de Linguistique romane*, 37, p. 1-24.
- [23] Short, Ian, 2007, *Manual of Anglo-Norman*, London.
- [24] Strauss, Trudie, and Maltitz, Michael Johan von, 2017, "Generalising Ward's Method for Use with Manhattan Distances", *Plos One*, 12-1, e0168288, DOI: 10.1371/journal.pone.0168288.
- [25] Stutzmann, Dominique, 2011, "Paléographie statistique pour décrire, identifier, dater... Normaliser pour coopérer et aller plus loin?", *Kodikologie und Paläographie im digitalen Zeitalter 2*, ed. Franz Fischer, et al., Norderstedt, p. 247-277, <https://halshs.archives-ouvertes.fr/halshs-00596970/>.
- [26] Tyssens, Madeleine, 1990, "Typologie de la tradition des textes épiques: les poèmes français", *Memorias de la Real Academia de Buenas Letras de Barcelona*, 22, p. 433-446.

- [27] Völker, Harald, 2003, *Skripta und Variation: Untersuchungen zur Negation und zur Substantivflexion in altfranzösischen Urkunden der Grafenschaft Luxemburg (1237-1281)*, Tübingen (doct. diss., Univ. of Trier).

## A Corpus

Sources: AND = *Anglo-Norman Source Texts*, ed. David A. Trotter, William Rothwell, Geert De Wilde, and Heather Pagan, Aberystwyth and Swansea, 2001, <http://www.anglo-norman.net/sources/>. GESTE [8]. NCA = *Nouveau Corpus d'Amsterdam: corpus informatique de textes littéraires d'ancien français (ca 1150-1350)*, ed. Anthonij Dees, Achim Stein, Pierre Kunstmann, and Martin Dietrich Gleßgen, Stuttgart, <http://www.uni-stuttgart.de/lingrom/stein/corpus>. OTA = *The University of Oxford Text Archive*, ed. University of Oxford IT Services, s. d., <http://ota.ox.ac.uk/>. TFA = *Textes de français ancien*, ed. Pierre Kunstmann and Mark Olsen, 2003, Ottawa, <http://artfl-project.uchicago.edu/content/tfa>. WIKIS = *Wikisource*, ed. Wikimédia Foundation, <http://en.wikisource.org/>.

We follow, when they exist, the identifier given in Möhren, Frankwalt, and Miller, Elena, 2010, *DEAFBibEl*, Heidelberg, [http://www.deaf-page.de/bibl\\_neu.php..](http://www.deaf-page.de/bibl_neu.php..)

Source	DEAF	ms base	Ed	placeWit	dateWit	placeText	dateText
TFA	AdenBuevH	Ars. 3142	Henry, 1953	Paris	1290pm10	flandr	1275
OTA	AimeriD	BL Roy. 20 B.XIX	Demaision, 1852	bourg	1270ca	nil	1210pm10
NCA+TFA	Aiol1NDeb	BnF fr. 25516	Normand et al., 1877	pic	1275pm25	pic	1160ca
TFA	Aiol2N	BnF fr. 25516	Normand et al., 1877	pic	1275pm25	pic	1210pm10
OTA	AliseW	Ars. 6562	Wienbeck et al., 1903	pic	1213pm13	pic	1190pm10
NCA+TFA	AmAmD	BnF fr. 860	Dembowski, 1969	lorrsept	1275pm25	nil	1200ca
AND	AmAmOctF	BL Roy. 12 C.XII	Fukui, 1990	agn	1335ca	agn	1190pm10
GESTE	Asprem C	Clerm.-Fer. AD 1F2	Camps	agn	1250pm16	agn	1180pm10
GESTE	Asprem P4	BnF, NAF 5094	Albarran & Camps	agn	1200pm20	agn	1180pm10
NCA	AyeB	BnF fr. 2170	Borg, 1967	nil	1300ca	norm	1200ca
TFA	BaudSebC	BnF fr. 12552	Crist, 2002	lorr	1387pm13	pic	1365ca
NCA	CharroiSch A1*	BnF fr. 774	Schoesler	frc	1263pm13	nil	1150pm17
NCA	CharroiSch A2*	BnF fr. 1449	Schoesler	frc	1263pm13	nil	1150pm17
NCA	CharroiSch A3*	BnF fr. 368	Schoesler	lorr	1325pm25	nil	1150pm17
NCA	CharroiSch A4*	Trivulz. 1025	Schoesler	frc	1283pm17	nil	1150pm17
NCA	CharroiSch B1*	BL Royal 20D XI	Schoesler	Paris	1335ca	nil	1150pm17
NCA	CharroiSch B2*	BnF fr. 24369-70	Schoesler	Paris	1335ca	nil	1150pm17
NCA	CharroiSch C*	Boul.-s.-M., BM 192	Schoesler	art	1295	nil	1150pm17
NCA	CharroiSch D*	BnF fr. 1448	Schoesler	lorrmérid	1275pm25	nil	1150pm20
NCA	CharroiSch fr.*	BnF NAF 934	Schoesler	nil	1250pm50	nil	1150pm17
TFA	ChGuillIM	BL Add. 38663	McMillan, 1949	agn	1250pm10	agn	1150pm16
TFA	CourLouisLe	BnF fr. 1449	Lepage, 1978	frc	1262pm13	nil	1150pm16
AND	DestrRomeF2	Hann. IV.578	Formisano, 1990	agn	1290pm10	agn	1250pm10
NCA	ElieB*	BnF fr. 25516	P. Bloem	pic	1275pm25	pic	1190pm10
TFA	EnfGarB*	BnF fr. 1460	A. Kostka, 2002	nil	1450pm10	pic	1300ca
GESTE	Fier-V	BAV Reg. lat. 1616	Camps	StBrieuc	1317	nil	1190ca
GESTE	FloovG	Montp., F. Méd. 441	Guessard, 1858	bourg	1325pm25	Sud-Est	1190pm10
NCA	FlorenceW	BnF NAF 4192	Wallenskoeld, 1907	Est	1300ca	pic	1213pm13
NCA	FlorOctOctV	Bodl. Hatton 100	Vollmoeller, 1883	pic	1290pm10	pic	1275pm25
NCA	GirVianeE	BL Roy. 20 B XIX	Van Emden, 1977	bourg	1270ca	champmérid	1210pm10
NCA	GormB	Brux., BR port. II 181	Bayot, 1931	agn	1213pm13	frc	1125pm25
NCA	GuibAndrM	BL Roy. 20 B XIX	Melander, 1922	bourg	1270ca	frc	1210pm10
GESTE	GuibRougG	Tours, BM 937	Guessard, 1858	nil	1250pm50	nil	1230ca
AND	HornP-C	Cambr. FF.VI.17	Pope, 1955	agn	1225pm25	agn	1170ca
AND	HornP-O	Bodl. Douce 132	Pope, 1955	agn	1250pm10	agn	1170ca
GESTE	MacaireAl2B	fragm. Loveday	Baker, 1915	agn	1250pm50	nil	1250pm50
TFA	MonGuill1C1	Ars. 6562	Cloetta, 1906	pic	1213pm13	picmérid	1150pm16
TFA	MonGuill1C2	Boul.-s.-M., BM 192	Cloetta, 1906	art	1295	picmérid	1180ca
TFA	MonRaineB	Ars. 6562	Bertin, 1973	pic	1213pm13	pic	1190pm10
WikiS	MortAymC	BL Roy. 20 B.XIX	Couraye, 1884	bourg	1270ca	nil	1213pm13
NCA	OrsonP	BnF NAF 16600	Paris, 1899	lorr	1290pm10	picmérid	1225ca
GESTE	OtinC A	Reg. lat. 1616	Camps	StBrieuc	1317	Nord-Est?	nil
GESTE	OtinC B	Bodmer 168	Camps	agn	1275pm25	Nord-Est?	nil
GESTE	OtinC M	BnF NAF 5094	Camps	agn	1200pm20	Nord-Est?	nil
Divers	PelCharlB	BL Roy. 16 E.VIII	Bonafin, 1987	agn	1290pm10	agn	1175pm25
NCA	PriseCordD	BnF fr. 1448	Densuianu, 1896	Meuse	1262pm13	lorr	1200ca
TFA	PriseOrabR1	BnF fr. 774	Régnier, 1986	Nord-Est	1262pm13	Nord-Est	1190pm10
NCA	RCambr1M	BnF fr. 2493	Meyer et al., 1882	pic	1225pm25	Nord-Est	1190pm10
NCA	RCambr2M	BnF fr. 2493	Meyer et al., 1882	Nord	1275pm25	Nord-Est	1190pm10
NCA	RoIS	Bodl. Digby 23	Segre, 1971	agn	1137pm13	Nord-Ouest	1100ca

# Stemmatology: an R package for the computer-assisted analysis of textual traditions\*

Jean-Baptiste Camps<sup>†</sup> and Florian Cafiero

1. Centre Jean-Mabillon

École nationale des chartes | Paris Sciences & Lettres

2. Institut Interdisciplinaire d'Anthropologie du Contemporain

École des hautes études en sciences sociales | Paris Sciences & Lettres

## Abstract

Given a set of witnesses of a text, determining their relations and reconstructing a *stemma codicum* is one of the fundamental purposes of textual criticism and philology. For this task, various computer-assisted procedures and methods have been described since the 1950's, some elaborating on traditional principles (Lachmannian, Queninian...), some borrowed from other fields such as phylogenetics. In this poster, we describe *Stemmatology*, a new open source package for the statistical software R, that implements procedures for the computer-assisted analysis of textual traditions. We have started implementation of stemmatological methods in the package by focusing, on one hand, on procedures derived from traditional textual criticism, the “Lachmannian” tradition in general, and particularly some of Eric Poole’s methodological insights (Poole [13, 14]); and on the other hand, we made use of methods for the detection of contamination and polygenesis, two major issues for genealogical analysis.

## 1 Introduction

Before the appearance of the printing press, in the West, the only way of reproducing and spreading a text in written form was manual copying. During this process, accidents, errors and intentional modifications occurred, progressively modifying the text of each witness. For the philologist interested in the study of a textual tradition or the restoration of the original text, it has been imperative to study the

---

\*The source code for the package is available on Github: Jean-Baptiste Camps & Florian Cafiero, *stemmatology* : an R stemmatology package, v. 0.2.2, 2014-..., <http://github.com/Jean-Baptiste-Camps/stemmatology>, DOI: 10.5281/zenodo.1117389.

<sup>†</sup>Corresponding author: [jbcamps@hotmail.com](mailto:jbcamps@hotmail.com)

different variants of the witnesses, to assess their genealogical relations, at least since the beginning of the scientific age of philology in the XIX<sup>th</sup> century. As a result, the method of common errors (often deemed “Lachmannian”) took progressively form during the XIX<sup>th</sup> and the beginning of the XX<sup>th</sup> century. Yet, different phenomena such as horizontal transmission (contamination) or the independent appearance of identical variation in different witnesses (polygenesis) cause major difficulties to this method<sup>1</sup>.

Dating back to the experiments of Dom Froger [8], various computational procedures have been used to help assess the genealogy of a textual tradition. These can be roughly divided between methods based on pre-existent textual criticism principles (Lachmannian, Quentinian or other) and methods inspired by fields other than textual criticism, such as phylogenetics or compression-based algorithms [17, 1]. Research in this specialised field is active; methodological contributions, in particular, are numerous, though it is out of our scope here to summarise them (see for instance the recent special issue on stemmatology in [11]).

Amongst these methods, we decided to start by implementing a method based on traditional philological principles, firstly because these are often less available or lack open-source and user-friendly implementations, secondly because, for now, classical methods (including non computerised ones), still seem to offer very satisfactory results when compared to others [17]. We chose to start with the approach designed by Poole [13, 14], extended by Camps & Cafiero [3]. It offered us an algorithmic and easy to compute transposition of the common error method. It also helped us addressing the major problems raised by contamination and polygenesis.

Beyond the methods implemented for now, this software package has been developed with a main objective: valuing the interactions with the researcher, allowing his or her insights to guide and enhance the results.

## 2 Data model

It is not our purpose here to decide which data model would be better to represent textual variations, nor to put constraints on the meaning given to the notions of *witness* or *variant location*. Moreover, the definition of the basic unit of variation, the inclusion or not of some types of variation, are all features that can deeply vary from one context to another, or between projects. A variety of models already exists, from the word-based collation table to the TEI encoded apparatus or the graph model [18], with the addition of local and project-specific models. The data themselves can be stored according to various implementations, including graph databases [1].

To stay as independent as possible from all these choices, we adopted a simple and abstract representation, with very few hard constraints for the user. In our data

---

<sup>1</sup>For a definition of the notions specific to textual criticism, see Duval [6], especially “Contamination”, “Erreur polygénétique” and “Polygène” (p. 88-89, 134-135 and 218). See also Macé, Roelli *et al.* [12], for English language definitions.

model, each column stands for a witness, and each line for a variant location. Each variant is given a numeric code (*NA* for not acquired, 0 for omission,  $1 \dots n$  for variants), as shown in table 1. The exact meaning to give to *variant location* and *variant* is defined by the user, according to his or her approach and the nature of the materials being analysed. The choice to consider omission as readings or not is also left to the user through an option (`omissionsAsReadings`, which can be set to `TRUE` or `FALSE`).

	$W_1$	$W_2$	$W_3$
$VL_1$	1	1	2
$VL_2$	0	1	2
$VL_n$	NA	1	2

Table 1: Tabular data model

To illustrate the flexibility of this approach, let us take as example the case of the potential combination of macro-structural and localised variants: the order of a few paragraphs or verses (or books, etc.) may be different in two groups of witnesses, while the paragraphs themselves also contain *varia lectio*. The user may then choose to:

1. consider only the variations in the order of paragraphs, and encode it as a single variant location, with each observed order taken as a variant and given a numeric code;
2. consider only the variations in content, and create a variant location for each of them, ignoring macro-structural variation;
3. do both, and encode successively a variant location for the change of order, and others for the variations in content.

This flattened approach can also be used for smaller inversions, and there is no limit in terms of depth or imbrication: if there is, in a given tradition, variation in the order of books, in the order of paragraphs inside those books, of sentences inside paragraphs and words inside sentences, along with localised variations in content, a user may very well decide to create separate variant locations to record separately variation of order at each of those levels, in addition to variant locations recording localised variants. Methodological choices in terms of alignment or segmentation are outside the scope of this package.

Since the principal purpose of our software is stemmatological analysis, and not data representation or manipulation, we do not include functions to collate texts or encode variants. Yet, to maximise interoperability, we offer a configurable and easily customised `xslt` stylesheet to transform a TEI-encoded parallel-segmentation apparatus to our data format. We welcome the contribution of other stylesheets to the repository<sup>2</sup>. To illustrate one of the possible transformations,

<sup>2</sup>Available on Github, Jean-Baptiste Camps, *stemmaology-utils*, <https://github.com/Jean-Baptiste-Camps/stemmaology-utils>, DOI: 10.5281/zenodo.1117181.

including cases of inversion, we can take as example this short sample from the beginning of v. 3686 of Chrétien de Troyes' *Chevalier au lion*:

H: Onques ne fu cil  
P: Onques chil ne fu  
V: Onques cil ne fu  
F: Cil ne fu onques  
G: Et cil ne fu pas  
A: Onques cil ne fu  
S: Onques cil ne fu  
R: Onques cil ne fu  
M: Onques cil ne fut

In this sample we find:

- inversions, at various levels: permutation of *Onques* and *cil ne fu*, as well as permutations of *cil* and *ne fu*;
- variants on function-words: *Onques* against *Et ... pas*, that can be decomposed, if need be, in a substitution (*Onques*/*Et*) and an addition/omission (*pas*);
- graphic (diatopic, diachronic) variation (*cil/chil*; *fufut*).

This can be expressed in TEI in the following way (among other possibilities). To account for displacement of one word, we decompose it into two simpler forms of changes, that is one deletion at a first location and one addition in a second location (but we link the two, and still indicate the content subvariants inside):

```
<1 n="3686">
  <!-- First variant, Onques vs. Et -->
  <app xml:id="VL_3686.1" type="functionWord">
    <rdg wit="#H #P #V #A #S #R #M #F">
      <app xml:id="VL_3686.1.1">
        <!-- Subvariant: inversion of Onques -->
        <rdg wit="#H #P #V #A #S #R #M">Onques</rdg>
        <rdg wit="#F" corresp="#inv_F_01"/>
      </app>
    </rdg>
    <rdg wit="#G">Et</rdg>
  </app>
  <!-- Graphical variant chil / cil-->
  <app type="graphic" xml:id="VL_3686.2">
    <rdg wit="#P">chil</rdg>
    <rdg wit="#F #V #G #A #S #R #M #H">
      <!-- H has 'cil' but at a different place, we nonetheless
      indicate that its reading is the same that FVGASRM
      -->
      <app xml:id="VL_3686.2.1">
        <rdg wit="#F #V #G #A #S #R #M">cil</rdg>
        <rdg wit="#H" corresp="#inv_H_01"/>
      </app>
    </rdg>
  </app>
```



```

</app>
ne <app type="graphic" xml:id="VL_3686.3">
  <rdg wit="#H #P #V #F #G #A #S #R">fu</rdg>
  <rdg wit="#M">fut</rdg>
</app>
<!-- And here we account for the inversion -->
<app type="functionWord" xml:id="VL_3686.4">
  <rdg wit="#H" xml:id="inv_H_01">cil</rdg>
  <rdg wit="#P #V #F #G #A #S #R #M"/>
</app>
<app type="functionWord" xml:id="VL_3686.5">
  <rdg wit="#G">pas</rdg>
  <rdg wit="#F" xml:id="inv_F_01">onques</rdg>
  <rdg wit="#H #P #V #A #S #R #M"/>
</app>
</l>

```

Using the xslt stylesheet provided this could then be converted, generating for instance the outcome presented in table 2.

	H	P	V	F	G	A	S	R	M
VL_3686.1	1	1	1	1	2	1	1	1	1
VL_3686.1.1	1	1	1	0	NA	1	1	1	1
VL_3686.2	2	1	2	2	2	2	2	2	2
VL_3686.2.1	0	NA	1	1	1	1	1	1	1
VL_3686.3	1	1	1	1	1	1	1	1	2
VL_3686.4	1	0	0	0	0	0	0	0	0
VL_3686.5	0	0	0	2	1	0	0	0	0

Table 2: Result from the automated conversion into the data format used by the package (all types of variation retained)

In this example, we make the assumption that the user chose to retain all types of variations. By default, only variant locations labelled as *substantive* are retained in the transformation, but it can be configured.

To facilitate quick experimentation, and reproducibility of procedures, we ship the package with various datasets, based on artificial or historical traditions. The datasets are provided in *.rda* format, as part of the package. They consist of tabular data, in the format used for analysis, and represent the selection of significant variant locations as used in [3]. The main datasets are *Fournival* (historical, [15]) and *Parzival* (artificial, [19]). Documentation and reference of datasets are included in the package manual.

### 3 Features of the package

To implement our package, we chose to use the open-source environment for statistical analysis R [16]. R is a well-established software, with a strong community, allowing us to build on a wide range of built-in or community-developed functions. Furthermore, the choice of R makes it very easy to fully document the procedures,

to include both commands and datasets, and to distribute the package. We adopted a copyleft license, the GNU General Public License v.3 [10].

In its current state, the package mainly consists in two sets of features. The first set is dedicated to exploratory analysis of a textual tradition, the second to the building of a stemma.

In the exploratory set of functions, we first implemented a procedure (the function `PCC.conflicts`) to identify contradictions in the variant locations' genealogical configurations, comparing their readings two by two. The underlying intuition is simple: if a variant location is in conflict with a relatively large number of variant locations, it can be considered as unreliable, and ruled out of the further computations. On the other hand, variant locations in conflict with an unreliable variant location can be considered as potentially reliable. The rest of the procedure will help the researcher to assess which variant locations are the source of the contradictions, through visualisation or computations.

To analyse this phenomenon, we represent the set of conflicting variant locations as dots (or “nodes”) on a graph. When there is a conflict between two variant locations, we draw an undirected link (or “edge”) between them. The user can then view the conflicts between variant locations as a network, that can be plotted and mathematically analysed. Current visualisation uses the classical Fruchterman-Rheingold [9] spatialisation.

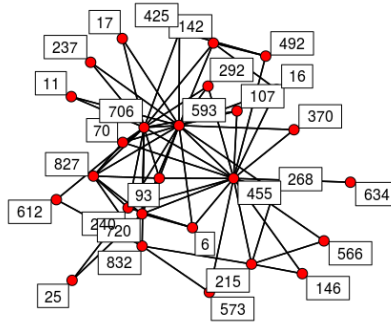


Figure 1: Graph of the conflicts between the variant locations in the *Parzival* dataset

The user is then guided in determining the level of conflictuality that seems acceptable in his corpus. The “degree centrality” [7] of the various nodes is computed and displayed. The nodes are clustered according to the value of their centrality. The higher is the value, the more uncertain is the interest of the associated variant location.

This step is meant to help the user to have an intuition about the “conflictuality” acceptable between variant locations. He is then invited to give the level he deems

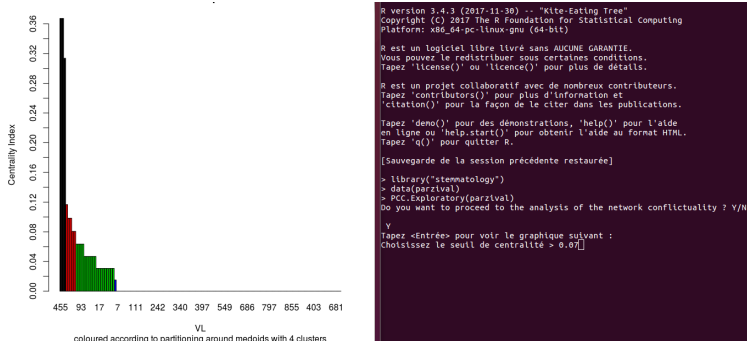


Figure 2: User interface of the package: visualising the centrality of the nodes

bearable (PCC.overconflicting function), and the variants generating too many contradictions according to this setting are displayed.

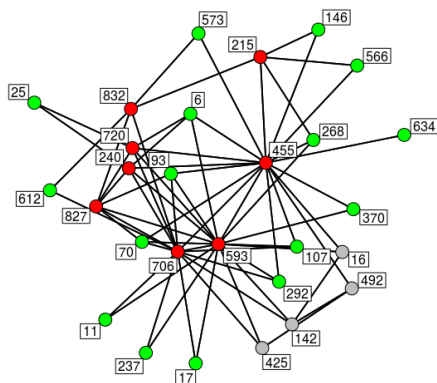


Figure 3: Overconflicting variants isolated - in red - in the *Parzival* dataset

PCC.elimination eventually gets rid of those variant locations, under the user's supervision.

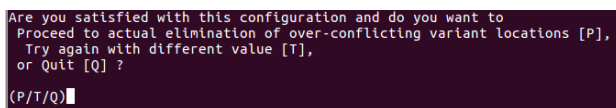


Figure 4: User interface of the package: assisted selection of variants

If contamination is suspected, the function `PCC.contam` can be called. It removes a witness from its calculations, and computes the number of conflicts between variant locations remaining without it. The function repeats the same computation for every witness in the textual tradition. If removing one specific witness induces a significant decrease in conflictuality between variant locations, the function offers to label this manuscript as plausibly contaminated.

At the end of this selection process, some uncertainty can remain about several variant location. Yet, in the event of algorithmically undecidable situations, the user should not be stuck. The function (`PCC.equipollent`) thus allows to create different databases corresponding to the competing configurations. From then on, these databases can be studied separately, and might result in different plausible stemmata.

The second set of functions (called by the general function `PCC.stemma`) allows the user to build one or several stemmata, depending on the input. For the construction of the stemma, only one method is implemented as of now, relying on the transformation of the common error method into a disagreement-based algorithm, formulated by Poole [13, 14]. As such, it is mainly based on disagreements opposing at least two witnesses to at least two others. The algorithm proceeds step by step, first assessing groups (`PCC.buildGroup`), then reconstructing or identifying their model, and then restarting from step one after eliminating the *codices descripti* from the database. It keeps going until there are less than four witnesses in the database.

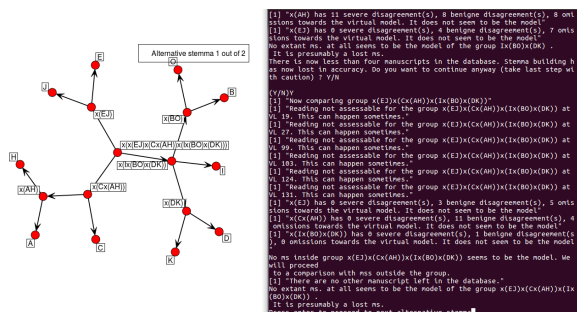


Figure 5: User interface of the package: building the stemma (*Fournival* dataset)

Even though the algorithm can compute a plausible configuration, its default setting incites the expert to make its own decision regarding the very top of the stemma, which is the most delicate to assess [2]. Warned about the loss of reliability of the algorithm's computations, the user can however demand to display its final results.

## 4 Further developments

This first version of the package calls for new developments and improvements. Some minor revisions to improve the different visualisations are undergoing. To highlight the variant locations' different levels of conflictuality, the default mapping could be changed to a circular [5] or radial axis layout, where nodes could be ordered by decreasing centrality. Various mappings could be accessible to the user via a dedicated option. Other visual options could also be implemented for the *stemma codicum*. Obtaining a clearer, more customizable, ready for publication graph, is one of our short-term objectives.

More importantly, we are in the process of implementing other functions. One of them could be dedicated to a better detection of contamination, with procedures such as “cardiograms” [4, 20]. In the future, we also would like to offer an easy access to a large set of existing methods for stemma-building, to facilitate comparisons or benchmarks. The availability of the open source code and its online repository make it easy for others to help us complete our goal and contribute to this software's development.

## References

- [1] Andrews, Tara, Gershoni, Ido, Imhof, Ramona, Kaufmann, Sascha, Schaerer, Jakob, Studer, Thomas, & Zumbunn, Severin, s.d., “Efficient Stemmology: a Graph Database Application in the Digital Humanities”, <http://home.inf.unibe.ch/~tstuder/papers/StemmaRest.pdf>.
- [2] Bédier, Joseph, 1928, “La tradition manuscrite du *Lai de l'ombre* : réflexions sur l'art d'éditer les anciens textes”, *Romania*, 54, p. 161-196 and 321-356.
- [3] Camps, Jean-Baptiste & Cafiero, Florian, 2015, “Genealogical variant locations and simplified stemma : a test case”, in *Analysis of Ancient and Medieval Texts and Manuscripts : Digital Approaches*, ed. Tara Andrews & Caroline Macé, Turnhout, p. 69-93 (Lectio, 1).
- [4] Den Hollander, A.A., 2004, “How shock waves revealed successive contamination : A cardiogram of early sixteenth-century Dutch Bibles”, in *Studies in Stemmology 2*, ed. P. Van Reenen, A.A. Den Hollander & M.J.P. Van Mulken, Amsterdam, p. 99-112.
- [5] Doğrusöz, Uğur, Belviranlı, M., & Dilek, A., 2012, “CiSE : A circular spring embedder layout algorithm”, *IEEE Transactions on Visualization and Computer Graphics*, DOI : 10.1109/TVCG.2012.178 .
- [6] Duval, Frédéric, 2015, *Les mots de l'édition de textes*, Paris (Magister).

- [7] Sabidussi Gert, 1966, “The centrality index of a graph”, *Psychometrika*, 31, p. 581–603.
- [8] Froger, Jacques, 1968, *La critique des textes et son automatisation*, Paris (Initiation aux nouveautés de la science).
- [9] Fruchterman, Thomas M. J., & Reingold, Edward M., 1991, “Graph Drawing by Force-Directed Placement”, *Software – Practice & Experience*, 21-11, p. 1129–1164, DOI : 10.1002/spe.4380211102.
- [10] GNU General Public License, version 3, 2007, Free Software Foundation, <http://www.gnu.org/licenses/gpl.html>
- [11] Heikkilä, Tuomas, & Roos, Teemu, 2016, “Thematic Section on Studia Stemmatalogica”, *Digital Scholarship in the Humanities* 31-3, p. 520-22, DOI : 10.1093/llc/fqw038.
- [12] *Parvum lexicon stemmatologicum - PLS - HIIT Wiki*, ed. Caroline Macé & Philipp Roelli, 2015, <https://wiki.hiit.fi/display/stemmatology/Parvum+lexicon+stemmatologicum>.
- [13] Poole, Eric, 1974, “The Computer in Determining Stemmatic Relationships”, *Computers and the Humanities*, 8-4, p. 207-16.
- [14] Poole, Eric, 1979, “L’analyse stemmatique des textes documentaires”, in *La pratique des ordinateurs dans la critique des textes*, Paris, p. 151-161.
- [15] Richart de Fornival, 1957, *Li Bestiaires d’Amours di maistre Richart de Fornival e li response du bestiaire*, ed. Cesare Segre, Milano & Napoli.
- [16] R Development Core Team, 2014, *R : A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, <http://www.R-project.org>.
- [17] Roos, Teemu, & Heikkilä, Tuomas, 2009, “Evaluating methods for computer-assisted stemmatology using artificial benchmark data sets”, *Literary and Linguistic Computing*, 24-4, p. 417-433.
- [18] Schmidt, Desmond, & Colomb, Robert, 2009, “A data structure for representing multi-version texts online”, *International Journal of Human-Computer Studies*, 67-6, p. 497-514, DOI : 10.1016/j.ijhcs.2009.02.001.
- [19] Spencer, M., Davidson, E. A., Barbrook, A. C., & Howe, C. J., 2004, “Phylogenetics of artificial manuscripts”, *Journal of Theoretical Biology* 227, p. 503–11, DOI : 10.1016/j.jtbi.2003.11.022.
- [20] Wattel, E., & Van Mulken, M.J.P., 1996, “Shock Waves in Text Traditions, Cardiograms of the Medieval Litterature”, in *Studies in Stemmatology*, ed. P. Van Reenen, M.J.P. Van Mulken, Amsterdam, p.105-121.

# Towards a Spatial Annotation Scheme for Basque based on ISO-Space

Ainara Estarrona and Izaskun Aldezabal  
IXA NLP group, University of the Basque Country  
{ainara.estarrona} {izaskun.aldezabal} @ehu.eus

## Abstract

The purpose of this paper is to create a preliminary spatial annotation scheme for Basque based on ISO-Space. To do this, we have first analysed in depth the ISO-Space annotation scheme, and then checked its suitability to apply it to Basque in order to expand the semantic tagging that is being developed in the IXA group. Although the typology of English and Basque are very different, we conclude that the model is also useful for Basque. However, as we went further on the scope of spatial structures, we noticed that this research field *per se* is still in an early stage and sometimes it was difficult to understand some concepts in the annotation scheme. Therefore, we have made our own proposals to tackle some of the problems that we have encountered.

**Key Words:** Semantic annotation, ISO-Space, spatiotemporal annotation, spatial relations, Basque.

## 1 Introduction

The annotation of spatial information is an important challenge for the development of advanced tools and applications such as machine translation, language learning or text summarization, among others. The aim of this paper is to create a preliminary scheme of spatial information for Basque based on ISO-Space. This paper is part of a more general ongoing project the IXA group<sup>1</sup> is pursuing about corpus-tagging frameworks. At the semantic level, the nouns have so far been tagged with Basque WordNet senses [2, 10] and the verbs have been annotated following the PropBank/VerbNet model [4]<sup>2</sup>. Regarding spatiotemporal annotation, a corpus that contains temporal information has been created (*EusTimeBank*) following the EusTimeML mark-up scheme [3]. Our goal now is to lay the foundations to continue this ongoing project with the annotation of spatial information.

The present paper develops as follows: Section 2 presents a brief outline of ISO-Space; in Section 3 we analyse the adequacy of the ISO-Space model to Basque. Section 4 discusses the theoretical problems and practical difficulties that we have found when analysing the ISO-Space scheme. Finally, Section 5 will outline some conclusions.

---

<sup>1</sup> <http://ixa.si.ehu.es>

<sup>2</sup> This semantic tagging makes use of the EPEC corpus (*Euskararen Prozesamendurako Erreferentzia Corpora-Reference Corpus for the Processing of Basque*) [1], which contains 300,000 words of standard written text.

## 2 Brief overview of ISO-Space

With the aim of creating a model for labelling spatial information for Basque, we have first analysed the work carried out in this field for English.

### 2.1 Development of the spatial annotation process within SemEval

The annotation of spatial information is a shared task (SpaceEval) within SemEval since 2012 [6]. We have reviewed the development of the annotation process starting with the SpatialML [8] mark-up scheme, following with the Space Role Labeling [5] and ending with ISO-Space [10], the annotation scheme adopted as a standard since SemEval 2015 [11]<sup>3</sup>.

SpatialML was the basis of the annotation of spatial information. It provide a robust platform for the subtask of geolocating geographic entities and facilities in text, and to do that, it uses basic tags to identify locations and toponyms in the text. However, the complexity of spatial language, motivates a more expressive mark-up scheme.

The Space Role Labeling (SpRL) in SemEval 2012 [6] had a focus on the main roles of trajectories, landmarks, spatial indicators, and the links between these roles which form spatial relations. The formal semantics of the relations are divided into three types: directional, regional (topological), and distal. The annotated corpus, contained mostly static spatial relations. In SemEval 2013, the SpRL task was extended to the recognition of motion indicators and paths, which are applied to the more dynamic spatial relations.

The annotation scheme was extended enriching the semantics in static and dynamic spatial information and including new tags and relations. In this way the ISO-Space annotation scheme was created and it is the standard adopted since SpaceEval 2015.

### 2.2 The ISO-Space mark-up scheme

The descriptive mechanism of ISO-Space consists of a set of six basic entities and a set of four spatial relations over them. The basic entities are: *place*, *path*, *spatial\_entity*, *motion\_event*, *spatial\_signal* and *measure*. All these entities are described by four spatial relations: *qualitative spatial link* (*qslink*), *orientation link* (*olink*), *movement link* (*movelink*) and *measure link* (*mlink*) (see Table 1).

---

<sup>3</sup> The IXA group participated in SemEval-2015 task8 (SpaceEval) for the automatic recognition of spatial information following ISO-Space [13].



Basic entities		Spatial relations
Location tags	Non-location tags	
Place	Spatial_entity	
Path	Spatial_signal	QSLINK
		OLINK
	Motion	MOVELINK
	Measure	MLINK

Table 1: The ISO-Space mark-up scheme.

The *place* tag is used for annotating geographic and administrative entities (lakes, mountains, towns, countries...). The *path* tag is used for locations where the focus is on the potential for traversal or functions as a boundary (road, coast, Pacific Coast Highway...). The *spatial\_entity* is a named entity that is not a location, but one which participates in an ISO-Space link tag. A *motion* tag is an spatial event involving change of location. The *spatial\_signal* tag annotates typically prepositions or other function words that trigger spatial relations between two ISO-Space elements. A *measure* is a tag that captures distances and dimensions and it consists of a numerical component and a unit component or of a relative measurement term (*near*, *close*, *far*...).

The tags for spatial relations capture information about relationships between those tagged elements that we have mentioned in the previous paragraph. There are four link tags:

- a) *qslink*: This tag is used to capture the topological relationship between two spatial objects and it is triggered by *spatial\_signal* tags.
- b) *olink*: This tag covers the relationships that are not topological and its trigger is a *spatial\_signal*.
- c) *movelink*: This tag connects all of the elements that participate in a motion event and it is introduced by a triggering *motion* tag.
- d) *mlink*: This tag can be used to capture the distance between two objects and also to describe the dimensions of a single object and it is commonly accompanied by a *measure* tag (but this is not a requirement).

The annotation scheme also specifies a list of attributes and their values for each of these entities and relations.

### 3 Descriptive adequacy of the ISO-Space annotation scheme to Basque

In order to create an annotation scheme of the spatial information for Basque, we have based on the annotation guidelines for SpaceEval 2015<sup>4</sup> and we have

<sup>4</sup> <http://jamespusto.com/wp-content/uploads/2014/07/SpaceEval-guidelines.pdf>

translated the annotation scheme into Basque to be able to use it in our internal works. However, we have kept the tags in English (see Table 2).

Attribute	Value	Extent
<b>Id</b>	s1, s2, s3...	Postposition
<b>Cluster</b>	Identifies the sense of the postposition <sup>5</sup>	
<b>Semantic_type</b>	DIRECTIONAL (1) TOPOLOGICAL (2) DIR_TOP (3)	
<b>Ex.:</b> - <i>Boston New York iparraldean dago</i> ('Boston is <u>north of</u> New York') - <i>Donostia Gipuzkoan dago</i> ('Donostia is in Gipuzkoa') - <i>Edalontzia mahai(aren) gainean dago</i> ('The glass is <u>on</u> the table')		

Table 2: The attributes and values for the *spatial\_signal* tag.

We have created similar tables for each tag, but in this case we have focused on this one because it gives us the opportunity to talk about one of the points in which English and Basque differ. In English spatial signals are typically prepositions while in Basque, which is an agglutinative language, they are postpositions. Therefore, words like *iparraldean* ('north of') in Basque should be segmented into lemmas and suffixes (postpositions) before annotating them (*iparralde* [lemma] + *an* [inessive suffix]). Although this typological contrast makes the annotation different, this does not represent a problem for the ISO-Space mark-up scheme, as demonstrated by Lee *et al.* [7] for another agglutinative language such as Korean.

Lexical differences may also result in dissimilar annotations as we can see with the motion verbs in Basque and English. In English some motion verbs such as 'bike' or 'walk' contain the information on the manner of moving inside them, but in Basque the manner of moving is usually expressed explicitly. For instance *oinez ibili* ('on foot go' = 'walk') or *bizikletaz ibili* ('by bike go' = 'bike'). This lexical feature can be annotated by ISO-Space without problems, because *motion* tag (see Table 3) has attributes for motion type (path or manner) and motion class (move internally, move externally, leave, reach, follow...). Therefore, ISO-Space mark-up scheme is adequate for the detailed annotation of various features associated with motion verbs [7].

---

<sup>5</sup> Preposition in English.

Attribute	Value	Extent
Id	m1, m2, m3...	Verb
motion_type	MANNER, PATH, COMPOUND, GOAL <sup>6</sup>	
motion_class	MOVE, MOVE_EXTERNAL, MOVE_INTERNAL, LEAVE, REACH, DETACH, HIT, FOLLOW, DEVIATE, CROSS	
motion_sense	LITERAL, FICTIVE, INTRINSIC_CHANGE	
mod	A spatially relevant modifier	
countable	TRUE/FALSE	
Ex.: Jon bizikletaz <i>iritsi</i> zen eskolara (‘Jon arrived at school by bycycle’)		

Table 3: The attributes and values for the *motion* tag

In the same way that in English some verbs of movement contain the way of moving within themselves, in Basque some verbs of movement may contain in themselves the information about the *final\_location* or *goal* of the movement (*etxerat* = 'home-to go'; *zelairatu* = 'the field-into go'). We propose to add a fourth value for the *motion\_type* attribute in order to tag this type of movement verbs in Basque. Therefore, we would have four values in the annotation scheme for Basque: *manner*, *path*, *compound* and *goal* (see Table 3).

## 4 Discussion

In this section we will focus on three topics. On the one hand, we will analyse both the '*non-consuming*' tags and the *motion\_class* attribute of the *motion* tag together, because we think that they are closely related; and on the other hand, we will talk about the *spatial\_signal* tag and the links that it triggers.

*Non-consuming* tags were created to capture spatially relevant locations or entities that are not directly referenced in the text. The extent of *non-consuming* tags is a null or empty string. In SpaceEval Annotation Guidelines<sup>7</sup> it is said that, generally, non-consuming tags are not necessary to capture relevant spatial objects and relations and that for this reason they should be used sparingly. In fact, they mention three situation where the use of these tags is necessary: i) locations referenced by a measure; ii) locations implied by 'cross' and 'across'; and iii) sets whose members are mentioned<sup>8</sup>.

<sup>6</sup> The *goal* value does not appear in ISO-Space. It is a value that we have added to tag a particular type of verbs that exist in Basque, the verbs that contain in themselves the information about the final location or goal of the movement.

<sup>7</sup> <http://jamespusto.com/wp-content/uploads/2014/07/SpaceEval-guidelines.pdf>

<sup>8</sup> This third situation is not mentioned in Pustejovsky [12].

The reason to use these tags in such situations and not in others is that these tags are necessary to fill the value of certain attributes in other tags or links. We may agree with this assumption, but we do not understand why in other situations, for example, in the case of *movelink* tags the *non-consuming* tags are not necessary. In the *motion* tag there is an attribute that is *motion\_class* (see Table 3 in section 3). The guidelines [12] specifies which attributes are required in the *movelink* tag for each *motion\_class* (see Table 4). We assume that if these attributes are required, they should be specified using a *non-consuming* tag, because they are necessary to fill the *source*, *goal*, *midpoint* or *ground* attributes of the *movelink* tags (see Table 5). However, the guidelines do not specify anything about it.

<i>motion_class</i> of trigger	Required Attributes
move	-
move_external	landmark <sup>9</sup>
move_internal	landmark
leave	source
reach	goal
detach	source
hit	goal
follow	pathID <sup>10</sup>
deviate	source
cross	source, midPoint, goal

Table 4: Required attributes in the *movelink* tag for each *motion\_class*.

Attribute	Value	Trigger
<b>Id</b>	mv11, mv12, mv13...	Movement verb
<b>trigger</b>	ID of a MOTION that triggered the link	
<b>source</b>	ID of a location/entity/event tag at the beginning of the event-path	
<b>goal</b>	ID of a location/entity/event tag at the end of the event-path	
<b>midPoint</b>	ID(s) of event-path	

<sup>9</sup> *Ground* in Pustejovsky [12].

<sup>10</sup> *Goal* in Pustejovsky [12].

	midpoint location/entity/event tags	
<b>mover</b>	ID of the location/entity/event tag whose location changes	
<b>ground</b>	ID of a location/entity/event tag that the mover participant's motion is relative to	
<b>goal_reached</b>	TRUE, FALSE, UNCERTAIN	
<b>pathID</b>	ID of a PATH tag that is identical to the event-path of the trigger MOTION	
<b>motion_signalID<sup>11</sup></b>	ID(s) of (an) MOTION_SIGNAL tag(s) that contributes path or manner information to the trigger MOTION	
Ex.: [ <i>Jonek<sub>se2</sub></i> <sup>12</sup> ] [ <i>autoz<sub>ms3</sub></i> ] [ <i>bidaiatu<sub>m1</sub></i> ] <i>zuen</i> ('Jon <sub>se2</sub> traveled <sub>m1</sub> by_car <sub>ms3</sub> ') trigger=m1; mover=se2; motion_signalID=ms3		

Table 5: The attributes and values for the *movelink* tag.

In order to help the annotators know when to use these *non-consuming* tags and when not, we think it is necessary to establish precise criteria. Hence we have made a proposal based on the verbal subcategorization of Basque that we have collected in our verbs lexicon *Basque Verb Index* (BVI)<sup>13</sup>. We propose to annotate with *non-consuming* tags all the subcategorised elements or arguments of a given movement verb that are not explicitly referenced, but which can be retrieved from the text using coreference<sup>14</sup> (1).

- (1) *Mikel Indiara joan zen oporretan.*  
Mikel to-India go.partc was on-holiday  
'Mikel went on holiday to India'

[Indiara] *iritsi bezain\_laster damutu zen.*  
[To-India] arrive.partc as soon as regret.partc was  
'As soon as he arrived [to India] he regretted it'

<sup>11</sup> *AdjunctID* in Pustejovsky [12].

<sup>12</sup> *Se* = spatial\_entity, *ms* = motion\_signal and *m* = motion.

<sup>13</sup> <http://ixa2.si.ehu.es/e-rolda/index.php>

<sup>14</sup> In some cases, we will need a wider context than the phrase in order to identify these elided elements.

In (1) we would create a non-consuming tag for the final location argument of the verb *iritsi* ('to arrive'), because we can retrieve it from the previous sentence. Nevertheless in (2) we would not create a non-consuming tag, because the final location of the verb *joan* ('to go', 'to leave') is underspecified and can not be retrieved from the text.

- (2) *Halaxe joan zen mundu honetatik Joanes, Bargotako*  
 Like this go.partc was world this-of Joanes, Bargota-of  
*Brujoa.*  
 Sorcerer-the  
 'This is how Joanes, the sorcerer of Bargota, left this world'

Following with the verbs of movement, the classification of the movement classes and their event-path structures presented in the guidelines is very interesting for crosslingual studies. As mentioned above, each *motion\_class* has an argument which is focused depending on the structure of the event-path. For example, while the focus for the 'leave' class is the 'source', the focus for the 'reach' class is the 'goal'. In our case, we have not yet done this analysis for Basque, and therefore, we find it difficult to classify the Basque verbs based on such criteria. For instance, the verb *joan* ('to go') can be both a 'leave' class and a 'reach' class verb depending on the underspecified locative arguments, that is, when the underspecified locative argument is the *goal*, it will be a 'leave' *motion\_class* verb. In fact, in this case the verb *joan* would be translated as 'to leave', and not as 'to go', in English. However, we think that in order to carry out a motion-class study for Basque it is necessary to annotate previously a sample of corpora following the proposed criteria, since both the omission and underspecification phenomena can only be identified at surface syntax level.

Finally, the *spatial\_signal* tag has caused us problems, since it has been sometimes difficult to differentiate between the three semantic types that are mentioned in the guidelines (*topological*, *directional* and *dir\_top*) and the relations that they introduce (*qslink* and *olink*). In Pustejovsky [12] when measuring the inter-annotator agreement for each tag, the tags *qslink* and *olink* are the ones that get the lowest results. This suggests to us that perhaps the distinction between them is not clear enough and that depending on the particular task or application in which the annotation will be applied it is likely that this distinction will not be necessary, or at least not in that degree of detail.

## 5 Conclusions

In this paper we have presented a brief overview of the ISO-Space annotation scheme and we have analysed its adequacy for labelling spatial structures in Basque.

The main conclusion we have drawn is that the ISO-Space mark-up scheme is suitable for tagging spatial information in Basque. However, as

Basque is an agglutinative language, the identification of markables needs sometimes to resort to smaller segments than word forms. In addition, we have enriched the annotation scheme with a new type of movement for the *motion* tag, the motion type *goal*, necessary to annotate this type of verb that exists in Basque.

In this paper we have presented the first attempt to adapt the ISO-Space mark-up scheme to Basque, and therefore, although this model can be taken as a general theoretical framework, in the future and with concrete applications in mind, it will be necessary to delimit the annotation scheme depending on those specific applications.

## Acknowledgements

This research has been supported by the University of the Basque Country (IXA group, (GIU16/16)) and the Ministry of Economy, Industry and Competitiveness of the Spanish Government (TUNER: TIN 2015-65308-C5-1-R; MUSTER: PCIN-2015-226).

## References

- [1] Aduriz, Itziar, Aranzabe, María Jesús, Arriola, Jose. María, Atutxa, Aitziber, Díaz de Ilarraza, Arantza, Ezeiza, Nerea, Gojenola, Koldo, Oronoz, Maite, Soroa, Aitor and Urizar, Ruben (2006) Methodology and steps towards the construction of EPEC, a corpus of written Basque tagged at morphological and syntactic levels for automatic processing. In Andrew Wilson, Paul Rayson and Dawn Archer (eds.), *Corpus Linguistics Around the World*. Book series: Language and Computers. Vol. 56, 1-15. Rodopi (Netherlands).
- [2] Agirre, Eneko, Aldezabal, Izaskun, Etxeberria, Jone, Izagirre, Izaskun, Mendizabal, Karmele, Pociello Eli and Quintian, Mikel (2006) A methodology for the joint development of the Basque WordNet and Semcor. In *Proceedings of the 5th International Conference on Language Resources and Evaluations (LREC)*. Genoa, Italy.
- [3] Altuna, Begoña, Aranzabe, María Jesús and Díaz de Ilarraza, Arantza (2017) EusHeidelTime: Time Expression Extraction and Normalisation for Basque. *Procesamiento del Lenguaje Natural*, n.º 59, 15-22.
- [4] Estarrona, Ainara, Aldezabal, Izaskun, Díaz de Ilarraza, Arantza and Aranzabe, María Jesús (2016) Methodology for the semiautomatic annotation of EPEC-RolSem, a Basque corpus labelled at predicate level following the PropBank/VerbNet model. Edward Vanhoutte (ed.) *Digital Scholarship in the Humanities* (2016) 31 (3): 470-492. DOI: <http://dx.doi.org/10.1093/llc/fqv010>. First published online: 17 June 2015 (23 pages). Published by Oxford University Press on behalf of EADH: The European Association for Digital Humanities.
- [5] Kordjamshidi, P., VanOtterlo, M. and Moens, M.F. (2010) SpatialRoleLabeling: Task Definition and Annotation Scheme. *Proceedings of the Seventh conference on International Language and Resources and Evaluation (LREC'10)*. 413-420. European Language Resources and Evaluation (ELRA).

- [6] Kordjamshidi, P., Bethard, S. and Moens, M.F. (2012) SemEval-2012 Task3: SpatialRoleLabeling. In *Proceedings of the 6th International Sorkshop on Semantic Evaluation (SemEval)*, 365-373.
- [7] Lee, Kiyong, Fang, Alex C. and Pustejovsky, James (2011) Multilingual Verification of the Annotation Scheme ISO-Space. *Fifth IEEE International Conference on Semantic Computing*, 449-458.
- [8] Mani, I., Doran, C., Harris, D., Hitzeman, J., Quimby, R., Richer, J., Wellner, B., Mardis, S. and Clancy, S. (2010) SpatialML: annotation scheme, resources, and evaluation. *Language Resources and Evaluation*, Volume 44 (3), 263-280. Springer.
- [9] Pociello, Eli, Agirre, Eneko and Aldezabal, Izaskun (2010) Methodology and Construction of the Basque WordNet. *Language Resources and Evaluation Journal*, 45:2, 121-142. Springer.
- [10] Pustejovsky, James, Moszkowicz, Jessica and Verhagen, M. (2012) A Linguistically Grounded Annotation Language for Spatial Information. *Traitement Automatique des Langues (TAL)*, Volume 53 – n° 2/2012, 87-113.
- [11] Pustejovsky, James, Kordjamshidi, P., Moens, M.F., Levine, A. Dworman, S. and Yocum, Z. (2015) Semeval-2015 task 8: Spaceeval. *Proceedings of the 9<sup>th</sup> International Workshop on Semantic Evaluation*. 884-894.
- [12] Pustejovsky, James (2017) ISO-Space Annotating Static and Dynamic Spatial Information. Nancy, Ide and James, Pustejovsky (eds.) *Handbook of Linguistic Annotation*, 989-1024. Springer Netherlands.
- [13] Salaberri, Haritz., Arregi, O. and Zapiain, Beñat (2015) IXAGroupEHUSpaceEval: (*X-Space*) A WordNet-based approach towards the Automatic Recognition of Spatial Information following the ISO-Space Annotation Scheme. *Proceedings of the 9<sup>th</sup> International Workshop on Semantic Evaluation (SemEvan2015)*, 856-861. Denver, Colorado, June 4-5, 2015. Association for Computational Linguistics.



# LitText: Realizing the "All Methods Applied to All Texts" Motto: Exploring a Corpus of Literary Texts With SPARQL

Andrew U. Frank and Christine Ivanovic

Geoinformation TU Wien, Comparative Literature University Vienna  
E-mail: frank@geoinfo.tuwien.ac.at, christine.ivanovic@univie.ac.at

## Abstract

Text-based digital humanities research should follow a pattern of (1) building a corpus of texts; and (2) analyzing the texts systematically. The design of LitText supports such a research approach and empowers digital humanities (DH) researchers to build their own corpus with minimal technical involvement. Text is prepared with minimal markup, is automatically processed by NLP services, and is converted to RDF triples following the Semantic Web recommendation for linked data. The DH researcher can then query the corpus with SPARQL queries. The design of LitText is different from other corpus workbenches, as it integrates widely used, standard NLP tools and triplestores, and the SPARQL query language. The approach is demonstrated with the construction of a 13 million word corpus and a query to identify texts where animals behave like humans (e.g. fables).

Keywords: linked data, Semantic Web, RDF, SPARQL

## 1 Introduction

Corpus-based research poses a challenge for comparative literature, and indeed for most disciplines within text-based humanities [5]. In order to bridge the gap between what could be achieved with computational linguistics (CL) and what is currently done in digital humanities (DH), one must step back from technical issues, and focus on the research practices of literary studies scholars. In general, research in text-based DH starts with (1) selecting texts which are to be analyzed; and then progresses to (2) reading the texts and identifying parts with properties relevant to the research goal.

LitText demonstrates that the traditional approach can guide the design of a corpus workbench. It shows how the currently available tools of computational linguistics (CL) can be combined to advance text-based humanities in general. Hardware and CL tools are powerful enough to apply thorough CL analysis to all texts

in a corpus automatically and without much involvement of the DH researcher. The results are stored in a triplestore for analysis with the SPARQL query language, and can be combined with other available resources as linked (Semantic Web) data.

The conceptual model to build a specialized corpus and exploit it with a query facility that can retrieve pieces of text and then manipulate them in a spreadsheet is conceptually close to current DH research methods in text-based humanities. It facilitates the application of the "all methods for all texts" maxim, and advances "good research practice" by (1) separating the selection of the texts which are the object of the study and which form the corpus from (2) the analysis. Performance is sufficient even on modest and outdated hardware with all available CL tools applied. We found that a 13 million word corpus (equivalent to more than 120 novels) can be re-processed in a single day on an ordinary current PC (see section 6 for hardware used). A query against such a corpus takes a mere 1 to 2 minutes. Re-processing the corpus after changes are made is therefore very feasible.

*LitText* makes it possible to build substantial corpora with texts in multiple languages, specific to focused research. Empowering text-based DH researchers to build their own corpora processed on their own hardware under their own control can also reduce fears of copyright infringement (for a technical solution, see: [10]).

*LitText* is a workbench for the construction of corpora (similar to IMS Open Corpus Workbench [3]) rather than an integration of CL and other tools like DARIAH or LAPPSgrid [4]. *LitText* is not a collection of tools, but a method to build and analyze a literary corpus with minimal CL knowledge required.

There are three novel aspects in *LitText*:

1. Processing texts in multiple languages, hiding the technical complexity from the DH researcher.
2. Storing the result of the NLP analysis as linked data.
3. Using the SPARQL query language to search for relevant text parts.

Section 2 compares the approach of *LitText* with other, similar efforts. Section 3 covers the corpus building, while Section 4 shows analysis with an example query. Section 5 presents the workflow, and Section 6 reports some performance figures. Section 7 mentions some future extension, leading to Section 8 for the conclusion.

## **2 Design of *LitText* in comparison to other projects**

Extant research literature discusses three different approaches to make CL tools more useful for DH: (1) workbenches with tight integration of CL tools with storage and query facilities for a corpus; (2) environments in which CL tools and storage facilities are made available to DH researchers; and (3) software which is highly integrated, for very specific research questions.

`LitText` is similar to IMS Open Corpus Workbench CWB<sup>1</sup> [3] and similar efforts to integrate CL tools and corpus management. It combines tools to process text, and builds from the processed, i.e. annotated text a corpus for querying and can serve as a foundation for other more specialised tools. Unlike IMS CWB, `LitText` (1) allows source text to include metadata and texts not to be processed, simplifying the management of sources; (2) stores the annotations in RDF triples which are stored in a triple store; (3) uses the SPARQL query language (rather than the proprietary CQP language).

Two major projects to bridge the gap between computational linguistics and digital humanities are reported in pertinent literature:

- DARIAH<sup>2</sup> is an European infrastructure incorporated in 2014, connecting multiple nationally funded projects to collect tools useful for a wide range of humanities fields to allow them to analyze natural-language text. It uses standard tools to process text, and allows for the storage of input text and results. The goal is to overcome non-trivial technical installation difficulties, and assure the availability of NLP tools — a first step for researchers from the humanities to apply these tools to their specific problems.
- LAPPSgrid<sup>3</sup> [4] has similar goals. It adds considerations for copyright restrictions attached to the texts to permit the holding of such texts, their processing, and making the results available in accordance with those restrictions (e.g. HathiTrust) [10].

These projects collect CL tools and make them more usable for DH researchers. `LitText` focuses on the DH researcher and structures his work into two familiar tasks: (1) collecting the corpus, and (2) searching the corpus. CL annotations and the management in a triplestore are automated and entirely transparent to the DH researcher. Rather than forcing the research organization to conform to a given technical solutions, the technical solution is instead matched to the research approach.

Lord et al.[6] followed a different approach: the researcher supplies some example texts she is interested in, and the search engine then finds all text similar to the given one. `LitText` instead forces the researcher to explicitly formulate the properties of the text pieces she is interested in through a query; `LitText` does not rely on an algorithm to pick up intended similarities between examples.

### 3 Corpus Building

The goal of `LitText` is to facilitate the collection of texts and automate the formation of a corpus from the texts relevant to a particular investigation. In order to

---

<sup>1</sup><http://cwb.sourceforge.net/>

<sup>2</sup><http://www.dariah.eu/>

<sup>3</sup><http://www.lappsgrid.org/>

collect many texts into a corpus, the effort to prepare the texts must be minimal; but it must also be flexible, in order to accommodate different text sources. Texts are stored in files, and may contain metadata and other textual material which is not part of the actual analysis; texts can be in different languages, and languages can change within a text.

### 3.1 Preparation and Preprocessing

Preparing a text means finding a source, assuring proper UTF-8 encoding, adding some markup for metadata (title, author, year of publication), and separating the actual text from other textual material included in the file (e.g. table of contents, preface, original file name, etc.)

Much of the material we used comes from Project Gutenberg<sup>4</sup>, and we prepared a small tool that, given a Gutenberg book number, downloads the respective texts and augments them with most of the necessary markup automatically; downloading and marking up a literary text from Project Gutenberg takes less than 2 minutes of person-time for editing the text file (on a current laptop or office PC).

Metadata is structured along the lines of the Dublin Core standard, but usually only partially available; the level of detail included can be pushed as far as is suitable for the respective project.

### 3.2 Processing with CL software

A text with markup is processed by NLP programs, and the result translated into the N-Triples format. NLP tools are connected as services, and the conversion of text to input format and the automated conversion of a text file to subject-predicate-object triples takes place in three phases:

1. The metadata, the layout, and the structure of the text intended by the author are converted (if desired, the full text can be captured and translated into the N-Triples format for later use in query formulation).
2. Using the markup, the text is split into parts of a single language, and sent to language-specific NLP services. Differences in language-specific NLP processing are handled internally.
3. The result of the NLP process, for example in the XML format produced by the Stanford coreNLP tools, is translated to triple structure, preserving the produced Treebank codes (currently the UD codes produced by the coreNLP package)[8].

NLP tools are connected as services; if necessary, a tool must be wrapped into a service wrapper (as was necessary for the lemmatization for German language texts,

---

<sup>4</sup><http://www.gutenberg.org/>

using TreeTagger [12]). `LitText` extracts the appropriate text parts and converts to the format expected by the tool, then analyses the results.

The differences between the dependency codes of different languages - even when using Universal Dependencies - are currently preserved, and must be dealt with in a query. We hope that language-independent annotations schemes, especially AMR<sup>5</sup>[1] will lead to a higher-level, more semantic annotation and query facility.

### 3.3 Building the corpus as a triplestore

The N-Triples are stored in one of the readily available triplestores. We constructed utilities for loading the N-Triple files into the store which relies on standardized HTTP protocol SPARQL update commands; they should work with any compliant triplestore.

### 3.4 Additional data sources

Literary analysis may require different types of additional input, e.g. Wikipedia<sup>6</sup> or moviedb<sup>7</sup>. Data in linked-data format, typically available through a public SPARQL endpoint, can be connected to the corpus in queries.

In the following example query (see section 4), we use hypernym relations imported from WordNet; which can be downloaded in triple format [7] based on the lemon schema, and loaded in a triplestore. The resources, especially the hypernym relations, can be accessed in SPARQL queries.

The relations of importance for the example query are extracted from the WordNet data and stored as new relations; stored relations reduce the complexity of later query formulation and speeds up processing. WordNet-like datasets are available for many languages<sup>8</sup>, but their coverage<sup>9</sup> and the data model may vary. Extracting the data required with a SPARQL construct query is helpful to bridge differences between different WordNet data models.

## 4 Analysis

A triplestore offers a HTTP protocol query facility, the so-called SPARQL endpoint, where queries are submitted and answers are returned. SPARQL is a powerful, general-purpose triplestore query language. It is patterned after SQL, but logically simpler because storage is always in binary relations.

---

<sup>5</sup><https://github.com/amrisi/amr-guidelines/blob/master/amr.md>

<sup>6</sup><http://data.linkedmdb.org/sparql>

<sup>7</sup><http://data.linkedmdb.org/sparql>

<sup>8</sup><http://globalwordnet.org/>

<sup>9</sup>[http://compling.hss.ntu.edu.sg/omw/\[2\]](http://compling.hss.ntu.edu.sg/omw/[2])

Statistical queries to determine word frequency, size of vocabulary, etc. are easier to perform in the same framework than more complex queries justified by literary study hypothesis.

We show here an example query to identify “animal fables”, considered here as texts in which animals are reported to behave like humans, i.e. they think and use intelligent communication. Examples are e.g. the classic fables ascribed to Aesop.

A systematic search reveals that modern literary texts include similar literary tropes. Such an analysis requires more language analysis than can be provided by simple word frequency analysis etc. The proposed algorithm to identify such texts finds sentences where the subject is an animal, and the verb expresses either “think” or “communicate”. The approach is to use the dependency tree to identify the subject and the verb, lemmatize those, and then test whether the subject noun has “animal” among its hypernyms, as well as the verb has “think” or “communicate” among its hypernyms.

```

prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX wn: <http://wordnet-rdf.princeton.edu/ontology#>
PREFIX nlp: <http://gerastree.at/nlp_2015#>
prefix lemon: <http://lemon-model.net/lemon#>
prefix lit: <http://gerastree.at/lit_2014#>

SELECT ?author ?werk ?sentform ?subjlemma ?verblemma
FROM <http://gerastree.at/c> #the corpus of literary text
FROM <http://gerastree.at/a1> #the graph with the hypernyms
WHERE {
  ?dep nlp:dependency "NSUBJ" .
  ?dep nlp:dependent ?subj .
  ?dep nlp:governor ?verb .

  ?subj nlp:lemma3 ?subjlemma .
  ?verb nlp:lemma3 ?verblemma .

  ?verb nlp:pos ?catv .
  filter (strstarts (?catv, "V")).

  ?subj nlp:pos ?cats .
  filter (strstarts (?cats, "N")).

  ?lexs lemon:writtenRep ?subjlemma .
  ?lexs lit:nounClass wn:Animal .
  ?lexv lemon:writtenRep ?verblemma .
  {?lexv lit:nounClass wn:Communicate}
  union {?lexv lit:nounClass wn:Cerebrate} .

  ?subj rdfs:partOf ?sent .
  ?sent nlp:sentenceForm ?sentform .

  ?sent lit:inWerk ?werk .
  ?werk lit:author ?author .
} order by ?werk

```

The actual text is stored in graph `c` and the hypernym relations extracted from the WordNet in graph `a1`; the query progresses top-down: first it fills the variable `?dep` with a node in a dependency tree which is a subject-noun phrase, and takes the subject and the verb part (`?subj` and `?verb`). Next, check that `?verb` is indeed a verb

and ?subj a noun, and retrieve (for subject and verb) the corresponding lemmata as ?subjlemma and ?verblemma. The next steps are tests whether the subject is an animal, and whether the verb indicates either communicating or thinking, using the previously stored wordnet hypernym relation. The remainder of the query is tracing from the sentence to the text (*werk*) and the author.

The output of the result is stated in the `SELECT` clause, and gives the qualifying sentences as well as the author and title of the text containing that sentence.

## 4.1 Quality of result

The statistical query uses a definition of "word" which is somewhat unusual for literary studies; it is defined by the NLP process and thus uniformly applicable across languages. It counts interpunction marks as words. Processing for the semantic query is simplistic, as it omits constructions where the subject is an animal but not a noun - using coreference information would probably reduce such errors. Commission errors lead to counting sentences in which a dog barks, a cat meows etc.; and constructions where a polysemous noun is not used to describe an animal even though its first sense is a hypernym of "animal". Nevertheless, texts which should indeed be considered animal fables (in the narrow definition) are characterized by frequent occurrences of such constructions, while other texts tend to show only a few stray occurrences.

## 5 Workflow

The programs are designed such that a DH researcher needs only insert a marked-up text file into a directory. A `LitText` process to take text files and convert them to the N-triple format, and a second process to take N-triple file and insert them into a corpus scan regularly for new files and process them. These processes can run on a server where the corpus is incrementally built while marked-up text files are added to a directory.

## 6 Performance

Testing was done on a PC with an Intel i5 processor clocked at 3.2 GHz, with 24 GB of memory<sup>10</sup>. The example corpus we built consists of more than 120 literary texts (mostly books) for a total of 12.7 million words.

- Download and markup of a text from Project Gutenberg takes less than 2 minutes person-time.

---

<sup>10</sup>The memory is mostly necessary for NLP processing of highly complex literary texts like "Ulysses" by J. Joyce; ordinary grammatical text is processed using less than 8 GB

- The preprocessing of the full corpus takes a total of 12 hours. Processing a literary text runs at about 270 words per second and produces about 17 triples per word, when the full text and all NLP results are stored.
- Storing the triples processes an average 7,000 triples per second, for a total of 8 hours 35 min for the 215 million triples.

Processing the semantic query takes 111 seconds (same corpus and hardware as before).

The code can be downloaded from a github account<sup>11</sup> and combined with a suitable triple store<sup>12</sup>. The SPARQL endpoint is accessible on the web<sup>13</sup>.

## 7 Future work

A number of improvements are envisioned:

- Adding support for more languages. At the moment, *LitText* is prepared to handle English and German text, using Stanford coreNLP. Two server processes are set up at two different ports; in addition, German text is lemmatized using TreeTagger as a service at a third port<sup>14</sup>. Adding other languages is an obvious extension (e.g. French, Japanese).
- Using coreference data produced by NLP programs.
- Using Abstract Meaning Representation Language (AMR) [1] to simplify queries across texts in multiple languages.
- Building multilingual corpora in order to facilitate computational comparative literature studies
- Construct a benchmark comparable to the one reported by Proisl and Uhrig[9] to assess performance issues.

## 8 Conclusions

*LitText* is a suite of programs for researchers in DH to build multi-language corpora for projects as linked data, and to explore them using the standardized SPARQL query language. Other data available as linked data, e.g. Wikipedia or WordNet, can be integrated seamlessly. The design goal was to minimize the amount of technical detail a DH researcher would need to use the system, and to follow an approach already familiar to a DH researcher:

---

<sup>11</sup>[andrewufrank.github.com/LitText](http://andrewufrank.github.com/LitText)

<sup>12</sup>[jena.apache.org/documentation/serving\\_data](http://jena.apache.org/documentation/serving_data)

<sup>13</sup><http://nlp.gerastree.at:3030/>; for details contact first author.

<sup>14</sup><http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/>



- build the corpus to be studied
- find pieces of text of interest to the research

Texts require minimal preparation with a simple markup language to add metadata about the text and identify the sections of text to be studied; multiple languages in a single file are possible, and the NLP processing is transparent to the user. The results of the NLP processing are stored in triples in a triple store, and can be queried with the standardized query language SPARQL. A DH researcher needs to learn a minimum set of NLP codes, e.g. a few Treebank codes. Further processing of results is possible with spreadsheet software. Texts can be added at any time by inserting files with markup into a directory. When queries are repeated, they search the increased corpus. The motto "all methods applied to all texts" [5] is respected: the full corpus can be reprocessed in one day and the reevaluation of queries takes only minutes.

Technically, *LitText* revolves around a linked-data triple store and the corresponding SPARQL update and query language. The main server takes a text file with markup, sends the pertinent text to the required, language-specific NLP processes, and converts the result to a triple format. Utilities to download text from research-specific sources and to automate markup as far as automatically possible are small extensions. Storage and query can be done with any SPARQL 1.1 conformant program; loading triples into triple-stores and accessing the SPARQL endpoint uses the HTTP protocol and can be done in a web browser. Additional resources which are available as linked data can easily be incorporated (e.g. WordNet, Wikipedia). Methods to safeguard copyright protection for texts are not yet included and will be added when necessary. The use of standards makes a plethora of additional data in linked data format and additional software available and reduces the number of utilities which must be specially programmed.

## 9 Acknowledgment

We benefited from discussion with Hanno Biber from the OeAW. Christine Ivanovic was supported by the Max Kade Foundation while at Brown University in Providence, RI. Comments from reviewers were helpful to improve the final version of the contribution.

## References

- [1] L. Banarescu, C. Bonial, M S. Cai, Georgescu, K. Griffitt, U. Hermjakob, K. Knight, P. Koehn, M. Palmer, and N. Schneider. Abstract meaning representation for sembanking. 2013.
- [2] Francis Bond and Kyonghee Paik. A survey of wordnets and their licenses. *Small*, 8(4):5, 2012.

- [3] Stefan Evert and Andrew Hardie. Twenty-first century corpus workbench: Updating a query architecture for the new millennium. 2011.
- [4] Nancy Ide, James Pustejovsky, Christopher Cieri, Eric Nyberg, Denise DiPersio, Chunqi Shi, Keith Suderman, Marc Verhagen, Di Wang, and Jonathan Wright. The language application grid. In *International Workshop on World-wide Language Service Infrastructure*, pages 51–70. Springer, 2015.
- [5] Christine Ivanovic and Andrew U Frank. Corpus-based research in Computational Comparative Literature. In Francesco Mambrini, Marco Passarotti, and Caroline Sporleder, editors, *Proceedings of the Workshop on Corpus-Based Research in the Humanities (CRH)*. Warsaw, Poland, pages 69–78, 2015.
- [6] Greg Lord, Martha Nell Smith, Matthew G Kirschenbaum, Tanya Clement, Loretta Auvil, James Rose, Bei Yu, and Catherine Plaisant. Exploring erotics in Emily Dickinson’s correspondence with text mining and visual interfaces. In *Digital Libraries, 2006. JCDL’06. Proceedings of the 6th ACM/IEEE-CS Joint Conference on*, pages 141–150. IEEE, 2006.
- [7] John McCrae, Christiane Fellbaum, and Philipp Cimiano. Publishing and linking WordNet using lemon and RDF. In *Proceedings of the 3rd Workshop on Linked Data in Linguistics*, 2014.
- [8] Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan T McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. Universal dependencies v1: A multilingual treebank collection. In *LREC*, 2016.
- [9] Thomas Proisl and Peter Uhrig. Efficient dependency graph matching with the IMS open corpus workbench. In *LREC*, pages 2750–2756, 2012.
- [10] James Pustejovsky, Marc Verhagen, Keongmin Rim, Yu Ma, Liang Ran, Samitha Liyanage, Jaimie Murdock, Robert H McDonald, and Beth Plale. Enhancing access to digital media: The language application grid in the HTRC data capsule. In *Proceedings of the Practice and Experience in Advanced Research Computing 2017 on Sustainability, Success and Impact*, page 60. ACM, 2017.
- [11] Roland Schäfer. Processing and querying large web corpora with the cow14 architecture. In *Proceedings of the 3rd Workshop on Challenges in the Management of Large Corpora*, pages 28–34, 2015.
- [12] Helmut Schmid. Probabilistic part-of speech tagging using decision trees. In *New methods in language processing*, page 154. Routledge, 2013.
- [13] Jörg Tiedemann. Parallel data, tools and interfaces in opus. In *LREC*, volume 2012, pages 2214–2218, 2012.

# Incorporating Hittite into PROIEL: a pilot project

Guglielmo Inglese<sup>1</sup>, Maria Molina<sup>2</sup> and Hanne Eckhoff<sup>3</sup>

<sup>1</sup>University of Pavia/University of Bergamo

<sup>2</sup>Institute of Linguistics, Russian Academy of Science

<sup>3</sup>Dep. of Modern Languages, University of Oxford

E-mail: guglielmo.inglese01@ateneopv.it;  
maria.lakhuti@gmail.com; hanne.eckhoff@mod-  
langs.ox.ac.uk

## Abstract

In this paper, we report the results of a pilot project aimed at the inclusion of Hittite texts in the PROIEL family of treebanks. The first challenge is that the PROIEL annotation scheme has been designed for Indo-European languages mostly written in alphabetic scripts, so that a way to annotate the complex cuneiform script on which Hittite tablets are recorded must be worked out. Moreover, Hittite also provides some interesting morphosyntactic features that require the adaptation of annotation strategies already in use for other languages in PROIEL. Overall, our preliminary findings show that Hittite can be easily integrated into the PROIEL enterprise, but also that future work is required to effectively achieve this goal.

## 1 Introduction

The *Pragmatic Resources in Old Indo-European Languages* (PROIEL) project set out in 2008 with the aim of investigating information packaging and related phenomena, e.g. word order and discourse particles, in ancient Indo-European (IE) languages (Haug *et al.* [6]; Eckhoff *et al.* [2]). The core of the project consisted in the creation of annotated linguistic resources, i.e. treebanks, for the languages under analysis. In its earliest phase, the PROIEL corpus included the Greek text of the New Testament, along with its translations in Latin, Gothic, Old Church Slavonic (OCS), and Armenian (Haug *et al.* [5]). The texts were annotated in a layered scheme including lemmatization, morphological annotation, syntactic dependency annotation, and information structure.

Since its beginning, the PROIEL project has continuously grown and has nowadays become a standard for the annotation of ancient IE languages (Eckhoff *et al.* [2]). First, the treebanks of Greek, Latin, and Armenian texts have been expanded thanks to the addition of new textual material. Second, the PROIEL family of treebanks has been enriched with the addition of several newly created resources: the TOROT treebank, which includes Old Russian and OCS (Eckhoff & Berdicevskis [3]), the ISWOC treebank, featuring Old English, Old French, Old Spanish and Portuguese texts (Bech & Eide [1]), as well as new treebanks for ancient Germanic languages (see Eckhoff *et al.* [2] for details), namely Old Icelandic (*Greinir skáldskapar*),

Old Norwegian (*Menotec*), and Old Swedish (*MABIR*). Moreover, the treebanks featured in PROIEL have been recently converted to Universal Dependencies (UD).<sup>1</sup>

In spite of this positive trend of growth, there is still room for improvement, and the PROIEL project can be enhanced by the inclusion of additional IE languages. In this paper, we report on the results of a pilot project aimed at the integration of Hittite texts in PROIEL. Even though Hittite is the most anciently attested IE language and therefore of great interest for Indo-Europeanists, it remains a rather under-resourced language (Giusfredi [4]). First, reliable digital editions of Hittite texts are available only for a sub-set of the extant corpus (cf. *Hethitologie Portal Mainz*; <https://www.hethport.uni-wuerzburg.de/>). A linguistically annotated corpus is still a desideratum, even though this gap is progressively being filled: Inglese [10] laid out the basis for the annotation of Hittite texts according to the UD framework, with a focus on Old Hittite material, and a corpus of Middle and New Hittite material is currently being annotated with a constituency-based grammar at the project of the *Hittite Corpus* (HC; <http://hittitecorpus.ru/>; Molin & Molina [15]; Molina [16]). Therefore, adding Hittite texts to the PROIEL will not only improve the language coverage of the project, but will also considerably contribute to the creation of a much-needed digital resource in the field of Hittitology.

The paper is organized as follows. In Section 2, we briefly sketch the outline of the project and the material employed. We also discuss important issues connected with the preparation of the texts for the annotation and the philological issues that one needs to be aware of before digitizing Hittite texts. Section 3 contains an overview of the main problems encountered in the linguistic annotation of Hittite texts following the PROIEL's guidelines on different levels: tokenization (3.1), lemmatization and morphology (3.2), and syntax (3.3). Also, we briefly touch upon the crucial issue of fragmentary texts (3.4). We summarize our conclusions in section 4.

## 2 The Hittite pilot project

The pilot project was carried out in July and August 2016 and focused on the adaptation of the existing PROIEL annotation scheme to the necessities of Hittite. As a pilot study, we worked on the annotation of three Hittite texts: two New Hittite letters (KUB 19.5 + KBo 19.79, KUB 14.3, ed. by Hoffner [8]) and one Old Hittite instruction text (KBo 22.1, ed. by Miller [14]), for a total of 108 sentences. The annotation was manually performed by two

---

<sup>1</sup> See the project website for details (<http://universalddependencies.org/>). One of the anonymous reviewers asked why we opted for the PROIEL annotation scheme rather than annotating our data directly in the UD format. The reason is two-fold. On the one hand, PROIEL provides a more detailed scheme, in which we can include more structural information. Also, it allows the annotation of the semantic and the pragmatic layer, which is currently unavailable in UD. Another advantage is that the PROIEL scheme is stable, while UD is to some extent a moving target, as the scheme is still under considerable restructuring.

independent annotators (Maria Molina and Guglielmo Inglese) by means of the PROIEL Annotator web interface (Eckhoff *et al.* [2]) and the results and the issues that emerged during the annotation process were subject to extensive group discussions.

## 2.1 Material employed

As remarked in section 1, there is still a substantial lack of comprehensive digital editions of Hittite texts, and philologically reliable editions are mostly scattered across different sources. Clearly, this seriously hampers the possibility to carry out in-depth corpus analyses of the language and constitutes a further stimulus for the creation of a well-structured treebank of Hittite. However, unlike languages currently featured in the PROIEL, for which “the availability of electronic editions [...] is relatively good” (Haug *et al.* [5]: 58), in the long run the inclusion of Hittite in PROIEL will require a good deal of manual digitalization of Hittite texts.

Texts for our pilot project have been kindly provided by Maria Molina from the HC, and are based on up-to-date philological editions. The texts were already split into sentences (see Molina [16] for the criteria behind sentence splitting; cf. Eckhoff *et al.* [2] on sentence splitting in PROIEL), and they were imported into the PROIEL annotation web interface by means of a script created by Hanne Eckhoff.

## 2.2 Text preparation: philological issues

The preparation of the texts for the annotation is not a trivial task, mostly owing to the philological complexity of the Hittite script. Unlike languages currently featured in PROIEL, such as Latin and Ancient Greek, which employ alphabetic scripts, Hittite is recorded in cuneiform script (see Hoffner & Melchert [7] for an overview), which poses several challenges for the digital annotation (Inglese [10]; Molina [16]).

The first issue is how to make the cuneiform script accessible to non-specialists of the language.<sup>2</sup> Two options are generally available: either texts are given in narrow transliteration, that is, each cuneiform sign is represented separately with hyphens as sign boundaries, as in *e-eš-zi* ‘he is’, or texts can be given in broad transcription, which is a rough phonological interpretation of the script, as in *ēšzi*. In our pilot, we have decided to give texts in broad transcription, which makes the corpus more readily available to users less acquainted with Hittite philology. However, it must be stressed that broad transcription requires a relatively high degree of normalization, so that most information about the cuneiform spelling is lost. Therefore, in the next steps of the project the narrow transliteration will be included in the corpus as well (see Inglese [10] for a possible solution), as it provides invaluable

---

<sup>2</sup> One of the anonymous reviewers asked why we have not decided to provide texts in cuneiform script with Unicode encoding. The reason is that texts in transcription are much more easily accessible to readers who have not been trained in Hittite philology. Moreover, Hittite cuneiform manuscripts are already digitized and freely available online at the *HPM*.

information on various linguistic facts, e.g. accent and vowel length, and spelling practices are worth investigating in their own right for various purposes (see e.g. Kloekhorst [13]).

Another peculiar feature of Hittite texts is that beside ‘syllabic’ signs, which stand for syllables in words written in Hittite and are commonly transliterated in lowercase italics, one also finds ‘logographic’ signs, i.e. signs which are read as Akkadian or Sumerian words, and are used as shortcuts for underlying Hittite words. As we discuss below, the annotation of Akkadograms and Sumerograms constitutes a remarkably tricky task. In addition, some Sumerograms, which are labelled ‘determinatives’, were graphically preposed to nouns to indicate the semantic class that a given noun belongs to.

To give an example of the complexity of the Hittite script, consider the passage in example (1), given in narrow transliteration. In this sentence, only the finite verb *ḥ[e]kta* ‘he bows’ is written in Hittite syllabic signs. As for the rest, one finds e.g. the Sumerian logograms LÚ standing for the Hittite nominative form *pešnaš* ‘man’, and the combination of the Akkadian preposition *ANA* ‘to’ with the Sumerogram LUGAL, which together stand for the Hittite dative form *ḥassui* ‘to the king’. Moreover, the determinative sign <sup>d</sup> preposed to the Sumerogram IM ‘Storm God’ indicates that the name refers to a deity.

- (1) LÚ <sup>d</sup>IM      A-NA LUGAL    *ḥ[é-e]k-ta*  
man storm.god to king bow.PRS.3SG.MID  
“The man of the Storm God bows in the presence of the king.” (KBo 20.10 + KBo 25.59 i 5)

### 3 Linguistic annotation

In this section, we illustrate the main problems that we encountered in the linguistic annotation of Hittite texts following the PROIEL scheme. We discuss each layer of annotation separately, and highlight the most problematic issues. Notably, in the pilot the pragmatic level was left out.

#### 3.1 Tokenization

Hittite scribes separated words through blank spaces, so that tokenization is a relatively trivial task. Still, some minor issues emerged in the course of the project. The first issue is how to tokenize and represent clitic chains in Wackernagel’s position (P2), which constitute a rarity among IE languages, but are systematic in Hittite. For instance, the graphic word *nu-wa-aš-ša-an* should be split up as *nu=wa=šan*, i.e. the sequence of the sentence initial connective *nu* plus the quotative particle *=wa* and the local particle *=šan*. For now, we have treated each item in the clitic chain as an independent token, and merely added the = sign to visually indicate token boundaries (cf. Eckhoff *et al.* [2] for the treatment of clitics in Old Portuguese in PROIEL). This is however a provisional solution, as one ideally needs a way to

automatically retrieve whether a given token is a clitic or not. This might be achieved by inserting a dedicated tag at the morphological level.

The tokenization of determinative signs constitutes a further issue. In principle, determinatives can be tokenized either as distinct tokens or as a word-feature (for the discussion of pros and cons of both approaches see Inglese [10]). In our pilot, we have consistently adopted the former option, and treated determinatives in the same way as articles in Ancient Greek (see the treatment of LÚ.MEŠ in Fig. 2, sec. 3.3). It is unclear whether this strategy will be effective in the long run, and the annotation of more material is needed to gain a full appreciation of the issue.

Finally, another problem that was encountered is the treatment of Sumerian and Akkadian multi-word expressions, such as LÚ<sup>GIS</sup>BANŠUR ‘table attendant’ and <sup>d</sup>UTU=ŠI ‘his majesty’, which stand for single Hittite lexemes but are formally made up of multiple tokens in the languages they are written in. For the time being, we have resorted to annotating each token individually and indicating on the syntactic level that the two words belong to a single multi-word expression.

### 3.2 Lemmatization and morphology

Based on our pilot experience, the lemmatization and the annotation of morphological features are the layers of annotation that require the least adaptation of the existing PROIEL scheme.

Concerning lemmatization, for the sake of uniformity we have decided to give lemmas according to Tischler’s glossary [12]. For most words, the stem form is used as the lemma, whereas for suppletive forms and *-r/n*-alternating stems the nominative is used instead. As common practice in PROIEL (cf. Eckhoff *et al.* [2]), homophonous lemmas are distinguished by storing them with variant numbers, e.g. *iya-#1* ‘make’ vs. *iya-#2* ‘march’.

As for the morphological annotation, the tagset of morphological features in use in the PROIEL scheme requires minor modifications only. On the one hand, some of the existing features are not needed and can be simply left out, Hittite being notoriously morphologically simpler than languages such as Ancient Greek and Latin. On the other hand, new features were required, e.g. the ‘ergative’ case for neuter nouns ending in *-anza* when they occur as the subject of a transitive verb.

Further consideration is required for the lemmatization and morphological analysis of logograms. In general, one should decide whether to annotate these tokens according to the features of their surface language or according to the hypothesized features of their Hittite underlying forms. For the lemmatization, this implies a choice between Akkadian/Sumerian and Hittite lemmas for logograms. As the PROIEL scheme allows for a single lemma for each token only, we provisionally employed Hittite lemmas whenever available, and Akkadian/Sumerian ones in the rest of the cases. In the long run, it is desirable to develop a system in which logograms can be assigned both their surface lemma and their underlying Hittite lemma.

The issue of the morphological annotation of logograms is more complex. Beside plural markers on nouns (e.g. MEŠ), Sumerograms tend not to show overt morphological features. Therefore, they can either be left untagged, or they can be annotated according to the morphological features of their putative underlying Hittite forms. The situation of Akkadian is more complex. Unlike Sumerian, Akkadian forms in Hittite texts display a wider range of inflectional features, and some of them do not match the Hittite underlying forms. A case in point is the gender of 3<sup>rd</sup> sg. possessive pronouns, as in Akkadian one finds a masculine/feminine gender distinction =ŠU ‘his’ and =ŠA ‘her’ that is unparalleled in Hittite.

In our pilot, we tried to annotate all logograms according to their underlying Hittite forms, but this proves an unsatisfactory solution, because it greatly limits the possibility to search for logograms and their features in the corpus. Further work is needed to develop a solution to this issue.

### 3.3 Syntax

In PROIEL, the syntactic annotation, which constitutes the core of the treebank, is based on a dependency-style grammar.<sup>3</sup> The scheme was developed for the annotation of ancient IE languages, and it is quite suitable to annotate the syntax of Hittite texts as well. Consider the annotation of the Hittite complex sentence in (2), exemplified in Figure 1.<sup>4</sup>

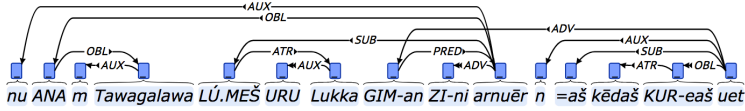


Figure 1: Annotation of a complex clause in Hittite

- (2) [nu] ANA <sup>m</sup>Tawagalawa LÚ<sup>MEŠ</sup> URU Lukka G[IM]-an  
 CONN to T. man(PL) city L. when  
 ZI-ni [a]rnuēr n=aš kēdaš  
 soul.DAT bring.PST.3PL CONN=3SG.NOM DEM.DAT.PL  
 KUR-eaš uet  
 land.DAT.PL come.pst.1sg  
 “As the men of Lukka notified Tawagalawa, he came into these  
 lands.” (KUB XIV i 3-4)

Unsurprisingly, some minor modifications were required to allow for a more precise treatment of Hittite language-specific phenomena. In the first place, Hittite features various Wackernagel’s (P2) clitic particles of partly unclear function, such as the so-called ‘local particles’, the connective

<sup>3</sup> See Eckhoff *et al.* [2] for a useful overview of the dependency grammar in use at PROIEL and the guidelines for details: <folk.uio.no/daghaug/syntactic\_guidelines.pdf>.

<sup>4</sup> In this paper, Hittite dependency trees are visualized with *Arborator* (<https://arborator.ilpqa.fr/>).



particles  $= (m)a$  and  $= (y)a$ , the quotative particle  $= wa(r)$ , and the particle  $= za$ . Following the PROIEL guidelines, since these items fail to show a syntactic function with respect to their head and loosely belong to the group of ‘grammatical’ words, we have consistently annotated them as AUX and assigned them a conventional head. However, we maintain that this annotation style is too opaque, as it does not allow a sufficient differentiation between items bearing the AUX relation. A more fine-grained tagset should be worked out. Similarly, we also annotate preverbs, which are never unverbated with the verbal stem they modify, as AUX, as in the case of *anda* ‘in’ in Fig. 2 below.

Another construction which deserves more attention is the relative clause. So far, relative clauses in PROIEL have been treated as embedded predications depending on a noun, and correlative relative clauses, which marginally occur in e.g. Latin, do not receive a dedicated annotation.

However, there is evidence that correlative relative clauses are not syntactically part of the main clause, as they do not modify an external head noun, nor can they fill in the valency frame of a predicate (cf. Inglese [11] with further references). As correlative clauses constitute the default relativization strategy in Hittite, we have devised a new annotation style to capture the linguistic reality of this phenomenon. In our scheme, the verb of the correlative clause depends on the verb of the main clause via the newly created *rel* tag. As an example, consider the annotation of the sentence in (3) given in Figure 2.

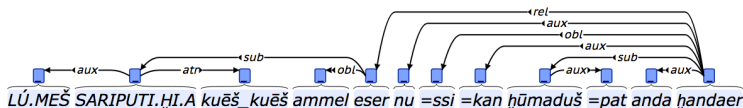


Figure 2: Annotation of correlative clauses

- (3) [LÚ].MEŠ SARIPUTÍ<sup>ĦI.A</sup> kuēš kuēš ammel  
purple-dyer(PL) REL.NOM.PL REL.NOM.PL 1SG.GEN  
eser [nu=ssi=kan ħūmaduš=pat anda  
be.PST.3PL CONN=3SG.DAT=PTC all.NOM.PL=FOC in  
ħandaer  
align.PST.3PL  
‘All the purple-dyers who were mine, they all joined him.’ (KUB  
19.5 + 10)

Finally, Hittite features different periphrastic constructions, or compound verb forms. In these cases, we follow PROIEL’s approach and treat constructions with *ħark-* ‘have’ and *eš-* ‘be’ plus participle as grammaticalized monoclausal constructions when they show a perfect or a passive reading (for discussion see Hoffner & Melchert [7]; Inglese & Luraghi [9]). The annotation of a perfect with *ħark-*, quoted in (4), is

exemplified in Figure 3. As the figure shows, the participle *ħazzian* ‘pierced’ is treated as the head of the predication, whereas the finite verb *ħarzi* ‘has’ is tagged as AUX.

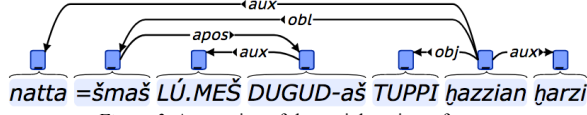


Figure 3: Annotation of the periphrastic perfect

- (4) *natta*=*šmaš* <sup>LÚ.MEŠ</sup>*DUGUD-aš* *TUPPI* *ħazzian*  
 NEG=2PL.DAT dignitary.DAT.PL tablet pierce.PTCP.N/A.N  
*ħarzi*  
 have.PRS.3SG  
 “(As my father keeps writing to you), has he not written the tablet to you dignitaries?” (KBo 22.1 i 23)

Conversely, in the case of the ‘stative’ *ħark-* and *eš-* plus participle and the ‘ingressive’ *dai-/tiya-* ‘put’ plus supine constructions, we take the finite verb as the head of the predication, and tag the accompanying verb as XOBJ. As an example of the treatment of the stative construction, consider the annotation of example (5) in Figure 4. The finite verb *ħarzi* is the root of the tree, and the participle *tamaššan* ‘oppressed’ depends on it as XOBJ.

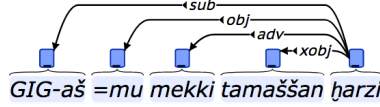


Figure 4: Annotation of the ‘stative’ periphrastic construction

- (5) *GIG-aš*=*mu* [*mekki*] *tamaššan* *ħarzi*  
 illness.NOM=1SG.ACC much oppress.PTCP.N/A.N have.PRS.3SG  
 “Illness keeps me severely prostrated.” (KUB 19.5 + i 5-6)

Notably, the ‘serial’ constructions with *pai-* ‘go’ and *uwa-* ‘come’ do not easily fit in either scheme: since we did not encounter them in our pilot, we leave the design of an appropriate annotation style for the future

### 3.4 Fragmentary texts

Another crucial issue concerns the annotation of fragmentary sentences, i.e. sentences which are only partly readable because of the poor conservation status of the manuscript. As discussed at length by Molin & Molina [15] and Inglese [12], various options are available for the annotation of fragmentary sentences. Given the complexity of the topic, in the pilot we have avoided the annotation of such sentences. In principle, we aim at an annotation halfway between what suggested by Molin & Molina [15] and

Inglese [12]: sentences should receive a sentence tag according to their ‘brokenness’ level, together with a more fine-grained sentence internal annotation of philological gaps as tokens. In the future, this will require a special adaptation of the PROIEL scheme, which so far does not allow the tagging of features at the sentence level.<sup>5</sup>

#### 4 Conclusions and future work

In this paper, we have reported on the preliminary results of a pilot project aimed at the inclusion of Hittite into the PROIEL enterprise. This is a much needed and welcome expansion of the resource. On the one hand, it will enrich the current language coverage of the PROIEL project, while at the same time ensuring the creation of the first dependency-based treebank for Hittite. We have shown that the PROIEL guidelines by and large easily lend themselves to the annotation of Hittite. However, Hittite texts presents several philological difficulties which requires further consideration, in order to provide a reliable and user-friendly digital resource. Finally, we have also discussed how the guidelines should be partly tailored to annotate a number of language-specific constructions of Hittite. Overall, our findings provide the necessary starting point for the creation of a Hittite treebank within the PROIEL framework.

#### References

- [1] Bech, K. & Eide, K. 2014. *The ISWOC corpus*. Department of Literature, Area Studies and European Languages, Oslo.
- [2] Eckhoff, H. M., Bech, K., Bouma, G., Eide, K., Haug, D. T. T., Haugen, O. E., Jøndal, M. 2017. The PROIEL treebank family: a standard for early attestations of Indo-European languages. *Language Resources and Evaluation*, <https://doi.org/10.1007/s10579-017-9388-5>.
- [3] Eckhoff, H. M. & Berdicevskis, A. 2015. Linguistics vs. digital editions: The Tromsø Old Russian and OCS Treebank. *Scripta & e-Scripta* 14–15: 9-25.
- [4] Giusfredi, F. 2014. Web resources for Hittitology. *Bibliotheca Orientalis* 71: 358-362.
- [5] Haug, D. T. T., Eckhoff, H. M., Majer, M., Welo, E. 2009. Breaking down and putting back together: analysis and synthesis of New Testament Greek. *Journal of Greek Linguistics* 9 (1): 56-92.

---

<sup>5</sup> As one of the anonymous reviewers suggests, fragmentary contexts might be handled in a similar way to disfluency phenomena in treebanks of spoken data (see e.g. the guidelines for the annotation of disfluency in UD). This similarity is based on the insight that the two phenomena involve the unexpected disruption of the ‘natural’ syntax of a sentence. However, we suspect that fragmentary contexts are more varied and complex to annotate than disfluencies, so that further work is needed to develop an *ad hoc* solution.

- [6] Haug, D. T. T., Jøhndal, M., Eckhoff, H. M., Welo, E., Hertenberg, M. J. B., & Muth, A. 2009. Computational and Linguistic Issues in Designing a Syntactically Annotated Parallel Corpus of Indo-European Languages. *Traitement Automatique des Langues* 50 (2): 17-45.
- [7] Hoffner, H. A. & Melchert, C. H. 2008. *A Grammar of the Hittite Language. Part I: reference grammar*. Winona Lake (Indiana): Eisenbrauns.
- [8] Hoffner, H. A. 2009. *Letters from the Hittite Kingdom*. Atlanta: Society of Biblical Literature.
- [9] Inglese, G. & Luraghi, S. Forthcoming. The Hittite Periphrastic Perfect. To appear in *Perfects in Indo-European languages*, vol. I, R. Crellin & T. Jügel (eds.). Amsterdam/Philadelphia: John Benjamins.
- [10] Inglese, G. 2015. Towards a Hittite Treebank. Basic Challenges and Methodological Remarks. In *Proceedings of the Workshop on Corpus-Based Research in the Humanities*, Passarotti, M., Mambrini, F., & Sporleder, C. (eds.), 59-68.
- [11] Inglese, G. 2016. La classificazione delle frasi relative in ittita arcaico: una prospettiva tipologica. *Studi e Saggi Linguistici* 54: 9-44.
- [12] Inglese, G. Forthcoming. Annotating the syntax of fragmentary sentences: the case of Hittite. To appear in *Formal Representation and Digital Humanities*, P. Cotticelli & F. Giusfredi (eds.). Cambridge: Cambridge Scholars Publishing.
- [13] Kloekhorst, A. 2014. *Accent in Hittite: A Study in Plene Spelling, Consonant Gradation, Clitics, and Metrics*. Wiesbaden: Harrassowitz.
- [14] Miller, J. 2013. *Royal Hittite Instructions and Related Administrative texts*. Atlanta: Society of Biblical Literature.
- [15] Molina, M. & Molin, A. 2016. In a Lacuna: building a syntactically annotated corpus for a dead cuneiform language (on the basis of Hittite). In *Proceedings of the International Conference "Dialogue 2016"*.
- [16] Molina, M. 2016. Syntactic Annotation for a Hittite Corpus: Problems and Principles. In *Proceedings of the Workshop on Computational Linguistics and Language Science*.
- [17] Tischler, J. 2001. *Hethitisches Handwörterbuch*. Innsbruck: Institut für Sprachwissenschaft.

# Towards an unsupervised learning method to generate international political event data with spatio-temporal annotations

VenuMadhav Kattagoni and Navjyoti Singh

Center for Exact Humanities

International Institute of Information Technology, Hyderabad, India

E-mail: [venumadhav.kattagoni@gmail.com](mailto:venumadhav.kattagoni@gmail.com), [singh.navjyoti@gmail.com](mailto:singh.navjyoti@gmail.com)

## Abstract

Event is one of the most important temporal phenomena. It is that which has a beginning and an end. Automatic coding or classification of events happening in international politics is an important part of social science data ecosystem. The last few decades have witnessed significant work in detecting political events in the international arena. Most of the current work involves classification of news based on pattern matching from a large set of verb patterns, political actors, compound nouns, compound verb phrases, reference to pronouns and deep parsing of sentences in news articles. Through this paper, we introduce a method for generating political events using news media data in the international arena. This method involves graph based unsupervised learning for extracting topics in the news articles and identifying events based on the actors listed as well as spatial and temporal entities.

## 1 Introduction

International relations are mostly framed by pronouncements, engagements, responses, comments or force postures made by the actors. Actors [1] in international relations include individuals, groups (including ephemeral groups like crowds), organizations (including corporate entities, both public and private) and all generally recognized countries (including states and related territories). The main source of such data is news media. We will be referring to news media as media through the rest of the paper. News can be used to analyze interactions between the actors and their relations and can also help in forecasting international conflicts. Breaking down complex events into a sequence of basic events would help researchers analyze international events statistically. We propose a new event model in this regard along with an unsupervised event detection methodology using media data which we perform on the International Relations domain. This can be fine-tuned to work for other domains as well. This model will help build a better storyline

filtering based on attributes in the model and also in further analysis of media bias (actor-centered perspectives) in the current multi-polar world.

We present a brief background on event ontologies and event coding in conjunction with media in Section 2. Our new event model is presented in Section 3. We then present our dataset in Section 4 and methodology for collecting, analyzing and modelling news articles to identify events, in Section 5. The study proceeds to describe the features used along with machine learning techniques for identifying events. In subsequent sections, the results are demonstrated, followed by a discussion. We end the paper with proposals on future work sparked by this study.

## **2 Background and Related Work**

The last few decades have witnessed a considerable escalation in studies which are directed at event coding ontologies in the political domain. This kind of research began during the 1970s with the purpose of forecasting International Conflict under the sponsorship of the U.S. Department of Defense Advanced Research Projects Agency (DARPA) [2], [3]. The kind of research that has been done is mainly on:

1. the political event data coding ontologies.
2. the generation of the political event data.
3. forecasting of international conflict.

The current focus of this paper is on the generation of Political Event Data. This data has existed in various ontologies which include WEIS [4], COPDAB [5], CAMEO [6], IDEA [7] etc. The WEIS Ontology is made up of 22 categories that encompass actions such as Request or Grant. The CAMEO ontology is an upgraded version of WEIS with mediation event types added to it. It is more fine-grained with 20 top-level categories which further contain fine-grained categories in a hierarchical manner. All these systems just give the actors involved and the type of mediation event but lack semantic analysis on the events detected.

Our work presented in this paper carves a similar problem by detecting events whose focus is on the topic of discussion (rather than the type of mediation as in earlier ontologies) and event co-reference resolution system inspired from an article by Donald Davidson about “The Individuation of Events” [8]. This event co-reference system is restricted to events in International Relations domain. Here, we consider only full co-reference of events.

## **3 An Event Model for analyzing International Relations**

Due to complexity of International Events, we started with a minimal conceptual model of event proposed by Westermann and Ramesh Jain [9]. They have mentioned seven elementary aspects of event description viz., temporal, causal, spatial,

experiential, informational and structural aspects. Most of the generic event models such as The Event Ontology [10], LODE [11] and other event models are event-centric and lack all the seven elementary aspects. Also, previous research on event detection focused on mediation types which do not aid semantic analysis of the events detected in the International Relations domain. We propose our own event model for the same as in figure 1.

Each event in the model has attributes such as date-time (Temporal aspect), location (Spatial aspect), actors, media-source, event-title, source-url, sentence (Causal, Experiential, Informational and Structural aspects) which helps in semantic analysis.

## 4 Dataset

Our system listens to 248 media feeds<sup>1</sup> for news articles daily. For this paper, news articles dated only between August 1-15, 2017<sup>2</sup> were used. We used Latent Dirichlet Allocation (LDA) model [12] which we trained on 2 lakh news articles to find topics of international relations and international politics. The number of topics is restricted to 20. Out of the 20 topics detected, we took 3 topics for International Relations which we verified manually. This model filtered a collection of 14846 scraped news articles to 3963 articles. We also manually verified all these 3963 articles to make sure all the articles belong to International Relations and International Politics domain.

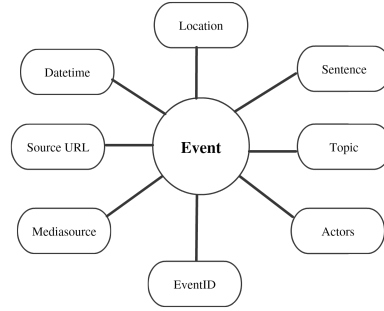


Figure 1: The Event Model

## 5 Methodology

The methodology is described in figure 2. A news article is passed through Graph based topic detection system which gives major topics of the article. The idea of graph based topic detection from an article is adapted from PageRank[13] and TextRank[14] which is explained below. TextRank[14] is a graph based ranking model for text processing specifically KeyPhrase Extraction and Sentence Extraction.

<sup>1</sup>[http://ceh.iiit.ac.in/international\\_relations/source.txt](http://ceh.iiit.ac.in/international_relations/source.txt)

<sup>2</sup>There was no particular reason to use this duration in particular. We started collecting articles from August 1.

1. Tokenization and POS tagging was performed. To avoid excessive growth of the graph size, only unigrams are used as candidates for graphs. Excessive growth of graph would result in computational difficulties.
2. Words which are dependents of International Actors are added as nodes to the graph with an initial rank as 1.
3. Edge is added between those lexical units which occur in the co-occurrence window of  $N$  words. Here, we have set co-occurrence window to 2 words.
4. The reranking algorithm was applied on the graph mentioned in TextRank[14] until it converges at a threshold of 0.0001.
5. Once the final score was obtained at each vertex in the graph, vertices are sorted in the reversed order of their score, and the top 20 vertices are retained.
6. The words of the top 20 vertices which are adjacent are clubbed. Only the Noun Phrases or Verb Phrases thus formed are considered as major topics.

Parallely, the same article is sent for sentence tokenization followed by constituency parsing of the sentence using the StanfordNLP parser. Named entities are then mapped to CAMEO Dictionaries to find International Actors by mapping Nouns and Noun Phrases tags of Penn Treebank POS tags[15] against CAMEO Dictionaries.

After actor detection, we ensure correct event detection by checking whether the major topics found through the graph are present in the sentence. If it is found to be successful, locations in the sentence are found using Gazetter [16] and temporal entities are detected. If temporal entities are relative, we use the published date of the article to find the absolute date and time. All these attributes together form the event.

The intuition behind this methodology is that any news article would report any topic happening connected with the chronological events happened earlier. TextRank [14] is a graph based unsupervised topic detection algorithm. Topic of discussion with the international actors [1] involved in it with spatial and temporal entities together constitute an event. The topic of discussion is identified using the modified TexRank algorithm. We modified TextRank to suit extraction of only International Events. We also have the sentence in which the event is detected so that

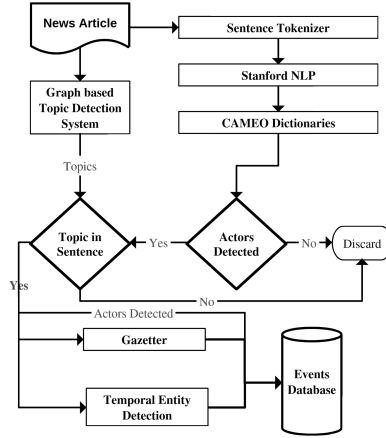


Figure 2: Pipeline of Methodology



further analysis can be done using the action verbs used in the sentence where the event is detected.

Every event that happens across the world is reported by many media sources. So, co-reference among events is required to prevent duplicates in event detection. This would further help in framing analysis of an event by different media sources.

### 5.1 Identifying Event Co-reference

We extrapolate Davidson's[8] theory of "Individuation of Events" and postulate conditions for event co-reference in international relations. If  $x$  and  $y$  are two international events, then  $x \equiv y$  if and only if

1. All the actors involved in  $x$  are identical to actors involved in  $y$ .
2.  $x$  and  $y$  happen at the same location.
3.  $x$  and  $y$  consume identical stretches of time.
4.  $x$  and  $y$  appear on the same topic.

Each media reports locations in the news articles it publishes in different levels granularity based on its audience. We compare parent states and countries obtained from Gazetteer[16] for location matching. For temporal detection, we resolve both absolute datetime and relative datetime.

## 6 Result and Analysis

A random snapshot of 2000 events from 8929 events detected by our system was given to 12 annotators to check for the accuracy of topic detection and event co-reference. The accuracy of topic detection and event co-reference are 86.84 and 90.93 respectively. The inter-annotator agreement numbers for topic detection is 0.78 and for event co-reference is 0.81 in terms of Fliess Kappa[19].

We compared our results with an existing mediation event detection system named PETRARCH [17](based on CAMEO [18]) as shown in table 1. We ignored finding the type of event because we are dealing with all the international relations articles whereas PETRARCH only focuses on mediation event types (CAMEO [18] categories).

We also analyzed the number of events detected by both the systems on 3963 articles. PETRARCH detected 8929 events without co-referencing whereas our system detected 9335 events with 5637 unique events(the remaining events co-refer).

The corpus generated by using this method will let the researchers analyze those events semantically. One such semantic analysis is the temporal analysis of events detected during the Qatar crisis between August 1, 2017 and August 15, 2017 as shown in figure 3.

Feature	PETRARCH	Current
Detection Method	Pattern Matching	Unsupervised
Topic of the event	NO	YES
Spatial Entities	NO	YES
Temporal Entities	NO	YES
Sentence detection	NO	YES
Type of event	YES	NO

Table 1: Comparison with PETRARCH.

## 7 Conclusion and Future Work

Through our work, we intend to give new scope to international relations domain researchers to do further semantic analysis of the events detected. To build on this, we are currently working on a semantic search engine for events happening in international relations domain. The vision of this research is also to find the parts in the news article where the conflict is framed and compare the conflict among various media sources. We also intend to work on a political ontology, which analyzes the relationship of various actors in this multi-polar world. Finally, we also want to analyze international actor interactions in real-time.

## References

- [1] Kan, H. 2009. "Actors in World Politics." *In Government and Politics*, Vol. II, edited by M. Sekiguchi, 242–259. Tokyo: Tokyo Metropolitan University.
- [2] Thomas W. Robinson. 1978. *Forecasting in International Relations: Theory, Methods, Problems, Prospects*. W.H. Freeman, edition 1.
- [3] Stephen J Andriole and Gerald W Hopple. 1988. Defense applications of artificial intelligence. *Lexington Books*.
- [4] Joshua S. Goldstein. 1992. A conflict-cooperation scale for weis events data. *Journal of Conflict Resolution* 36(2):369–385. (URL: <https://doi.org/10.1177/0022002792036002007>.)
- [5] Edward E. Azar. 2009. Conflict and peace data bank (copdab), 1948-1978. *ICPSR Data Holdings* (URL: <https://doi.org/10.3886/icpsr07767.v4>.)
- [6] Deborah J. Gerner, Rajaa Abu-Jabr, Philip A. Schrodt, and mr Yilmaz. 2002. Conflict and mediation event observations (cameo): A new event data framework for the analysis of foreign policy interactions. *In of Foreign Policy Interactions. Paper presented at the International Studies Association*.

- [7] Doug Bond, Joe Bond, Churl Oh, J. Craig Jenkins, and Charles Lewis Taylor. 2003. Integrated data for events analysis (idea): An event typology for automated events data development. *Journal of Peace Research* 40(6):733–745. (URL: <http://www.jstor.org/stable/3648388>.)
- [8] Davidson, D. (1986b). The individuation of events. In: *D. Davidson. Essays on actions and events* (pp. 163-180). Oxford: Oxford University Press. (Reprinted from *Essays in honour of Carl G. Hempel*, pp. 216-234, by N. Rescher, Ed., 1969, Dordrecht: Reidel)
- [9] Utz Westermann and Ramesh Jain. 2007. Toward a common event model for multimedia applications. *IEEE MultiMedia* 14(1):19–29. (URL: <https://doi.org/10.1109/MMUL.2007.23>.)
- [10] A. Abdallah Y. Raimond. 2006. The event ontology. (URL: <http://motools.sourceforge.net/event/event.html>.)
- [11] Raphaël Troncy, Bartosz Malocha, and André T. S. Fialho. 2010. Linking events with media. In *Proceedings of the 6th International Conference on Semantic Systems. ACM, New York, NY, USA, I-SEMANTICS '10*, pages 42:1–42:4. (URL: <https://doi.org/10.1145/1839707.1839759>.)
- [12] Blei, David M.; Ng, Andrew Y.; Jordan, Michael I (January 2003). Lafferty, John, ed. "Latent Dirichlet Allocation". *Journal of Machine Learning Research*. 3 (4–5): pp. 993–1022.
- [13] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. *Technical Report 1999-66, Stanford InfoLab, November 1999*.
- [14] Mihalcea, R., & Tarau, P. (2004). Textrank: Bringing order into texts. In Lin, D., Wu, D. (Eds.), *Proceedings of EMNLP 2004*, pp. 404–411 Barcelona, Spain. Association for Computational Linguistics.
- [15] Beatrice Santorini, "Part-of-Speech Tagging Guidelines for the Penn Treebank Project (3rd Revision)", . July 1990.
- [16] <http://www.geonames.org/>
- [17] Norris et al., (2017). PETRARCH2: Another Event Coding Program. *Journal of Open Source Software*, 2(9), 133, doi:10.21105/joss.00133
- [18] Gerner, Deborah J., Philip A. Schrod, Omur Yilmaz, and Rajaa Abu-Jabr. 2001. "Conflict and Mediation Event Observations (Cameo): A New Event Data Framework for the Analysis of Foreign Policy Interactions."
- [19] Fleiss, J. L. and Cohen, J. (1973) "The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability" in *Educational and Psychological Measurement*, Vol. 33 pp. 613–619

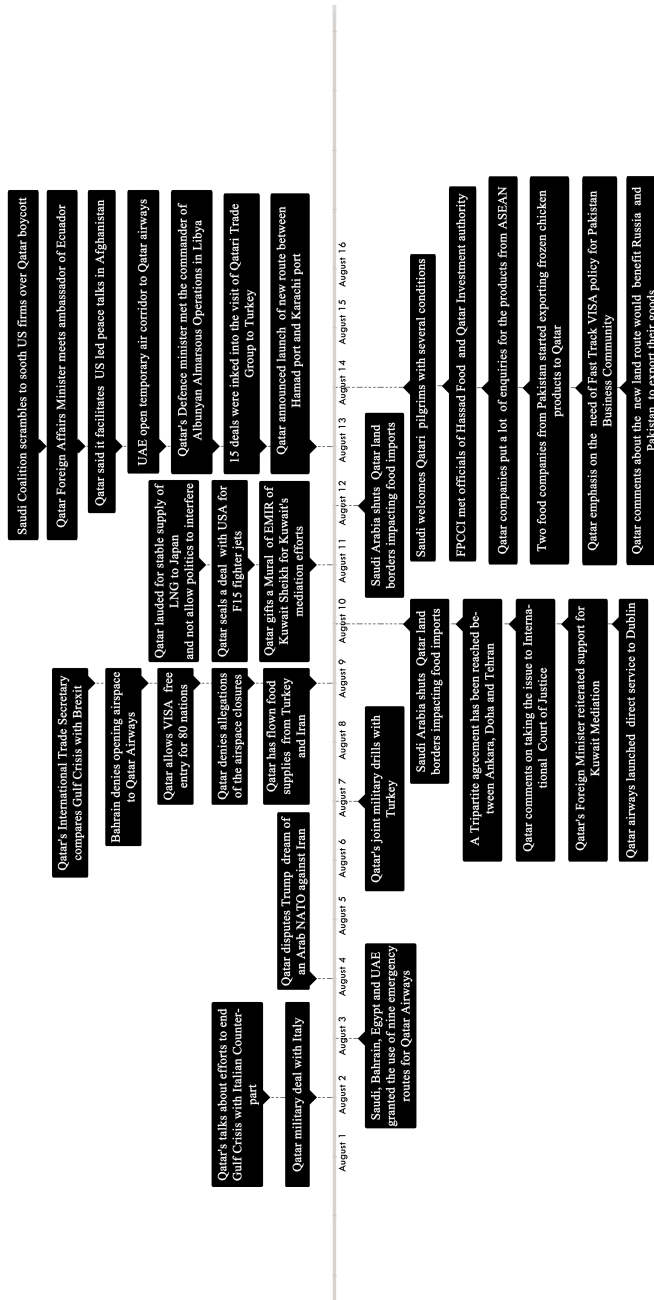


Figure 3: Temporal Analysis of Qatar Analysis.

# Tagging spatial and temporal PPs with two-way prepositions in adult-child and adult-adult conversation in German in Austria

Katharina Korecky-Kröll<sup>1,2</sup> and Lisa Buchegger<sup>1</sup>

<sup>1</sup>Department of Linguistics and <sup>2</sup>Department of German  
University of Vienna

E-mail: katharina.korecky-kroell@univie.ac.at

## Abstract

Expressions of time and space may show high regional, but also considerable inter- and intra-individual variation in corpora of spoken German. Therefore, developing an appropriate tagging system covering all relevant aspects of this variation is a challenge for corpus linguists. On the basis of three Austrian corpora of spontaneous adult-child and adult-adult conversation of participants from different socioeconomic backgrounds, we present a detailed tagging system for PPs expressing mainly spatial and temporal relations, with a focus on two-way prepositions (i.e. prepositions that govern both accusative and dative case). The system differentiates between standard case-marked forms (e.g., accusatives expressing change of location such *in den Kindergarten* ‘to (the) kindergarten’ and datives expressing static location such as *im Kindergarten* ‘in (the) kindergarten’) and colloquial or dialectal case-neutralized forms (*in Kindergarten* ‘to/in kindergarten’). Results show a significant effect of socioeconomic status in child-directed speech and child speech indicating that Austrian children from higher SES families get a more diverse and less ambiguous input with respect to case marking and thus develop case distinctions earlier than their lower SES peers. Likewise, different groups of adult native speakers of rural regions also show differences with respect to case marking in PPs with two-way prepositions.

## 1 Introduction

Prepositional phrases expressing primarily spatial relations may often also be used with metaphorically spatial or temporal meanings, e.g. in expressions such as *er lebt in der Vergangenheit* ‘he lives in the past’ or *ihr Geburtstag fällt auf einen Sonntag* ‘her birthday falls on a Sunday’ (Draye [3]: 96). Particularly interesting cases in German are PPs with two-way prepositions (*an, auf, hinter, in, neben, über, unter, vor, zwischen*) that govern both dative and accusative case and usually have the following semantically different meanings (cf. Duden [4]: 620): Whereas the dative usually expresses a static location, the accusative indicates a dynamic and directional change of location<sup>1</sup>.

---

<sup>1</sup> According to Draye [3], this relationship is more complex and consists in fact of a dichotomy of a marked emerging relationship expressed by the accusative and an

The acquisition of this distinction is a challenge for all children acquiring German as their first or second language (Mills [12], Turgay [17]). Especially children growing up in Upper German areas such as Austria are confronted with varieties that do not always distinguish between accusatives and datives (e.g., Zehetner [19], Weiss [18] for Bavarian) or use constructions that are fundamentally different from those used in Standard German (e.g., Seiler [16] for specific prepositional constructions in certain Bavarian and Alemannic varieties).

Thus, developing an appropriate tagging system for PPs with two-way prepositions in spoken corpora of German in Austria that covers all relevant aspects of this potential variation is a challenge for corpus linguists. We present a first proposal for such a system.

## 2 Method

### 2.1 The tagging system

Table 1 shows the different categories of the tagging system that were added as columns to the Excel files (see also 2.3):

No	Column	Tags (examples)	Comment
1	PREP_form	in, an, auf,...	actual form of the preposition
2	PREP	p	p if there is a preposition, otherwise empty
3	DET_form	d-er, d-ie, d-as,...	actual form of the determiner
4	DET	def, indef, poss,...	type of determiner
5	ADJ	blau-e	form of the attributive adjective (if present)
6	DET_use	DET+N_ADULT, DET+N_OTHER, OCORR, OMISS	correct determiner incorrect determiner correct non-use of determiner determiner omission
7	N_DP_PP	N, DP, PP	bare noun, determiner phrase, prepositional phrase
8	Case	ACC, DAT,...	case (accusative, dative,...)
9	Semantic_cat	spat, spat_met, temp_met, other	semantic category (spatial, metaphorically spatial, metaphorically temporal, other)
10	Stat_dir	stat, dir	static or directional PP
11	Std_stat_dir	stat, dir	Standard German static or directional PP in this context
12	Std_stat_dir_agr	1,0	1 if actual preposition agrees with the Standard w.r.t. static/directional
13	Std_PREP	auf	Standard German preposition for the context (e.g. <i>auf</i> instead of <i>an</i> )

---

unmarked non-emerging relationship expressed by the dative. But for the present analysis, the simpler definition given by the Duden [4] is largely sufficient.

No	Column	Tags (examples)	Comment
14	Std_PREP_agr	1, 0	1 if actual preposition agrees with the Standard form, 0 if not
15	DET_completeness	full, red, contr	completeness of the actual determiner: full determiner, reduced determiner (e.g. <i>dn</i> instead of <i>den</i> ), contraction of preposition and determiner (e.g. <i>im</i> = <i>in dem</i> )
16	Std_DET	d-as	Standard German determiner for the actual context
17	Std_DET_agr	1, 0	1 if actual determiner agrees with the Standard form, 0 if not
18	Std_ADJ	blau-e	Standard German form of the attributive adjective
19	Std_ADJ_agr	1, 0	1 if actual attributive adjective agrees with the Standard form, 0 if not
20	Std_DET_use	categories see 6	Standard German determiner use for the actual context
21	Std_DET_use_agr	1, 0	1 if actual determiner use agrees with the Standard, 0 if not
22	Std_case	ACC, DAT,...	Standard German case
23	Std_case_agr	1, 0	1 if actual case agrees with the Standard German case, 0 if not
24	Std_N_DP_PP_agr	1, 0	1 if all categories 12, 14, 16, 18, 20 agree with the standard, 0 if not

Table 1: Categories of the tagging system for German two-way PPs

This detailed tagging system helps to differentiate between categories that may show different degrees of deviations from Standard German. For example, in some varieties of German in Austria, the two prepositions *an* and *auf* may be used interchangeably in specific contexts. However, the use of the determiner may still be standard-like in these cases.

## 2.2 Participants

### A. Adult-child speech (urban)

Twenty-nine German-speaking parent-child dyads<sup>2</sup> living in Vienna were recorded at their homes at four data points: The children had a mean age of 3;1 (age range: 2;11 – 3;3) at the first recording and mean ages of 3;4, 4;4 and 4;8 at the three follow-up recordings.

Each recording lasted for one hour, and the best 30 minutes with the richest parent-child interaction were selected for transcription. As parents

---

<sup>2</sup> The corpus analyzed for this study is part of the larger INPUT project (“Investigating Parental and Other Caretakers’ Utterances to kindergarten children”, WWTF SSH11-027) financed by the Vienna Science and Technology Fund (WWTF) and led by Wolfgang U. Dressler from March 2012 to September 2016.

were asked to continue with their normal activities, situations showed considerable variation: Some parents asked their children to play a game, others decided to read storybooks, still others just engaged in spontaneous conversation.

The children were nearly balanced for socioeconomic status (SES) and gender (see Table 2).

Child SES	Child gender	N of children	Subtotal SES
HSES	female	8	15 HSES
HSES	male	7	
LSES	female	6	14 LSES
LSES	male	8	
Total		29	

Table 2: Child participants

Following other studies on language acquisition (cf. Ensminger and Fothergill [5]), SES was mainly assessed by the main parental caretaker's highest educational level (cf. OECD [13]): The LSES group included ISCED-97 levels 1 to 3b (i.e. from compulsory school to apprenticeship and vocational schools, but without high school diploma), whereas the HSES group had ISCED-97 levels 3a to 6 (i.e. from high school diploma up to PhD), see also Czinglar et al. [2].

For the present paper, the entire corpus of child speech (CS) as well as of child-directed speech (CDS) was analyzed.

#### B. Adult-adult speech (urban)

This corpus consists of two parts:

- 1) Twenty-nine 30-minute interviews with the same Viennese parents as investigated in corpus A<sup>3</sup> (topics: their daily routines and preferred activities with their children, their children's development and future,...)
- 2) Spontaneous mealtime conversations during family celebrations (e.g. birthdays, Christmas) of a Viennese family<sup>4</sup> consisting of 6-10 (mostly 8) adults of different ages and socioeconomic backgrounds (topics: politics, religion, education, travelling, health,...), see also Korecky-Kröll [9].

We successfully started tagging the corpus, but are not able to present final results yet.

<sup>3</sup> This corpus is also part of the larger INPUT project (see fn. 2).

<sup>4</sup> Transcription of this corpus was supported by the project "Morphologische Verbfamilien im Hebräischen und im Deutschen" financed by the "Gesellschaft der Freunde der Universität Tel Aviv in Österreich".



### C. Adult-adult speech (rural)

This corpus<sup>5</sup> also consists of two parts:

1) One-hour interviews with old and young dialect speakers from different rural areas all over Austria (topics: language biography, current language use, attitudes and perception of dialect and standard language, historical and future language change and contact, preferred activities, place attachment):

As the interviewers spoke Standard Austrian German, the participants usually spoke their intended standard language (Lenz [10]).

2) One-hour spontaneous conversations of the same dialect speakers as in C1 with their friends (same topics as in C1): As their friends were from the same dialect area, they spoke the intended local dialect (Lenz [10]).

So far, a subsample of spontaneous adult-directed speech (ADS) conversations of four dialect speakers (2 younger people from HSES, 2 older people from LSES backgrounds were analyzed).

### 2.3 Procedure

Two different systems were used for transcription and part-of-speech and morphology tagging, namely CHILDES (MacWhinney [11]) for corpus A and B (Korecky-Kröll [9]) as well as EXMARaLDA (EXMARaLDA [6]; Schmidt and Wörner [15]) for corpus C.

In order to insert the tags for the spatial and temporal expressions, the tagged transcripts were first imported into MS Excel as .csv files by using a JavaScript program (Korecky [7]). Another JavaScript program (Korecky [8]) performed a first automatic tagging of all noun phrases (including PPs). The relevant two-way PPs containing spatial and temporal as well as other expressions were filtered and checked and further tags were added to new columns (see categories described in section 2.1). The tagged Excel files were saved in csv format and imported in R.

For the statistical analysis, we used the lme4 package (Bates et al. [1]) of R (R Core Team [14]) to conduct generalized linear mixed effects analyses (glmer) of the relationship between Standard German marking of PPs (the dependent variable) and SES (the main independent variable) to discover group differences between participants. Data point (DP) was another fixed factor for the analyses of corpus A (as there were four data points for spontaneous speech recordings in children's homes), but not for the analyses of corpus C which comprised only recordings of one data point. The total number of noun tokens (log-normalized) per speaker and recording was included as a normalizing variable to account for different amounts of speech of different speakers: The more nouns a speaker uses in one recording, the

---

<sup>5</sup> This corpus is part of the larger SFB "German in Austria. Variation – Contact – Perception" (F60, principal investigator: Alexandra N. Lenz) financed by the Austrian Science Fund (FWF) since January 2016. It belongs to project part (PP) 03 "Speech Repertoires and Varietal Spectra", which is also led by Alexandra N. Lenz.

more PPs containing nouns he or she will use as well. The speaker ID (e.g., the ID of the child in the CS data) as well as the ID of the conversation partner (e.g., the ID of the parent in the CS data) were entered as random factors in all models in order to account for individual variation and also for two special cases (e.g., for the case of a mother of twins who was the conversation partner of both children). The dependent variable (e.g., Std\_case\_agr) was binomial: If the actual PP coding agreed with the Standard German PP coding, it was coded as 1, whereas it was coded as 0 if there was some deviation from Standard German (see also Table 1).

To test the tagging system, we performed two example analyses for corpus A and C: first on the category of standard-like case marking (Std\_case\_agr, see section 3.1) and second on the overall category of standard-like PP use (Std\_N\_DP\_PP\_agr, see section 3.2).

### 3 Results

#### 3.1 Standard-like case marking in PPs

Results for the category of standard-like case marking in PPs in corpus A show a significant effect of SES in CDS ( $\beta = -0.859$ ,  $SE = 0.285$ ,  $p = 0.003$ ) indicating that Austrian children from lower SES families get a less standard-like input w.r.t. case marking in PPs than children from higher SES families (see Table 3). Apart from a significant effect of the normalizing variable of noun tokens, we also find an effect of data point, namely more standard-like case forms at DP 3 compared to the three other data points. This may be due to the fact that more parents chose to read books to their children at this data point (and book-reading yields a more standard-like language use).

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-1.9644	1.6604	-1.183	0.23677
SES <sub>low</sub>	-0.8586	0.2850	-3.012	0.00259 **
DP 2	0.2276	0.2476	0.919	0.35812
DP 3	0.6497	0.2733	2.377	0.01747 *
DP 4	0.2070	0.2474	0.837	0.40262
N_TOK_log	2.1643	0.7027	3.080	0.00207 **
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

Table 3: GLM of standard-like case marking in PPs (CDS): fixed effects

A related SES effect is also found in the children’s output ( $\beta = -1.119$ ,  $SE = 0.436$ ,  $p = 0.010$ ). Thus, children from lower SES families also produce a less standard-like output with respect to case marking in PPs than children from higher SES families (see Table 4). In contrast to the parents, we do not find a significant effect of the normalizing variable of noun tokens, showing that children that use many nouns do not necessarily use many PPs with standard-

like case marking. However, the effect of data point found in parents is also replicated in the children: A higher amount of book-reading also leads to a more standard-like use of case marking in children's PPs.

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	0.4993	1.5424	0.324	0.74617
SESlow	-1.1186	0.4361	-2.565	0.01031 *
DP 2	0.5158	0.3214	1.605	0.10851
DP 3	1.0536	0.3303	3.190	0.00142 **
DP 4	0.4345	0.2805	1.549	0.12138
N_TOK_log	0.4423	0.7436	0.595	0.55194
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

Table 4: GLM of standard-like case marking in PPs (CS): fixed effects

Similar results for the category of standard-like case marking in PPs are found in the preliminary analysis of corpus C (rural ADS, see Table 5) showing that older participants of lower SES produce fewer standard-like case forms than younger higher participants ( $\beta = -0.969$ ,  $SE = 0.274$ ,  $p < 0.001$ ).

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-1.0973	1.9565	-0.561	0.57491
SESlow	-0.8970	0.2839	-3.159	0.00158 **
N_TOK_log	1.0630	0.7151	1.486	0.13715
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

Table 5: GLM of standard-like case marking in PPs (ADS): fixed effects

Nevertheless, more data from younger and older people from different educational backgrounds will be required to tease apart the effects of SES and age for corpus C.

### 3.2 Standard-like PP use

The second example analysis investigates the overall PP use w.r.t. Standard German: A PP will only get a tagging of 1 (standard-like) if all subcategories listed in Table 2 also agree with Standard German. Therefore, this analysis is stricter than the analysis of case marking presented in 3.1.

In contrast to the results on case marking, we do not find a significant SES effect in parents' input (Table 6), indicating that overall use of PPs is largely similar in urban CDS, regardless of parents' SES. Nevertheless, we find a significant SES effect in urban children's output (Table 7) and also in rural adult-directed speech (Table 8). As far as effects of data point are concerned, DP 3 yields again a more standard-like PP use in both parents (Table 6) and children (Table 7). In children, this is also true for DP 4, which

might be an evidence of an age effect, insofar as older children show a more advanced and standard-like use of PPs.

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-3.1864	1.1379	-2.800	0.005105 **
SES <sub>low</sub>	-0.1741	0.1904	-0.914	0.360746
DP 2	0.2967	0.1741	1.704	0.088301 .
DP 3	0.6907	0.1889	3.657	0.000256 ***
DP 4	0.1933	0.1737	1.113	0.265741
N_TOK_log	2.1100	0.4832	4.366	1.26e-05 ***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

Table 6: GLM of standard-like PP use (CDS): fixed effects

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	1.0476	1.3291	0.788	0.4306
SES <sub>low</sub>	-0.6988	0.3354	-2.083	0.0372 *
DP 2	0.4596	0.2754	1.669	0.0952 .
DP 3	0.6625	0.2603	2.545	0.0109 *
DP 4	0.5175	0.2491	2.078	0.0377 *
N_TOK_log	-0.1992	0.6449	-0.309	0.7574
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

Table 7: GLM of standard-like PP use (CS): fixed effects

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-0.5593	1.9096	-0.293	0.769620
SES <sub>low</sub>	-0.9693	0.2739	-3.539	0.000402 ***
N_TOK_log	0.8357	0.6963	1.200	0.230070
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

Table 8: GLM of standard-like PP use (ADS): fixed effects

## 4 Discussion and conclusion

Although our results are only preliminary, they show interesting and plausible evidence with respect to SES differences in language use of people of different age groups living in Austria.

Furthermore, our tagging system proved to be appropriate for covering all relevant aspects of spatial and temporal PPs with two-way prepositions in German in Austria so far. Nevertheless, more data of corpus C from further regions of Austria will show whether the system is indeed sufficient or whether new categories must be added.

A future goal is the creation of another program that exports the new tags back to the original systems (CHILDES, EXMARaLDA) by inserting new

tagging tiers that may be searched automatically by the program-specific commands (e.g. COMBO for combining groups of words in CHILDES, cf. MacWhinney [11]). Although MS Excel is a convenient tool that allows even less experienced student assistants to add new tags very fast and to check these tags for consistency, its performance will decrease as soon as the corpora get too big (on MS Excel for Mac, this holds already for corpus A). As corpora B and C are still growing, this step will be necessary to ensure that our system is also fit for the future.

## Acknowledgements

We thank all funding organizations, namely the Vienna Science and Technology Fund WWTF, the “Gesellschaft der Freunde der Universität Tel Aviv in Österreich” and the Austrian Science Fund FWF, for their financial support of the projects as well as the University of Vienna for funding our positions. Furthermore, we are deeply grateful to the project leaders, Wolfgang U. Dressler and Alexandra N. Lenz, for allowing us to use their great corpora for our analyses. We also thank numerous students of the Department of Linguistics and the Department of German of the University of Vienna for their thorough transcription work. Finally, we thank three anonymous reviewers for their helpful comments.

## References

- [1] Bates, Douglas, Mächler, Martin, Bolker, Ben and Walker, Steve. (2015) Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1), 1-48.
- [2] Czinglar, Christine, Korecky-Kröll, Katharina, Uzunkaya-Sharma, Kumru and Dressler, Wolfgang U. (2015). Wie beeinflusst der sozioökonomische Status den Erwerb der Erst- und Zweitsprache? Wortschatzerwerb und Geschwindigkeit im NP/DP-Erwerb bei Kindergartenkindern im türkisch-deutschen Kontrast. In Köpcke, Klaus-Michael & Ziegler, Arne (eds.), *Deutsche Grammatik in Kontakt. Deutsch als Zweitsprache in Schule und Unterricht*, pp. 207-240. Berlin: de Gruyter.
- [3] Draye, Luk (2014) German two-way prepositions and related phenomena. In Delbecque, Nicole, Lahousse, Karen, Van Langendonck, Willy (eds.) *Case and grammatical relations across languages*, vol. 6, *Non-nuclear Cases*, pp. 95-125. Amsterdam: Benjamins.
- [4] Duden (2016) Die Grammatik. 9<sup>th</sup> edition. Edited by Angelika Wöllstein and the Dudenredaktion. Berlin: Dudenverlag.
- [5] Ensminger, Margaret E. and Fothergill, Kate E. (2003) A decade of measuring SES: What it tells us and where to go from here. In Bornstein, Marc H. and Bradley, Robert H. (eds.) *Socioeconomic*

- status, parenting and child development*, pp. 13–27. Mahwah, NJ: Erlbaum.
- [6] EXMARaLDA (URL: <http://www.exmaralda.org>)
  - [7] Korecky, Paul C. (2015) CLANTOCSV [Computer software].
  - [8] Korecky, Paul C. (2016) NP-Kodierer [Computer software].
  - [9] Korecky-Kröll, Katharina (2017) Kodierung und Analyse mit CHILDES: Erfahrungen mit kindersprachlichen Spontansprachkorpora und erste Arbeiten zu einem rein erwachsenensprachlichen Spontansprachkorpus. In Resch, Claudia and Dressler, Wolfgang U. (eds.) *Digitale Methoden der Korpusforschung in Österreich*, pp. 85-113. Vienna: Austrian Academy of Sciences Press.
  - [10] Lenz, Alexandra N. (2003) Zur Interpretation des Intendierten Ortsdialekts. In Lenz, Alexandra N., Radtke, Edgar and Zwickl, Simone (eds.) *Variation im Raum. Variation and Space*, pp. 113-131. Frankfurt a.M. et al.: Lang.
  - [11] MacWhinney, Brian (2000) The CHILDES Project: Tools for Analyzing Talk. 3<sup>rd</sup> edition. Mahwah, NJ: Erlbaum.
  - [12] Mills, Anne E. (1985) The acquisition of German. In Slobin, Dan I. (ed.) *The crosslinguistic study of language acquisition, Vol 1: The data*, pp. 141-254. Hillsdale, NJ: Erlbaum.
  - [13] OECD (1999) *Classifying Educational Programmes. Manual for ISCED-97 Implementation in OECD Countries*. OECD: 1999 Edition.
  - [14] R Core Team (2015) R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. (URL: <http://www.R-project.org/>)
  - [15] Schmidt, Thomas and Wörner, Kai (2014) EXMARaLDA. In Durand, Jacques, Gut, Ulrike, Kristoffersen, Gjert (eds.) *The Oxford Handbook of Corpus Phonology*, pp. 402-419. Oxford: Oxford University Press.
  - [16] Seiler, Guido (2002) Prepositional Dative Marking in Upper German: A Case of Syntactic Microvariation. In: Barbiers, Sjef, Cornips, Leonie and van der Kleij, Susanne (eds.) *Syntactic Microvariation*. Amsterdam: Meertens Institute Electronic Publications in Linguistics, vol. II (URL: <http://www.meertens.knaw.nl/books/synmic/pdf/seiler.pdf>)
  - [17] Turgay, Katharina (2011) Der Zweitspracherwerb der deutschen Kasus in der Präpositionalphrase. *Zeitschrift für Germanistische Linguistik* 39, 24-54.
  - [18] Weiß, Helmut (1998) *Syntax des Bairischen. Studien zur Grammatik einer natürlichen Sprache*. Tübingen: Niemeyer.
  - [19] Zehetner, Ludwig (1978) Kontrastive Morphologie: Bairisch/Einheitssprache. In Ammon, Ulrich, Knoop, Ulrich and Radtke, Ingulf (eds.): *Grundlagen einer dialektororientierten Sprachdidaktik*, pp. 313-331. Weinheim: Beltz.

# Annotation and Classification of Locations in Folktales

Matthias Lindemann<sup>1</sup>, Stefan Grünewald<sup>1</sup> and Thierry Declerck<sup>2</sup>

<sup>1</sup> Department of Language Science and Technology  
Saarland University

E-mail: {mlinde|stefang}@coli.uni-saarland.de

<sup>2</sup> DFKI GmbH, Multilingual Technologies  
E-mail: declerck@dfki.de

## Abstract

In the context of a software project dedicated to the automated classification of folk and fairy tales, we focused on their segmentation by scenes and their respective locations. In contrast to novels, fairy tales are often taking place at the same types of locations, such as castles, in the forest, in a small hut, and the like. That is, locations can be considered as a feature for supporting the general classification of folktales. In this paper, we describe our first annotation approaches for supporting the automatic detection of locations in folktales that are in German language. To our knowledge, this is the first work on automatically detecting locations in folktales.

## 1 Introduction

In the context of a software project conducted at the Department of Language Science and Technology of the Saarland University we were dealing with the classification of folktales along the lines of schemes proposed by [2], [5] or [6]. One group focused on testing the relevance of segmenting tales by their described scenes. An important aspect of a scene is the location in which it takes place. Contrary to other literary genres, fairy tales seem to have recurrent locations across stories, like castles, forests, small huts, etc. The occurrence of locations can thus be considered as a feature that supports the classification of tales.

We started an investigation on this topic and concentrated in a first step on creating a corpus the annotation of which aiming at supporting the automated detection of locations in folktales written in German language. The task of location detection can be divided into three subtasks, whereas in this paper we only cover the first two subtasks:

1. Recognition of the scene boundaries (“segmentation”)
2. Recognition of the type of location where the scene takes place (“classification”)
3. Recognizing whether two identical locations from different scenes are the same location (“identity”)

For the creation of the corpus we wrote a crawler and downloaded text from 41 collections of tales, with a total of 1880 stories in German from all over the world. The main source is *Projekt Gutenberg*<sup>1</sup>. We excluded very small collections and lyric folktales because they differ much in style. The corpus we assembled from the web crawl contains about 4,3 millions tokens.

For our work on location detection, we first needed to check for which types of locations we could gain enough training data for applying a statistical approach. For this, the corpus has been tagged with the help of the TreeTagger<sup>2</sup> and we looked for the most frequent nouns expressing a location. We were also interested in knowing if a scene is occurring within or outside a location. However, for most types of locations we decided that they are too infrequent and would result in sparsity issues. Therefore, we make this distinction only for the locations “house” and “castle”. This corpus is the basis for the different types of annotation we are providing: manual and automated.

## 2 Manual Annotation

### 2.1 Annotation Guidelines

We established annotation guidelines for the annotation of segments and locations in tales. The main objective was to find and mark segments in which maximally one location is “involved”. But we also allow to mark segments in which more than one location is “involved”, in case it is not possible to avoid it.

Following those guidelines three tales have been annotated by six project participants. One tale was taken from the Grimm collection, one tale is by Andersen and one tale was taken from the “One Thousand and One Nights”. After this first annotation exercise, we adapted the guidelines in order to respond to encountered issues and problematic cases.

In the new version of the guidelines, a more precise specification for “segment” was given: segment boundaries are given by punctuation signs (excepting commas) and paragraph boundaries. This made it easier to agree on the same level of granularity. We derived 24 different types of locations from the corpus<sup>3</sup>. Table 1 illustrates examples from the guidelines, here translated into English.

<sup>1</sup><http://gutenberg.spiegel.de/genre/marchen-fairy>

<sup>2</sup>cf. [4] and <http://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/treetagger.html>.

<sup>3</sup>Turm, Wüste, Küche, Saal, Schloss\_innen, Schloss\_aussen, Wald, Haus\_aussen,



Location	Description
Desert	A desert, sand/stone, not in a metaphorical sense
Kitchen	A kitchen (i.e. a room on its own) e.g. in a castle or house
Hall	Hall, ballroom, throne room in a castle,...
Castle_inside	bedroom, private room or study in castle, e.g. a chamber, possibly also a corridor, stairs,....
Castle_outside	castle, palace, villa from outside, i.e. in the open air, balcony or inner courtyard
Church	Church, religious buildings
Nowhere	No place; as if it hadn't been annotated or if having an off-voice, like in a film.

Table 1: Excerpt from the guidelines: which places are to be interpreted in which way.

## 2.2 Inter-Annotator Agreement

To calculate the inter-annotator agreement, we have chosen Cohen's  $\kappa$ , which is computed pairwise between the annotators<sup>4</sup>. However, it is not directly applicable to cases where each instance (i.e. segment) contains more than one label (type of location). [3] adjust the calculation of the  $\kappa$  statistics so that instances can have a main label and a secondary label with different weights. As this is not the case in our work, we just generalize the calculation for  $n$  labels per instance, each with the same weight. However, it is not possible to directly compare two scenes because their corresponding segments will only be the same if the annotators fully agree on the boundaries. To work around the problem, we have used every word occurring in a selected segment as an instance that carries the labels. Table 2 shows the details of the pairwise inter-annotator agreement computation on a small sample of three stories.

	Annot_1	Annot_2	Annot_3	Annot_4	Annot_5
Annot_2	<b>0.686</b>				
Annot_3	0.569	<b>0.718</b>			
Annot_4	0.519	<b>0.587</b>	0.545		
Annot_5	0.469	<b>0.572</b>	0.505	0.445	
Annot_6	<b>0.655</b>	0.563	0.499	0.405	0.27

Table 2: Pairwise inter-annotator agreement at word level on three folktales after the adaptation of the guidelines

Haus\_innen, Weg, Stadt, Garten, Feld, See, Fluss, Meer, Höhle, Zelt, Stall, Kirche, Gefängnis, Wirtshaus, Mühle, Nirgendwo. Which translates to *Tower, Desert, Kitchen, Hall, Castle\_inside, Castle\_outside, Forest, House\_outside, House\_inside, Way, City, Garden, Field, Lake, River, Sea, Cave, Tent, Stable, Church, Prison, Inn, Mill, Nowhere*.

<sup>4</sup>See [1] for a discussion of Cohen's  $\kappa$  and other methods for measuring agreement among corpus annotators.

## 3 Towards the Automatic Segmentation

### 3.1 Features

We represent a segment by a bag of features. To reduce data sparsity compared to using bare words, we tagged and lemmatized all words with the TreeTagger [4] in a preprocessing step, focusing then mostly on open class words. We also use the lemmatized words to record the information whether they have occurred within literal speech (adding to the lemma a quotation mark). This way, we can differentiate between locations being mentioned by the narrator or by a character of the tale. Moreover, prepositions and their corresponding noun phrase's heads seem too important for the identification of locations to lose their connection by the bag of features assumption. Thus, we merged the preposition and its noun phrase's head.

For example, from the passage "*Wie kannst du es wagen," sprach sie mit zornigem Blick, "in meinen Garten zu steigen?"*"<sup>5</sup> the following features are extracted: {können", wagen", sprechen, zornig, Blick, in\_Garten", steigen"} ({can", dare", say, angry, gaze, in\_garden", climb"}). In addition, we also mark when a noun is modified by a negative expression like "kein" (*none*), as in "Aber es war kein Meer zu sehen" (*but there was no sea to be seen*). Here the extracted feature is !\_Meer and not Meer.

### 3.2 Segmentation

The basic idea for the automatic segmentation is that scenes are associated with characters through time and space and that scenes boundaries correspond somehow to changes of temporal and location information. That is, as soon as a movement of the main characters or some time is passing/jumping, a scene boundary must exist. Additionally, we assume there can be no scene boundary within direct speech. Regular expressions are used to determine whether there is any movement, for example "( |heim|zurück|um|wieder) (kehren|gekehrt)"<sup>6</sup>

A list of verbs expressing movements was extracted from the corpus also considering frequency information. Imperative forms of such verbs are extracted from the literal speech. Another strategy will have to be implemented for detecting time jumps, as those are typically marked by phrases.

In automatic segmentation by locations, the input text is first separated at punctuation marks or paragraph boundaries on which, according to the guideline, it is possible to segment (see Section 2.1) and, in a second step, it is reassembled anywhere where neither movement nor a time jump are observed. If there indeed is a movement or time jump detected, the task is to decide whether the segment in question (the one with movement or time jump) should be attached to the previous or to the following segment. To illustrate this procedure, the following example has

<sup>5</sup>In English: "How dare you," she said with angry gaze, "to climb into my garden?"

<sup>6</sup>This matches infinitive and participle forms of *return* and *return home*

three provisional segments one of which contains a movement. Since the movement verb **go** is used *at the beginning* of the second segmentation unit, the segmentation algorithm decides to transfer the second block to the third one<sup>7</sup>.

“[...] setz dich darunter und warte, bis die Nacht kommt, so wirst du schon das Gruseln lernen.”

Da **ging** der Junge zu dem Galgen, setzte sich darunter und wartete, bis der Abend kam.

Und weil ihn fror, machte er sich ein Feuer an;

## 4 Classification

We implemented three approaches for the classification: rule-based, statistical and hybrid. The rule-based classifier applies a keyword-spotting method on the features. To classify a segment as a location, at least one feature of the segment must match a specific regular expression. At the same time, it must not match another regular expression (a kind of blacklist). We call a feature a *key feature* if it fulfills these requirements. This blacklisting is used to cope with German compound nouns. For instance, a simplified rule is<sup>8</sup>:

$(ins?|im)_{-}.*[Hh]aus\wedge\neg[(Gottes|Schnecken|Vogel)haus]\rightarrow HAUS\_INNEN$

*Key features* of this rule are for instance "in\_Haus", "im\_Räuberhaus", but "Haus", "in\_Schneckenhaus" etc. are rejected. If several rules apply to the same segment, the rule-based classifier chooses the location with highest prior probability.

The generated corpus is a necessary prerequisite for the use of statistical methods. However, as we do not have labeled training data in necessary quantity, the rule-based classifier must be used to first annotate the corpus. We selected a Naïve Bayesian approach for training the model, and for this crossed the corpus with a window of seven features both the left and to the right. Whenever a key feature appears in the centre of the window, the content of the window is evaluated as a joint observation of the classified location with the features. We have observed that models are better when they use a context window that distributes weight unevenly so that features further away from the key feature in the middle have a lower weight.

We implemented two versions of this approach, a “simple” one and one with two stages that first performs a binary classification task (BUILDING or NON-BUILDING) to narrow down the set of possible classes which the simple approach has to choose from. Since the training data for this classifier is also generated with a

<sup>7</sup>In English: ““[...] sit underneath it and wait for the night to come, so you will find yourself to learn the fear.” Then the boy **went** to the gallows, sat underneath and waited until the evening came. And because he was freezing cold, he started a fire;

<sup>8</sup>Translations: Gotteshaus - house of prayer, Schneckenhaus - snail shell, Vogelhaus - birdhouse

rule-based system, we can now write rules that can identify buildings but are not specific enough to identify the type of the building, for instance<sup>9</sup>:

. \* ([Zz]immer| [Dd]ach| [Ff]enster) → BUILDING

Instead of keeping the rule-based and statistical approach separate, we also combine them, since the rule-based approach is relatively precise, but in return does not make a statement for some segments. The classification procedure is as follows: First, the rule-based classifier is applied. If there is exactly one result, this location is predicted; if there are multiple results, the statistical classifier with two stages is applied but restricted to the set of locations that the rule-based classifier found. If the rule-based approach does not find any location at all then all location types are taken into account by the statistical model.

## 5 Evaluation

For the purpose of evaluation, we created a development set consisting of the three folktales (Annot\_2, see Section 2.2) and 10 additional locally segmented and annotated tales from the corpus (180 annotated segments). There are two simple methods of evaluation, the first one being the evaluation of the classification with the usual metrics. For that, the segmentation has to be given. The second one is a joint evaluation of segmentation and classification, i.e. calculating agreement. We pursue both methods.

For calculating accuracy we consider a classification to be correct if the predicted label is in the set of the annotated labels.<sup>10</sup> On average, there are 1.144 labels per segment. Here we display in Table 3 a small summary of the evaluation when the segmentation is given.

Approach	Accuracy	Mean F-Score
Majority class	0.15	0.01
Rule based	0.45	0.41
Naive Bayes	0.43	0.38
Two-stage Naive Bayes	0.43	0.32
Hybrid	<b>0.53</b>	<b>0.46</b>

Table 3: Accuracy and (arithmetic) mean f-score over all classes of different approaches

The good performance of the hybrid model can be explained: as long as there is only one key feature, the rule-based classifier is applied. In the case of several results, the statistical approach is applied among the hits for taking a better informed decision.

<sup>9</sup>Translations: Zimmer - room, Dach - roof, Fenster - window

<sup>10</sup>This is a simplification, of course. It might be, that the location of a scene cannot be disambiguated but it has to be consistent over the tale.

		Annot_1	Annot_2	Annot_3	Annot_4	Annot_5	Annot_6
Manual	Annot_2	<b>0.686</b>					
	Annot_3	0.569	<b>0.718</b>				
	Annot_4	0.519	<b>0.587</b>	0.545			
	Annot_5	0.469	<b>0.572</b>	0.505	0.445		
	Annot_6	<b>0.655</b>	0.563	0.499	0.405	0.27	
Automatic	Two_stage_NB_seg	0.327	0.382	0.369	0.362	0.271	0.259
	Two_stage_NB	0.279**	0.295**	0.361**	0.319**	0.057	0.231**
	Hybrid_seg	0.395	<b>0.55</b>	<b>0.526</b>	<b>0.418</b>	<b>0.312</b>	0.296
	Hybrid	0.243	0.264	0.26	0.24	-0.008	0.225
	NB_seg	0.309	0.363	0.312	0.343	0.29	0.218
	NB	0.214	0.24	0.226	0.22	0.197**	0.194
	Rule-based_seg	<b>0.421</b>	0.5	0.405	0.329	0.272	<b>0.33</b>
	Rule-based	0.272	0.276	0.249	0.256	0.152	0.224

Table 4: Agreement between annotators and models. *NB* stands for Naive Bayes and *seg* means that a gold segmentation was given. The best agreement without the gold segmentation marked with \*\*.

Table 4 compares the inter-annotator agreement on the three folktales between manual annotation and automatic annotation. Classifiers with *seg* don’t have to call the automatic segmentation but receive the segmentation of Annot\_2. The most striking difference in agreement is to be noticed when comparing the same classifier with automatic segmentation and with gold segmentation. Consistent with the good accuracy of the hybrid model on the development set, it performs well in terms of agreement. When comparing agreement of different classifiers to each other, one should be aware that they get the same segmentation (gold or automatic) and their difference in performance is a combination of accuracy against the human annotator and the length of correctly annotated segments, since our way of calculating agreement favors agreement on long segments more than agreement on short segments.

## 5.1 Error Analysis

There are two major sources of errors that can be identified. Firstly, the automatic segmentation can ignore an actual boundary or detect a boundary where there actually is none. The latter case is especially bad because it results in many small segments that particularly hard (if at all) to classify. Secondly, there can also be errors that originate from the classification.

A large source of error in the automatic segmentation is the coarse way we detect movements, which does not take mood into account and does not disambiguate verbs that can express a movement. For instance, *came to his mind* does of course not entail an actual movement. Similarly, the intention of returning does not necessarily mean a movement.

Endlich **kam** es ihm in den Sinn, er wollte zu seinem Vater **zurückkehren**.

Finally, it **came** to his mind that he wants to **return** to his father.

Finally, we currently cannot disambiguate whether it is a main character or a minor character that moves (something) to a different location.

Unter Andern **ging** auch einer des Weges dahin, der eine Kuh zu Markte **trieb**.  
*One of the people who was **going** along the road **drove** a cow to the market.*

## 6 Towards a Visualization of the Segmentation by Locations

Related to the investigations described in the preceding sections, some work has been dedicated in setting the bases for a possible automated visualization of the provided annotations. We focused on two aspects:

1. A representation of the scenic structure of a tale
2. a visualization of interactions between characters

### 6.1 Scenic Structure

The scenic structure of a tale can be represented as a linear graph: The individual scenes form the nodes of the graph, and two nodes are connected by an edge if one scene immediately succeeds the other. Optionally, the graph can also be labeled: Nodes are then annotated with the type of location of the scene as well as the characters involved in it, while edges are annotated with the text of the scene transition. Furthermore, the types of locations can be illustrated with clip art images (e. g. a drawing of a castle for the location type “castle”). Figure 1<sup>11</sup> shows a part of such a graph for the tale Hänsel und Gretel.

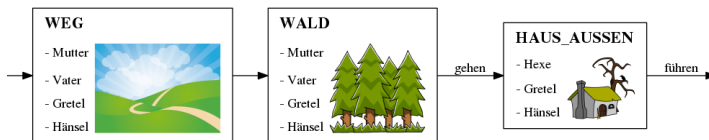


Figure 1: Representing a succession of scenes in Hänsel und Gretel, automatically generated from the annotations. Including types of locations and using clip art images for representing those.

In order to generate a graph as displayed in Figure 1, we are using a Python script that iterates over the annotated scenes of the tale and that for each scene creates a node and its labels, using the DOT graph description language. We use the advanced feature of HTML node syntax to properly arrange the various parts of the node (locations, characters, image). In a final step, code is created which links the nodes with edges to form a linear chain.

### 6.2 Interactions between Characters

The interactions between the characters in a tale can also be represented in a graph. In this case, every character in the tale is represented by exactly one node, and a edge is drawn from character A to character B if A talks to B at least once over the course of the narrative. The edge is then labeled with the number of times B is addressed by A. Additionally, nodes are

<sup>11</sup>The locations in the nodes are Path (WEG), Forest (WALD) and Outside\_House (HAUS\_AUSSEN).

positioned in such a way as to minimize the distance between characters who interact with each other more frequently. Naturally, unlike the scene graph, such a character interaction graph will in general not be linear. Figure 2 shows an example graph of this kind for the tale “Die Bremer Stadtmusikanten” (*Town Musicians of Bremen*)<sup>12</sup>

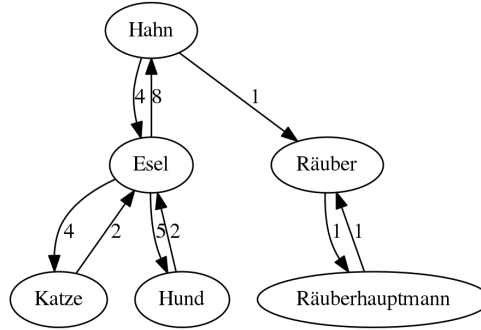


Figure 2: Representation of the interaction between characters in a tale, taking into consideration the frequency of such interactions.

As with the scene graph, we use Graphviz to create the character interaction graph. To extract the necessary information from a tale, we use a nested loop to iterate over all dialogue acts in each scene, ignoring passages spoken by the narrator. For each ordered pair (A,B) of characters, we count how often A talks to B. We then create a node for each character and link it via outgoing edges to all the nodes corresponding to characters they talk to at least once. By adjusting the weight attribute of the edges, we assure that characters that interact frequently are positioned close to each other.

Character interaction will be investigated in more details, as we assume that characters of a folktale interacting with each other are sharing a location, a feature that can improve our current algorithms for their detection.

## 7 Conclusion

We presented current work in establishing a corpus for supporting the automated classification of locations in folktales. The classification of locations can probably play a relevant role in the classification of tales along the lines of widely used classification systems for narratives. Automatic classification can help with that and provide means of finding spatial patterns in the structure of folktales. We are working on improving the currently implemented classification approaches and extending it to identifying identity of locations. We started also to apply basic algorithms for visualizing tales along their segmentation by locations. We are also aiming at adapting and integrating our annotation scheme with work proposed for example by [7].

<sup>12</sup>The characters in the nodes are a rooster (*Hahn*), a donkey (*Esel*), a cat (*Katze*), a dog (*Hund*), robbers (*Räuber*) and a robber chief (*Räuberhauptmann*).

## References

- [1] Ron Artstein and Massimo Poesio. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596, December 2008.
- [2] Vladimir Propp. *Morphology of the folktale*. Trans., Laurence Scott. 2nd ed., University of Texas Press, 1968.
- [3] Andrew Rosenberg and Ed Binkowski. Augmenting the kappa statistic to determine interannotator reliability for multiply labeled data points. In *Proceedings of HLT-NAACL 2004: Short Papers*, HLT-NAACL-Short '04, pages 77–80, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.
- [4] Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. In Daniel Jones and Harold Somers, editors, *New Methods in Language Processing*, Studies in Computational Linguistics, pages 154–164. UCL Press, London, GB, 1997.
- [5] Stith Thompson. *Motif-index of folk-literature: A classification of narrative elements in folktales, ballads, myths, fables, medieval romances, exempla, fabliaux, jest-books, and local legends*. Revised and enlarged edition, Indiana University Press, 1955–1958.
- [6] Hans-Jörg Uther. *The Types of International Folktales: A Classification and Bibliography. Based on the system of Antti Aarne and Stith Thompson*. Suomalainen Tiedekatemia, 2004.
- [7] Gabriel Viehhauser-Mery and Florian Barth. Towards a digital narratology of space. In *Digital Humanities 2017: Conference Abstracts*, Montréal, Canada, August 2017.



# What Can We Find Out about Time and Space in the ForFun Database?

Marie Mikulová, Eduard Bejček, Jarmila Panevová

Charles University,  
Faculty of Mathematics and Physics,  
Institute of Formal and Applied Linguistics  
E-mail: {mikulova, bejcek, panevova}@ufal.mff.cuni.cz

## Abstract

We present a description of time and space modifications in Czech sentences based on the ForFun database. ForFun is a new resource built on annotated corpora of Czech—Prague Dependency Treebanks—for inspecting thousands of real examples categorized by their form as well as by their deep syntactic function. Based on the database, we perform a detailed description of meanings of time and space modifications including a list of formal means with real examples coming from both written and spoken texts. It should be emphasized that the study is data-oriented rather than theory-oriented.

*“Space and time are the framework within which the mind is constrained to construct its experience of reality.” (Immanuel Kant)*

## 1 Introduction

Spatial and temporal modifications are by far the most frequently occurring modification types (cf. Table 1). While the time and space meanings (expressed usually by adverbials) belong to the language universals, their subdivision into subtle meanings is language specific. Studies of these specificities is needed both for applied tasks such as NLP as well as for the language typology (cf. the differences between the Czech *mezi* on the one side and English *between* and *among*, or Russian *meždu* and *sredi* on the other). The present study is not concerned with the philosophical or physical structure of space and time (as might be judged from the above quotation) but we concentrate here only on the forms used for the temporal and spatial meanings in Czech.<sup>1</sup>

---

<sup>1</sup>We do not deal with reconstructing the chronological order of events (as in TimeBank, see Pustejovsky et al. [14]; or Setzer – Gaizauskas [16], Katz – Arosio [6]) or with the notion of time taxis capturing all pointers for sequence of events in the text (Chrakovskij [2]).

Sentences containing:	All texts	%	Written texts	Spoken texts
Space modifications	74 164	29	43 089	31 075
Time modifications	66 503	26	42 266	24 237
Manner modifications	31 583	12	21 752	9 831
Casual modifications	26 569	10	18 022	8 547
Other modifications	50 425	20	35 967	14 458
No modification	99 564	39	60 060	39 504

Table 1: Number of sentences in ForFun with at least one free modification. Hypothetically, in a Czech text of 100 sentences, there would be 61 sentences containing a free modification (or several different modifications) and out of these sentences there would be: 29 sentences with spatial modification(s), 26 with temporal modification(s), 12 with manner modification(s), 10 with casual modification(s) and 22 with other free modification(s). For numbers of selected modifications disregarding sentences see Table 2.

The description of functions and forms of sentence units, in particular of adverbials, has a long lasting tradition in linguistics (e.g. for Czech: Šmilauer [18]; Daneš et al. [3]; Panevová et al. [13]; for English: Hasselgård [5]). Nowadays, a corpus-based approach brings about a number of challenges largely caused by the enormous range of meanings that adverbials and other modifications can convey. In the paper, we present a description of time and space modifications based on a well-developed dependency syntax theory which is known as the Functional Generative Description (FGD; see Sgall et al. [17]). The ideas of FGD were applied in the original annotation scenario of the Prague Dependency Treebanks (PDTs). Moreover, there is the Prague Database of Syntactic Forms and Functions (ForFun) now available for inspecting thousands of real examples categorized by their form as well as by their deep syntactic function. In the paper, we also give some statistics gained from ForFun since we think that an observation of frequency has an important place in a description of language because it displays linguistic choices made by speakers and writers.

## 2 Data Resources

### 2.1 Prague Dependency Treebanks

The main resource of the data for our study is the ForFun database constituted on the data of PDTs. PDTs are complex linguistically motivated treebanks with inter-linked hierarchical layers of standoff annotation. Their texts come from different sources: daily newspaper articles, Czech translations of the Wall Street Journal, rewritten dialogs and short, often vulgar segments typed into a web translator. For details about PDTs see Hajič et al. [4]. Altogether, the treebanks contain around 180,000 sentences with their morphological, syntactic and semantic annotation.

## 2.2 ForFun: Prague Database of Syntactic Forms and Functions

The ForFun database (Mikulová – Bejček [8]) draws on the complex linguistic annotation of PDTs and arranges morphological and syntactical annotation into new tool which gives a possibility to search quickly and in a user-friendly way all forms (almost 1,500 items) used in PDTs for particular function and vice versa to look up all functions (66 items) expressed by the particular forms. For any form and function, there are plenty of examples classified according to the word-class of the governing unit, and the source of text data, accompanied by the frequency in the particular corpora. Although outstanding amount of data is a great source for a linguistic study, it is not intended for statistical methods since this resource contains many interesting but rather rare phenomena. ForFun is provided as a digital open resource accessible to all scholars via the LINDAT/CLARIN repository.<sup>2</sup>

## 3 Temporal and Spatial Modifications based on ForFun

The basic semiotic relation between the function and form (terms known from Saussure's structural linguistics [15] as the relation between “signifié” and “signifiant”) is in FGD perceived as a relation between two language layers. Concerning the relation between syntactic functions and forms, we deal with the deep syntactic layer (for functions) and surface layers (for forms). The deep syntactic layer presents a rich linguistic annotation that combines syntax and semantics in the form of semantic labelling, coreference annotation, and argument structure. The types of the (deep) dependency relations (i.e. the functions) are represented by the functor attribute attached to all nodes.<sup>3</sup> The lower layers contain surface syntax and morphological annotation. Among others they contain information about the formal realizations of sentence units (parts of speech, cases, etc.) in the form of morphological tags assigned to all tokens.

In the following subsections, we describe the time and space modifications as they are treated in FGD and thus captured in PDTs and ForFun. In FGD, the repertory of temporal and spatial functors is used not only for the modifications dependent on verbs, adjectives and adverbs (i.e. of traditional adverbials, e.g. *Domy byly postaveny v minulém století*. ‘The houses were built in the last century’), but also for the modifications dependent on nouns (e.g. *domy z minulého století* ‘houses from the last century’). All these modifications (dependent on verbs, adjectives, adverbs, and nouns) with their corresponding meanings are objects of our description.<sup>4</sup>

---

<sup>2</sup><http://hdl.handle.net/11234/1-2542>

<sup>3</sup>For a full list of all functors with their description and labeling see Mikulová et al. [10].

<sup>4</sup>The distribution of temporal and spatial modifications according to word class of their parent node is shown in Table 5.

Functors for spatial meanings			112 778
LOC	where?	<i>We are in Vienna.</i>	75 210
DIR1	where from?	<i>I went from Prague.</i>	13 737
DIR2	which way?	<i>I went through Brno.</i>	1 483
DIR3	where to?	<i>I went to Vienna.</i>	22 348
Functors for temporal meanings			98 349
TWHEN	when?	<i>He arrived at five o'clock.</i>	68 781
TPAR	during what time?	<i>During our holiday not once it rained.</i>	3 040
TSIN	since when?	<i>The exposition has been open since yesterday.</i>	2 326
TTILL	till when?	<i>Till the evening I will be in Prague.</i>	3 348
TFRWH	from when?	<i>We have a lot of sweets from Christmas.</i>	1 144
TOWH	to when?	<i>He postponed the meeting to Friday.</i>	481
THO	how often?	<i>I work on that every day.</i>	7 179
TFHL	for how long?	<i>He came for a month.</i>	1 366
THL	how long?	<i>He managed to do it in a week.</i>	10 684

Table 2: Functors for temporal and spatial meanings and their raw frequency (number of examples) in the ForFun database.

### 3.1 Functions of Temporal Modifications

In the framework of FGD (and thus in PDTs and ForFun), temporal modifications establish a set of semantically differentiated functors. The individual functors differ according to which of the possible questions about time they answer; see Table 2. We concentrate here on core temporal meanings which express various temporal points or intervals on the chronological time axis.<sup>5</sup>

The basic ideas of FGD were formulated before the large language resources were available and the subdivision of the temporal (and also spatial) meanings in principle reflects the description of these types of modifications in reference grammars of Czech (Daneš et al. [3], Šmilauer [18]). However, in these handbooks the respective functions are exemplified by several examples with most typical forms. The functors thus were developed as relatively general categories and for the exhaustive description of the language, a more subtle division is needed. Therefore, in the FGD framework, it is assumed that functors would be more precisely subcategorized into the smaller units called subfunctors (cf. Panevová [12]).<sup>6</sup>

The material we have nowadays at our disposal thanks to the ForFun database brought us new stimulation for the checking of the list of functors as well as for their subcategorization into subfunctors. The functor TWHEN is proposed to be replaced by three functors. Our arguments for their introduction are connected with the necessity to follow the hierarchy between them and their corresponding sub-

<sup>5</sup>The modifications with meanings of duration and frequency (which are labelled by THO, THL, and TFHL functors) are not discussed in this paper.

<sup>6</sup>During the PDTs annotation, some subfunctors were assigned automatically (Mikulová et al. [10]). Illustrative examples of subfunctors are also given in Panevová et al. [13]

tle meanings (subfunctors). The *TWHEN* functor represents an answer for the most general question about the temporal circumstances: When something (action/state) happens/ed?, e.g., *Jan běhá v neděli*. ‘John jogs on Sunday’ but the question may be answered also by *Jan běhá před snídaní*. ‘John jogs before breakfast’ or *Jan běhá po obědě*. ‘John jogs after lunch’ or *Jan běhá během prázdnin* ‘John jogs during holidays’. The temporal modifications place the governing event on the time axis with regard to a time point expressed by the given modification: they place governing event as current, ongoing, preceding or following this time point. The different meanings (“at”, “before”, and “after the given time”) split the general functor *TWHEN* into three new functors *TAT*, *TBEFORE*, and *TAFTER*, respectively and they should be introduced in the revised version of PDTs scenario. Together with the *TPAR* functor (“during the given time”; the time period covers the state or action expressed by the governing event fully), these four functors constitute core of the temporal modifications.

	<i>TAT</i>	<i>TBEFORE</i>	<i>TAFTER</i>	<i>TPAR</i>
0	on Monday	before Monday	after Monday	during Monday
begin	at the beginning of the week	before the beginning of the week	after the beginning of the week	during the beginning of the week
end	at the end of the week	before the end of the week	after the end of the week	during the end of the week
middle	in the middle of the week	before the middle of the week	after the middle of the week	during the middle of the week
approx	around Christmas	–	–	–
between	between the games	–	–	–
assoon	–	–	as soon as I get back	–

Table 3: General arrangement of core temporal functors and their subfunctors

This conclusion is supported by other arguments: These functors can be further subcategorized into subtle meanings, such as “beginning”, “middle” or “end of the given time period”, or “approximate of the time period” (e.g., *Jan běhá začátkem léta*. ‘John jogs at the beginning of summer’, *Jan běhá uprostřed léta*. ‘John jogs at the middle of summer’, *Jan běhá koncem léta*. ‘John jogs at the end of summer’, *Jan běhá kolem druhé hodiny* ‘John jogs at about two o’clock’; see a general arrangement of temporal functors and their subfunctors in Table 3). However, the subcategorization of functors into subtle units is strictly connected with the determination of the class of secondary prepositions. The study of criteria for their determination with regard to the examples from the corpora as well as to the results proposed in the papers and monographs (e.g., Kroupová [7], Blatná [1]) is needed. Within a wide conception of secondary prepositions, the functor *TAT* is modified by the subfunctors *begin*, *middle*, *end*, *approx*, *between* (see also Table 4). However, it seems that in Czech, the new added functors *TBEFORE* and *TAFTER* and also

TPAR functor are not compatible with some proposed subfunctors; e.g., *Jan běhá \*kolem*-approx *před snídání*-TBEFORE, *Jan běhá \*uprostřed*-middle *po obědě*-TAFTER). This supports our hypothesis (mentioned above) that while the time and space meanings belong to the language universals, their subdivision into subtle meanings is language specific.

subfunctor forms	examples
0	<i>na+4</i> <i>na podzim</i> ‘in autumn’ <i>na+6</i> <i>na jaře</i> ‘in spring’ <i>v+6</i> <i>v minulém století</i> ‘in the last century’ <i>v+4</i> <i>v pátek</i> ‘on Friday’ <i>při+6</i> <i>při příležitosti narozenin</i> ‘on the occasion of the birthday’ <i>o+6</i> <i>o Velikonocích</i> ‘at Easter’ <i>u+2</i> <i>u snídani</i> ‘at breakfast’ <i>za+2</i> <i>za první světové války</i> ‘in the period of the first world war’ Accus <i>Sejdeme se příští týden.</i> ‘We shall meet next week’ Instr <i>Zákon vstupuje v platnost dnem podpisu.</i> ‘A law comes into effect by the day of signatur’ Adverb <i>Zítřka bude pršet.</i> ‘Tomorrow it will rain’ <i>když+vfin</i> <i>Když babička dovyprávěla, děti už spaly.</i> ‘By the time granny finished the tale, children were asleep’
begin	<i>začátkem+2</i> <i>začátkem sezóny</i> ‘by the beginning of the season’ <i>počátkem+2</i> <i>počátkem měsíce</i> ‘at the beginning of the month’
mid	<i>uprostřed+2</i> <i>uprostřed týdne</i> ‘in the middle of the week’
end	<i>koncem+2</i> <i>koncem roku</i> ‘by the end of the year’ <i>závěrem+2</i> <i>závěrem sezóny</i> ‘at the end of the season’
approx	<i>kolem+2</i> <i>kolem poledne</i> ‘around noon’ <i>okolo+2</i> <i>okolo druhé hodiny</i> ‘around two o’clock’
between	<i>mezi+7</i> <i>mezi dvěma válkami</i> ‘between the two wars’

Table 4: Subfunctors for TAT functor with list of most common forms for each subfunctor

The other argument in favor of splitting of TWHEN functor is the asymmetry in the distribution of the subfunctors within the generalized functor TWHEN. This is demonstrated by the hierarchy between functor and subfunctors with the different forms for subordinated verbal expressions for meanings “at”, “before”, “after”, and “during”, e.g., *Když je.*TAT *pěkné počasí*, *Jan běhá.* ‘When it is nice weather, John jogs’, *Než si zlomil.*TBEFORE *nohu*, *Jan běhal.* ‘Before John has broken his leg, he jogged’, *Poté, co si zlomil.*TAFTER *nohu*, *přestal Jan běhat.* ‘After John has broken his leg, he stopped jogging’, *Zatímco ona spala.*TPAR, *Jan běhal.* ‘While she is sleeping, John is jogging’. On the other hand, the meaning “immediately” (subfunctor assoon in Table 3) is compatible only with TAFTER functor (*Jakmile skončí přednáška*, *Jan běhá.* ‘As soon as the lecture is finished, John is jogging’).

The functors TTILL and TSIN, TOWH, TFRWH as specialized temporal meanings

also belong to this domain. The functors *TSIN* and *TTILL* represent the meaning applied as proper answer for the questions “from when” and “till when”, respectively, and they express the time point or interval in which the event either begins or ends. The functors *TOWH* and *TFRWH* do not place the event directly on the time axis, they add other temporal circumstances and they can be combined with all “*TWHEN* functors”, as well as with *TTILL* and *TSIN* functors (e.g., *Včera-TAT přeložil výuku z pátku-TFRWH na pondělí-TOWH* ‘Yesterday (he) postponed class from Friday to Monday’). Moreover, the temporal modifications can also relate directly to nouns not denoting events (e.g. *cukroví na Vánoce-TOWH* ‘sweets on Christmas’). These specialized functors can also be subcategorized into subfunctors (cf. *till the beginning of the week-TTILL-begin*, *from the middle of the week-TFRWH-middle*, etc.).

### 3.2 Functions of Spatial Modifications

Spatial modifications express either location or direction. The subdivision of their specific meanings is determined by the question specifying the location or direction as a criterion; see Table 2. Modifications assigned with *LOC* functor represent simple localization. *DIR1*, *DIR2*, and *DIR3* are used for directional meanings, they express starting point, path, and destination, respectively. Spatial modifications localize most often events or states, therefore they primarily modify verbs (see Table 5). However, they can also modify nouns and adjectives denoting events, and even nouns that do not denote events (e.g. *stůl v pokoji* ‘table in the room’; *kamínek z moře* ‘pebble from the sea’.)

Also the spatial functors need to be subcategorized into subfunctors (the different meanings of examples: *na stole* ‘on the table’, *pod stolem* ‘under the table’, etc. are covered by a single functor with the meaning “where” (*LOC*)). A proposal of subfunctors for spatial meanings was given in Mikulová et al. [9].

### 3.3 Forms of Temporal and Spatial Modifications

As we demonstrated above, temporal and spatial modifications are a heterogeneous category formally as well as semantically. In ForFun, the following formal means have been noted: adverb, prepositional case, prepositionless case, subordinate clause, and some others. As shown in Table 6,<sup>7</sup> the most frequent form by far is the prepositional case. Adverbs come second, followed by a subordinate clause. The least frequent form is a prepositionless case. The explanation for this is evident: prepositional case (including secondary prepositions) can exactly express fine-grained meanings that free modifications bear whereas a prepositionless case of noun is primarily the form for arguments. The only exception is Instrumental case for *DIR2* functor and Accusative case for *TWHEN* functor (cf. Table 6 with forms in Table 5 or Table 4). We can also observe that the forms are not equally distributed over the individual functors. For example, *DIR1*, *TSIN*, *TFRWH*, and *TOWH*

<sup>7</sup>The most frequent forms can also be observed in the Table 5 where the forms are distinguished according to the word class of the governing word.

	verb	noun	adjective	adverb
LOC	<b>69%</b> 18500× v+6 16251× Adv 8720× na+6 2536× u+2	<b>25%</b> 10983× v+6 3625× na+6 804× u+2 557× Adv	<b>3%</b> 1045× v+6 382× na+6 102× u+2 94× Adv	<b>1%</b> 184× v+6 145× Adv 130× na+6 48× u+2
DIR1	<b>34%</b> 3745× z+2 503× Adv 314× od+2 13× z+1	<b>53%</b> 6702× z+2 242× od+2 39× Adv 24× z+6	<b>9%</b> 1058× z+2 54× od+2 40× Adv	<b>1%</b> 56× od+2 17× z+2
DIR2	<b>81%</b> 409× Instr 307× po+6 253× přes+4 68× Adv	<b>15%</b> 66× po+6 62× Instr 54× přes+4 9× mezi+7	<b>2%</b> 17× Instr 5× přes+4	<b>0%</b>
DIR3	<b>81%</b> 7373× do+2 4357× Adv 2891× na+4 1741× k+3	<b>14%</b> 1605× do+2 548× na+4 369× k+3 198× Adv	<b>2%</b> 189× do+2 76× na+4 72× k+3 26× Adv	<b>1%</b> 62× do+2 30× k+3 27× Adv 25× na+4
TWHEN	<b>86%</b> 29110× Adv 9364× v+6 3339× když+vfin 3014× po+6	<b>8%</b> 1462× v+6 480× před+7 451× po+6 447× Adv	<b>3%</b> 592× Adv 500× v+6 253× Gen 135× po+6	<b>1%</b> 193× když+vfin 116× Adv 91× v+6 80× po+6
TFRWH	<b>84%</b> 252× z+2 7× od+2	<b>10%</b> 785× z+2 7× Gen	<b>3%</b> 10× z+2	<b>0%</b>
TOWH	<b>66%</b> 215× na+4 30× na+2 21× pro+4 18× do+2	<b>27%</b> 94× na+4 24× pro+4 5× na+2 5× do+2	<b>6%</b> 19× na+4	<b>0%</b>
TPAR	<b>87%</b> 902× během+2 652× při+6 288× Adv 142× v_průběhu+2	<b>9%</b> 108× při+6 87× během+2 18× v_průběhu+2 12× v+6	<b>3%</b> 36× během+2 27× při+6 5× za+4 5× za+2	<b>0%</b>
TSIN	<b>77%</b> 1637× od+2 32× Adv 20× Instr 18× ode+2	<b>15%</b> 191× od+2 132× z+2 10× ode+2	<b>6%</b> 121× od+2 6× z+2	<b>0%</b>
TTILL	<b>82%</b> 1589× do+2 762× Adv 196× dokud+vfin 97× než+vfin	<b>7%</b> 209× do+2 8× než+vfin 6× Adv 5× po+4	<b>7%</b> 157× Adv 60× do+2	<b>1%</b> 20× do+2

Table 5: Syntactic dependency of time and space modifications and the most frequent forms. (About 2% of modifications depend on technical nodes without a word class.)



Formal realization	Prepositional case	Adverb	Clause	Prepositionless case
LOC	74 %	23 %	1 %	2 %
DIR1	95 %	4 %	0 %	0 %
DIR2	58 %	5 %	0 %	36 %
DIR3	77 %	21 %	0 %	1 %
TWHEN	36 %	45 %	8 %	11 %
TERWH	98 %	0 %	1 %	1 %
TOWH	98 %	2 %	0 %	0 %
TPAR	82 %	10 %	5 %	3 %
TSIN	96 %	2 %	1 %	1 %
TTILL	61 %	28 %	10 %	1 %

Table 6: Distribution of formal realization types of time and space modifications in ForFun.

functors are almost exclusively realized by prepositional cases. For core temporal meanings (TWHEN), an adverb is a typical form. Subordinate clauses are most common for TTILL (conjunctions *dokud* ‘as long as’ or *než* ‘until’), and TWHEN (*když* ‘when’) and TPAR (*zatímco* ‘as soon as’). The variability of forms is notably among the subfunctors within one functor as we can observe in Table 4. The individual subtle meanings (except for the basic 0 subfunctor) are almost exclusively realized by limited number of forms. Prepositional cases (with secondary prepositions) dominate.

## 4 Conclusion

We presented a description of Czech time and space modifications based on the framework of the Functional Generative Description. The study is enriched by the quantitative observation gained from the ForFun database. ForFun is created for a manual processing using already annotated data of Prague Dependency Treebanks. We demonstrate that ForFun can be used for fundamental linguistic research, e.g. for a development of a fine-grained set of subtle meanings of modifications.

## Acknowledgments

The research has been supported by the project GA17-12624S of the Czech Science Foundation and by the LINDAT/CLARIN project of Ministry of Education, Youth and Sports of the Czech Republic LM2015071. The work has been using language resources developed, stored and distributed by the later project (LM2015071).

## References

- [1] Blatná R. (2006). *Víceslovné předložky v současné češtině*. Prague: Lidové noviny.
- [2] Chrakovskij V. S., ed. (2016). *Typology of Taxis Constructions*. München: LINCOM Studies in Theoretical Linguistics 58.
- [3] Daneš F. et al. (1987). *Mluvnice češtiny 3*. Prague: Academia.
- [4] Hajič J., Hajičová E., Mikulová M., Mírovský J. (2017) Prague Dependency Treebank. In *Handbook on Linguistic Annotation*. Dordrecht: Springer.
- [5] Hasselgård H. (2010). *Adjunct adverbials in English*. Cambridge University.
- [6] Katz G., and Arosio F. (2001). The Annotation of Temporal Information in Natural Language Sentences. In *ACL Workshop on Spatial and Temporal Reasoning*. Toulouse.
- [7] Kroupová L. (1985). *Sekundární předložky v současné češtině*. Prague: ČSAV.
- [8] Mikulová M. and Bejček E. (2018). ForFun 1.0: Prague Database of Syntactic Forms and Functions – An Invaluable Resource for Linguistic Research. In *Proceedings of the 11th International Conference on LREC*. Miyazaki, Japan.
- [9] Mikulová M. et al. (2017). Subcategorization of Adverbial Meanings Based On Corpus Data. *Linguistic Journal*. Vol. 68, No. 2, 268–277.
- [10] Mikulová M. et al. (2006). *Annotation on the tectogrammatical level in the Prague Dependency Treebank*. TR 2006/30, UFAL MFF UK, Prague.
- [11] Panevová J. (1980) *Formy a funkce ve stavbě české věty*. Prague: Academia.
- [12] Panevová J., Benešová E., Sgall P. (1971) *Čas a modalita v češtině*. Prague: Charles University.
- [13] Panevová J. et al. (2014). *Mluvnice současné češtiny 2*. Prague: Karolinum.
- [14] Pustejovsky, J. et al. (2006) *TimeBank 1.2*. LDC2006T08. Philadelphia: LDC.
- [15] Saussure, F. de (1916). *Cours de linguistique générale*. Paris: Payot.
- [16] Setzer A. and Gaizauskas R. (2002). On the Importance of Annotating Event-Event Temporal Relations in Text. In *Annotation Standards for Temporal Information in Natural Language, Proceedings of LREC 2002*. Gran Canaria.
- [17] Sgall P., Hajičová E., Panevová J. (1986) *The Meaning of the Sentence in Its Pragmatic Aspects*. Reidel.
- [18] Šmilauer V. (1969). *Novočeská skladba*. 2nd edition, Prague: SPN.

# On the Impact of Time Proximity on the Alignment of Spelling Variants in Old English Bibles: A Case Study

Maria Moritz

Institute of Computer Science  
University of Goettingen  
E-mail: mmoritz@etrap.eu

## Abstract

To reinforce research in historical text reuse detection. We investigate how a text is modified when it is reused in order to understand the broader context in which a reuse happened. Our long-term goal is to build a formalism behind the transformation of reuse. Previously, we investigated two datasets of Greek and Latin Biblical reuse to analyze how reuse is modified and how linguistic resources support this task. In this work, we focus on a technique to align historical variants that can not be normalized by common preprocessing techniques for Early Modern English. We use a monolingual, parallel English Bible corpus constituted of Bibles from 1500 until 1900 to investigate if time proximity can help to associate writing variants and, accordingly, can help to normalize Biblical text that is several hundred years apart from each-other. We conjecture that the writing variants between temporally close Bibles are better recognizable than between Bibles that are published with several centuries in between. We also elaborate first evaluation data that present a classification of word alignment errors and an alternative alignment procedure. This work is part of a bigger effort to apply alignment techniques that again help to investigate how historical text reuse is modified during its transmission.

## 1 Introduction

Text reuse is the repetition of text, be it literal or paraphrastic. Automated historical text reuse detection is not deeply investigated yet. To reinforce its research, we investigate how a text is modified when it is reused to understand the broader context in which a reuse happens. Our long-term goal is to build a formalism behind the transformation (modification) of reuse. Previously, we investigated two smaller datasets of Bible reuse to understand how reuse is generally modified (in terms of operations performed on word pairs) and how linguistic resources support this task [8].

In this paper, we investigate a prerequisite for the investigation of (paraphrastic) reuse modifications, namely the word alignment of our corpus that is challenged by historical writing variants. We report on work applied to a monolingual, parallel English Bible corpus constituted of Bibles from 1500 until 1900. We investigate if time proximity can help to associate writing variants and accordingly can help to normalize Biblical text easier than between text that was published several hundred years apart from each-other. Precisely, we use temporally close Bible translations to, i) investigate the spelling modifications between them, and ii) find out if the time proximity of these Bible translations can help to improve the alignment and, hence, the normalization of writing variants in old Bibles. We also showcase typical error cases that enable us to precisely improve our alignment method that we follow-up with an alternative alignment procedure. This is especially important when we need to align Bibles that are paraphrastic versions of each other. This paper is a work-in-progress report of our broader effort which to investigate how paraphrased text is modified in detail. A preparation step is to word-align our parallel Bible corpus.

## 2 Related Work

VARD is possibly the “goto” software in Early Modern English normalization: Baron and Rayson [2] present a tool that combines a known variants lookup, replacement rules and phonetic matching. Levenshtein’s distance is used to find precision and recall numbers, and an f-score is used to weight possible replacements. Yang and Eisenstein [12] investigate the application domain adaptation techniques to work with historical texts. Precisely, they apply part-of-speech tagging domain adaptation techniques to tag Early Modern English and Modern British English texts from the Penn Corpora of Historical English, and find that embedding the entire feature space outperforms word embeddings of individual words. Combined with spelling normalization they yield a 5% raw improvement in tagging accuracy on Early Modern English texts. Archer et al. [1] report on (re)training the UCREL semantic and part-of-speech (POS) analysis system to cope with Early Modern English using news texts from 1653 and 1654 totaling in 613,000 words. They introduce a rule-based component for spelling normalization and template rules to identify morphologically modified words that are ambiguous in terms of POS. They achieve correct POS-tags of about 94% when applying the system the to a held-out dataset.

## 3 Study Design

Our methodology uses a parallel monolingual corpus of Bible translations to map historical writing variants. Our goal is driven by diachronic data represented by temporally close Bibles. We investigate if time proximity of Bible editions can help

to map historical word variants to modern writing using only a simple character-distance measure.

### 3.1 Research Questions

We seek to find out: RQ1) Does the use of temporally close Bibles improve the alignment of historical writing variants?, RQ2) Whether and how does time proximity in historical texts (i.e., text that are published within short period) help to normalize old variants of text to modern spelling?, and RQ3) What are specific problems to align a historical Bible corpus?

### 3.2 Methodology

We describe briefly the steps that our experiment concludes: We word-align two time-proximate Bibles each by allowing relationships represented in terms of operations as displayed in Table 2. Our alignment is especially focussed on the explicit type of relationship (e.g., morphological modification, synonym replacement). We lemmatize the texts using MorphAdorner [3] to make sure to identify variants that the state-of-the-art can handle.

We manually evaluate ten verses pairs for each alignment to give an overview of challenges that come with aligning historical text, but more importantly, to understand why a word-relation-based alignment, i.e., by associating relationships between tow words and how they are modified or replaced (opposed to statistical alignment).

### 3.3 Data

We use a set of historical, English Bible translations that are already aligned on the verse level. We have fourteen full English Bibles available from three different resource: i) Parallel Text Project (ptp) [7], ii) Mysword (mys) [10], and iii) Bible Study Tools (bst) [11]. These range from 1500 to 1900. However, we select only a Bible subset that we think is suitable for the task. We exclude literal Bible translations—i.e., translations that follow literally the Ancient Greek and Hebrew primary source text—such as Young’s literal translation, Smiths Literal Translation and the English Septuagint by Brenton, because these Bible editions have a very diverse vocabulary. We also exclude the Darby Bible (1890), because a majority of its text was translated from other languages [6]. Table 1 lists them next to the year of publication.

The text of the upper four Bibles (TCB, MATT, GREAT and GEN) is written in Early Modern English. This means that words appear different than today (e.g., “daye”, “deuyde” instead of “day”, “divide”) or they are in old spelling (e.g., “heauen” instead of “heaven”). MorphAdorner [3], which we use for the lemmatization, is able to cover such variants only when they follow certain rules. For example “catell” (TCB) is correctly normalized to “cattle”, “kynde” (TCB) to “kind”, and “likenes”

(MATT) to “likeness”. But “lycknesse” (MATT) and “licknesse” (TCB) remain untouched. The lower five Bibles (RHE, DRC, KJV, WBT, ERV) do not contain a lot of historical writing. They contain a couple of words holding the typical archaica ending “eth”, e.g., creepeth, yieldeth, etc.

<b>Bible</b>	<b>date</b>
Matthew Bible (MATT) (mys)	1537
Great Bible (GREAT) (mys)	1539
Geneva Bible (GEN) (mys)	1560
Douay-Rheims	1582-1609
Catholic Bible (RHE) (bst)	
Douay-Rheims,	1749-1752
Challoner Revision (DRC) (mys)	
King James (KJV) (ptp)	1611-1769
The Webster Bible (WBT) (bst)	1833
English Revised Version (ERV) (mys)	1881-1894

Table 1: Overview of used Bibles

<b>operation verbose</b>	<b>operation name</b>
perfect match	NOP(word1,word2)
lower-casing matches	lower(word1,word2)
lemmatizing matches	lem(word1,word2)
short levenshtein matches	lev(word1,word2)
words are synonyms	syn(word1,word2)
word1 is hypernym of word2	hyper(word1,word2)
word1 is hyponym of word2	hypo(word1,word2)

Table 2: Transformation operations

## 4 Data Alignment

This section describes our methodology in greater detail and an assessment strategy to avoid error propagation. We show first results and discuss possible shortcomings.

### 4.1 Pairwise Bible Alignment

We align the Bibles pairwise in a chronological manner using a set of possible substitution and modification operations shown in Table 2.

We first align words of each verse in two time sequential Bibles each allowing for a set of associations (in form of operations, see Table 2 [8]) that two coupled words can have: NOP (two words are identical), lower (aka case-folding), lem (lemma of both words is identical), lev (words with an edit-distance[5] of 2/7),

syn, hyper, hypo (words are synonyms, hypernyms or hyponyms of each other). Given the lemma of a word, we use information retrieved from BabelNet[9] to identify relationships that are represented by the latter three operations. Only those resulting couples related by the operations “lev” and “lem” are considered to measure variants in our parallel Bibles especially when they cannot be normalized by the pre-processing tool we use, namely MorphAdorner.

## 4.2 Preliminary Results

We present first results of our work. In Table 3, the matching alignments that are enabled by lemmatizing words are displayed under “known lemmas”. We distinguish word types from the source and target Bible, and tokens. Under “newly found edits” we list word types from the source and the target Bible, and tokens that we can align by allowing a strict edit distance that requires a minimum length of six characters for matching word candidates. Because of its strictness, our distance measure works especially well for mapping proper names. We can align about half as many types with our measure compared to the types that can be aligned after lemmatization with MorphAdorner. Alignment between RHE and DRC, and KJV and WBT is particular similar (almost not differences between verses). This is the case because by in both cases the target Bible is a direct revision of its predecessor.

When we compare the overall alignments (union set) with the identified types and tokens between MATT (our oldest Bible) and ERV (our most recent Bible), we can align about four times as many types with the “lev” operation and about three times to twice as many word types with the “lem” operation. This tells us that we indeed can find more matches when we use the advantages of temporally close Bibles.

source Bible	target Bible	known lemmas ( <i>lem</i> )			newly found edits ( <i>lev</i> )		
		source types	target types	tokens	source types	target types	tokens
MATT	GREAT	8,595	7,939	110,779	4,683	4,508	9,795
GREAT	GEN	7,531	6,105	147,671	3,178	2,753	9,359
GEN	RHE	5,300	4,534	115,027	1,471	1,424	6,296
RHE	DRC	392	406	777	349	359	1,212
DRC	KJV	2,713	2,747	24,206	1,235	1,199	4,316
KJV	WBT	706	717	7,242	594	592	2,233
WBT	ERV	1,734	1,816	11,908	974	958	2,772
sum		16,311	15,094	417,610	10,587	9,915	35,983
MATT	ERV	8,137	5,317	181,451	2,682	2,160	8,561

Table 3: Results of types and tokens identified between two Bibles each during alignment for the operations “lem” and “lev”

A side product of our work is a dictionary with 5,803 entries that contains types of our aligned words where the key entry is chosen to be the first appearance of a word that closes an alignment chain, i.e., the word from the youngest Bible. The other variants that are stored next to the key entry are all other types of words that appear in one or more alignment chains. We generate this dictionary only based

on verses that appear in every Bible. Because we do not differ part of speech, this dictionary is not aware of mixed part of speech information in one dictionary entry. Here is one example:

- offering
  - offreth offeryng offring offereth offeringe offer offered offred offerynge offrynges offryng offerings offrynge
- require
  - requier requyre requyreth requireth requere

As we can see from the example, we cannot ensure that the leading entry in our dictionary is actually a lemma, but we still find a lot of variants that we store together in one set.

### 4.3 Error Classification

We manually evaluate ten verses from each Bible alignment pair. Table 4 shows how precise our edit distance works, how well its recall is, and how often lemmatizing enables a correct alignment. It also lists which other operations are identified, and it shows a first classification of errors that we found during the evaluation of alignments.

Bible		lem alignments		lev alignments			other operations			error types		
source	target	correct	wrong	true pos	false pos	false neg	syn	hyper	hypo	WN	PP	AUX
MATT	GREAT	32	0	2	0	3	2	1	0	3	2	0
GREAT	GEN	56	1	0	0	4	2	2	0	1	2	2
GEN	RHE	33	0	1	0	0	9	0	3	0	0	2
RHE	DRC	2	0	0	0	0	0	0	0	0	0	0
DRC	KJV	5	0	0	0	0	6	2	0	1	0	2
KJV	WBT	1	0	0	0	0	0	0	0	0	0	0
WBT	ERV	7	0	1	0	0	1	1	0	0	0	0

Table 4: Detailed list of error classes, manually evaluated between the alignment

In general, we can see that alignment by lemmatization works well with one exception of a false positive. Alignment by lev has a high precision but due to the strict condition a comparably bad recall.

We distinguish three error classes: i) WN (word net) errors, ii) PP (pre-processing) errors such as wrongly tokenized words, and iii) AUX (auxiliary) errors. The first class contains errors where two words can not be aligned with each other, simply because the synset database that we use does not store these words in the respective relations, or does not contain all of the words. The latter is the most frequent error. It appears when two auxiliary verbs are aligned, because their lemmas are identical. In many cases, however, these associations represent false couples. We list examples of each error class in Table 5.



source	swalowe	my	Selah	for	faythfulnes
target	eate	me	Sela.	forth	treuth
error class	WN error	recall error	PP error	recall error	WN error

---

source	shall	wold	eate	vp	shall
target	will	would	swallowe	-	wil
error class	AUX error	recall error	WN error	-	AUX error

Table 5: Error class examples. In the example below, it appeared that our algorithm aligned “wold” and “will”, which is wrong, and further could not align “shall/will” and “shall/wil”

The next section reports on the alignment accuracy of another experiment in which we insert a statistical alignment at the end of our pre-processing step.

#### 4.4 Statistical Alignment as Pre-processing

We align our Bibles, which are already aligned on the verse-level (by their verse identifier) on the token level. To this end, we use Berkeley Word Aligner [4], a statistical, unsupervised word aligner that was originally designed for machine translation. It combines two asymmetric alignment models based on Hidden Markov Models that are trained jointly to maximize their agreement in a combined symmetric alignment model. This mechanism especially makes the prioritized order of applying an operation as a relation between words obsolete.

Bible		lem alignments		lev alignments				other operations				error types				
source	target	correct	wrong	true	pos	false	pos	false	neg	syn	hyper	hypo	co-hypo	WN	PP	AUX
MATT	GREAT	30	0	2	0	2	2	0	0	4	0	2	0			
GREAT	GEN	53	0	0	0	3	2	0	0	2	0	2	0			
GEN	RHE	30	0	1	0	0	8	0	2	2	0	0	0			
RHE	DRC	2	0	0	0	0	0	0	0	0	0	0	0			
DRC	KJV	4	0	0	0	0	6	2	0	2	0	0	0			
KJV	WBT	1	0	0	0	0	0	0	0	0	0	0	0			
WBT	ERV	4	0	1	0	0	0	0	0	0	0	0	0			

Table 6: Detailed list of error classes, manually evaluated between the alignment with statistical pre-alignment

As Table 6 shows, we can reduce the alignment errors drastically. In fact, only pre-processing errors remain. We see minor differences in the numbers of relations. This can be attributed to the following reasons.

First, we enable co-hyponyms that we disabled in the former experiment to reduce false positives. This enabling allows words that are placed on a similar position within the sentences to be rather related as co-hyponyms than via the lev operation. E.g., “my-me” is now aligned via the co-hyponym relation whereas it was a false negative alignment of lev before (due to our minimum word length). It also compensates the WN error from the former experiment. Especially lemma, hyponym, and

hypernym relations are decreased now. This is a clear disadvantage of using a statistical pre-alignment. Word-couples such as “13:13 syn(performeth,done);” and “2:2 hypo(layd,prepared);” could not be aligned because their language model frequencies differ too much from each other. Depending on a sentence’s available alignment candidates (i.e., if the words among two sentences remain the same to a high degree) a word couple such as “12:9 lem(him,he);” is aligned or not. In our sample both happens once. Further word couples with distant positions in the sentences as “8:11 lem(he,him);”, “0:3 lem(Exalt,exalted);” are, likewise, not aligned. However, this also contributes to an accuracy increase of the local alignment. Specifically, in the former alignment experiment, often function words are aligned with each other even when they have a highly different sentence position. This sometimes causes false positives, but can be prevented by statistical pre-alignment.

In summary, we find that using the Berkeley Aligner as a pre-processing step does not yield us too many disadvantages, but assures precision with the disadvantage that words, especially those with a POS change are not aligned anymore (recall decrease). It is worth investigating if a mixed approach can combine advantages of both our approaches. I.e., to align words that can not be aligned by Berkeley Aligner with only our operation relations as a post-alignment step.

## 5 Conclusion

We reported our work-in-progress on optimizing the alignment of historical writing variants. Such alignments are relevant for analyzing the properties and modifications of text reuse—a task that is especially difficult for historical texts. In future work, we plan to improve the mechanisms used for the alignment. For example, we could combine a statistical alignment as a pre-processing step together with a post-processing step to associate words that were not aligned by the statistical alignment. The use of word family dictionaries might also help to align words with different distributions that have also different part of speeches.

## Acknowledgments

This work is funded by the German Federal Ministry of Education and Research (grant 01UG1509).

## References

- [1] Dawn Archer, Tony McEnery, Paul Rayson, and Andrew Hardie. Developing an automated semantic analysis system for early modern english. In *Corpus Linguistics 2003 conference*, pages 22–31, 2003.

- [2] Alistair Baron and Paul Rayson. Vard2: A tool for dealing with spelling variation in historical corpora. In *Postgraduate conference in corpus linguistics*, 2008.
- [3] Philip R Burns. Morphadorner v2: A java library for the morphological adornment of english language texts. <http://douglasduhaime.com/blog/cross-lingual-plagiarism-detection-with-scikit-learn>, 2013.
- [4] John DeNero and Dan Klein. Tailoring word alignments to syntactic machine translation. In *Proceedings of the Annual Meeting on Association for Computational Linguistics*, volume 45, page 17, 2007.
- [5] Vladimir I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Doklady Akademii Nauk SSSR*, 163(4):845–848, 1965. (1966) Russisch, Englische Übersetzung. In: Soviet Physics Doklady Vol. 10, No. 8: 707–710.
- [6] Michael Marlowe. John Nelson Darby’s Version. <http://www.bible-researcher.com/darby.html>. Accessed: Oct. 2017.
- [7] Thomas Mayer and Michael Cysouw. Creating a massively parallel bible corpus. In *Proceedings of LREC’14*. European Language Resources Association (ELRA), 2014.
- [8] Maria Moritz, Andreas Wiederhold, Barbara Pavlek, Yuri Bizzoni, and Marco Büchler. Non-literal text reuse in historical texts: An approach to identify reuse transformations and its application to bible reuse. In *Empirical Methods in Natural Language Processing (EMNLP’16)*, Austin, TX, USA. Association for Computational Linguistics, 2016.
- [9] Roberto Navigli and Simone Paolo Ponzetto. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artif. Intell.*, 193:217–250, December 2012.
- [10] Riversoft Systems. Mysword. [www.mysword.info/](http://www.mysword.info/), 2011–2017.
- [11] Bible Study Tools. Bible study tools. <http://www.biblestudytools.com/>, 2017.
- [12] Yi Yang and Jacob Eisenstein. Part-of-speech tagging for historical english. *CoRR*, abs/1603.03144, 2016.



# A Toolkit for lemmatising, analysing, and visualising Middle English Data

Michael Percillier

English Department  
University of Mannheim (Germany)  
E-mail: percillier@uni-mannheim.de

## Abstract

The paper describes a collection of resources and tools aimed at inserting and utilising lemma information of lexical verbs in a Middle English corpus. The resources and tools introduced are bundled in a web platform, which is also presented. The use of the platform's resources and tools is illustrated with an example query investigating the transfer of argument structure from verbs copied from Old French/Anglo-Norman to native Middle English verbs. The platform is highly relevant for any study investigating semantic verb classes in Middle English or language contact between Middle English and Old French.

## 1 Introduction

The present paper introduces a collection of resources, tools, and methods developed within a research project entitled *Borrowing of Argument Structure in Contact Situations: The Case of Medieval English under French influence* (henceforth BASICS),<sup>1</sup> which investigates the copying of argument structures alongside lexical verbs in the medieval contact situation between Anglo-Norman (henceforth AN) and Middle English (henceforth ME). The project requires analyses that focus not only on syntactic structures, but also on individual verbs and semantic verb classes. The currently available resources for ME, notably the *Penn-Parsed Corpus of Middle English* (henceforth PPCME2) [4], are designed for syntactic queries rather than lemma-based or semantic queries. To this end, lexical verbs in the PPCME2 were lemmatised, and methods for querying verb classes within this lemmatised version of the corpus were developed. The present paper describes a web platform called the *BASICS Toolkit*<sup>2</sup> which hosts these resources, tools, and methods, and provides a walkthrough using a concrete example query.

---

<sup>1</sup>The project website is accessible at <https://sites.google.com/site/dfgbasics/>.

<sup>2</sup>Accessible at <http://terrano.philosophie.uni-stuttgart.de/BASICStoolkit/>.

## 2 Verb Lemmatisation

This section briefly outlines the lemmatisation process. As a lemmatised version of the corpus cannot be provided for licensing reasons, possibilities for users to apply the procedure to their own copy of the corpus will be described. A more detailed description of the lemmatisation process is offered in [8].

### 2.1 Form-lemma Inventory

The lemmatisation of lexical verbs in the PPCME2 relies on a list of form-lemma correspondences, obtained by extracting all verb forms from the corpus and manually assigning lemmata as listed in the *Middle English Dictionary* (henceforth MED) [7]. Each entry of the MED possesses an identifying number (henceforth MED-ID), with which homonyms can be disambiguated. The contents of the form-lemma inventory can be queried in the platform under the tab LEMMA SEARCH.

### 2.2 Insertion of Lemma Information

The lemma information is inserted into the corpus by directly appending it to the verb form, so as to still comply with the *Penn-Treebank* format, whose basic structure is shown in Example (1).<sup>3</sup> Given the format of the corpus, its exploitation is tied to its related software *CorpusSearch* [10], meaning that more recent XML-based search tools such as *TigerSearch* [6] or *ANNIS* [3] cannot be used.

```
(1) ( (IP-MAT (ADVP-TMP (ADV Then))
      (NP-SBJ (D the)
               (N child))
      (VBD became)
      (ADJP (ADJR happier)
             (CONJ and)
             (ADJR happier))
      (E_S .)))
```

Each piece of inserted information is demarcated by @ characters and specified by an attribute. Verb lemmas are specified by the attribute *l*, and MED-IDs by the attribute *m* (see Example (2)). The identification of verb origin is based on a list of 2,026 English verbs copied from French between 1066 and 1500, obtained in cooperation with the *Oxford English Dictionary* (henceforth OED) [9]. For verbs occurring in this list of French-based verbs, the attribute *e* (for *etymology*) is defined as *french* (see Example (3)), whereas other verbs have this attribute set as *nonfrench*. The attribute *w* (for *warning*) indicates that the lemma was not found in the form-lemma inventory, and was matched using either spelling substitution or stemming methods (see Examples (3) and (4) respectively), or that the manual form-lemma match was deemed doubtful (see Example (5)). For verbs spelt

---

<sup>3</sup>Example adapted from <http://corpussearch.sourceforge.net/format.html>.

as multiple words, the information is appended to the final element (see Example (6)). In cases where no form-lemma correspondence could be found even after the substitution and stemming methods, the lemma and MED-ID are marked as *NA* (see Example (7)).

- (2) (VAG settyng@l=setten@m=39654@e=nonfrench@)
- (3) (VAG consyderyng@l=consideren@m=9387@e=french@w=substitution@)
- (4) (VB tellyn@l=tellen@m=44693@e=nonfrench@w=stemming@)
- (5) (VBI wilne@l=wilnen@m=52815@e=nonfrench@w=doubt@)
- (6) (VBP21 vnder) (VBP22 stont@l=understonden@m=48362@e=nonfrench@w=substitution@)
- (7) (VAN iii@l=NA@m=NA@)

### 2.3 Application of the Lemmatisation Process

As previously mentioned, a lemmatised version of the corpus cannot be provided, nor can any search function within the corpus be offered for licensing reasons. Instead, the form-lemma inventory and the *Python* script used to insert the lemma information into the corpus are available for download under the tab LEMMATISER, along with instructions, thus allowing users to lemmatise their own version of the corpus.

## 3 Resources and Tools of the Platform

The present section provides a walkthrough of the resources and tools that the platform offers for the exploitation of the lemmatised version of the PPCME2, using an analysis of prepositional objects with ME synonyms of *please* as an illustrative example. The example was chosen to mirror findings from previous studies [1][11], in which argument structures from Old French/Anglo-Norman (henceforth OF/AN) verbs, as shown in Example (8),<sup>4</sup> are said to be transferred to native ME verbs.

- (8) ces qui volent a Deu pleisir e le suen regne deservir (Purg S Pat MARIE 43)  
[12]

“Those who want to please God and merit/serve his reign.”

---

<sup>4</sup>Compare with Modern French *plaire à quelqu'un*.

### 3.1 Verb Classes

The semantic verb classes by Levin [5] group lexical verbs on a semantic basis and list possible syntactic alternations for each verb class. Under the tab VERB CLASSES, the platform offers a searchable index of Levin’s verb classes, from which a list of verbs from the desired verb class can be extracted and exported. For example, performing a search for the verb class “amuse” will return the 220 verbs listed in Levin’s class “31.1. Amuse Verbs”, a subgroup of class “31. Psych Verbs (Verbs Of Psychological State)”.

### 3.2 Reverse Lookup

As Levin’s model applies only to Present-Day English (henceforth PDE) verbs, ME equivalents need to be determined. The tool REVERSE LOOKUP provides assistance for this task by querying the MED for verb entries in whose definition the PDE verbs occur, thus providing a list of likely synonyms.

In addition to verb lists corresponding to Levin’s verb classes as generated by the VERB CLASSES tool, users can also define their own verb lists by uploading a text file with one verb per line. For example, if one wishes to focus on the single PDE verb *please* rather than the entirety of the “Amuse” group, uploading a text file consisting of the string “please” will perform an MED query that returns 44 potential ME synonyms of PDE *please*.

The tool offers downloads of the results in the HTML format, as well as a CSV table to ease the process of manual verification. Upon verification, 23 of these verbs can be considered as ME synonyms of *please* (see Table 1 for a sample).

Table 1: Sample from the verification process of potential ME synonyms of PDE *please*

MED Lemma	MED-ID	OED Lemma	OED-ID	French-based	ME Synonym
cheren	7450	cheer	31144	FALSE	FALSE
comforten	8533	comfort	36891	TRUE	TRUE
enjoien	13877	enjoy	62406	TRUE	FALSE
mirthen	27917	mirth	119118	FALSE	TRUE

Verbs such as *cheren* and *enjoien* are listed as matches because their MED entries contain the string “please”. However, the definitions containing this string are “2. [...] to be glad or pleased” for *cheren* and “1. (c) to enjoy or be pleased by (something)” for *enjoien*, which suggests that in the argument structure of these verbs, the semantic role of the experiencer corresponds to the syntactic role of the subject. This places the verbs in Levin’s class “31.2. Admire Verbs”, rather than in the class “31.1. Amuse Verbs”, in which the experiencer corresponds to the direct object, and to which *please* belongs. In spite of their semantic similarity, *cheren* and *enjoien* are not considered ME synonyms of PDE *please* due to their different



argument structures.

In contrast, verbs such as *comforten* and *mirthen* are considered ME synonyms of PDE *please* as they are a match in terms of meaning as well as argument structure. The relevant definitions are “6. To entertain or amuse (sb.); please (sb.)” for *comforten* and “(a) To comfort or console (sb.), distract (sb.) from sorrow; amuse or entertain (sb.); please (sb.)” for *mirthen*.

### 3.3 Corpus Query

In order to perform a search in the PPCME2, a query for the tool *CorpusSearch* needs to be formulated. The tool QMAKER generates such a \*.q query file based on a verb list created by the REVERSE LOOKUP tool, or any CSV table containing a column named “MED-ID” in which the MED-IDs of the lemmata to be queried are listed. Users can apply the generated query file as is, as in Example (9),<sup>5</sup> or can edit it, as shown in Example (10), in which the query is limited to verbs co-occurring with a prepositional phrase headed by *to* (including spelling variants).

- (9) node: IP\*  
 query:  
 (\*m=886@\*|\*m=1194@\*|\*m=1798@\* exists)
- (10) node: IP\*  
 query:  
 (IP\* iDoms V\*)  
 AND (V\* iDoms \*m=886@\*|\*m=1194@\*|\*m=1798@\*)  
 AND (V\* hasSister PP)  
 AND (PP iDoms P)  
 AND (P iDoms to|two|tu|te|tho|ta|tol)

The adapted query shown in Example (10) makes it easy to identify instances of prepositional objects with French-based verbs (11) and native verbs (12).

- (11) Y shal pleise **to** our Lord in þe kyngdom of þe leueand (CMEARLPS,142.6249)  
 “I shall please our Lord in the kingdom of the Levant.”
- (12) **to** zome ha wyle queme (CMAYENBI,23.342)  
 “It will please some.”

### 3.4 Graphical Representation

As previously mentioned, the PPCME2 cannot be used with XML-based tools such as *TigerSearch* and *ANNIS*, which offer graphical representation of output in addition to search functions. Other graphical tools such as *grammarscope* [2] require

<sup>5</sup>In order to save space in the display of Examples (9) and (10), the list of ME synonyms of *please* is limited to three verbs instead of displaying the entire list of 23 verbs.

yet another annotation format (Stanford Parser), and are therefore equally incompatible with the PPCME2.

In order to enable graphical representation to assist the analysis of the *Penn-Treebank* formatted output, the \*.out file generated by *CorpusSearch* can be converted to graphical syntactic trees by the PENN2SVG tool, offered by the web platform. For each result listed, a syntactic tree is generated in the SVG (*Scalable Vector Graphics*) format, preceded by a plain display of the sentence as present in the corpus, with the lexical items matching the query displayed in red. This plain display is in turn preceded by a “pretty” display of the sentence, in which escaped ME characters are shown as the actual characters, e.g. as shown in Figure 1 for  $+d \rightarrow \delta$  and  $+t \rightarrow p$ . Both sentence displays allow the user to view the lemma information of lexical verbs by hovering the cursor over underlined verbs, which shows the lemma information in white over a black background, also shown in the “pretty” display of Figure 1. The SVG trees and both sentence displays are embedded in a HTML file, so as to be viewable in any modern web browser.

hit61: CMKATHE,27.131

pe king wes swiðe icwemetlemmatised @l=iqumen@m=23304@e=nonfrench@ (CMKATHE,27.131)

+te king wes swi+de icwemet

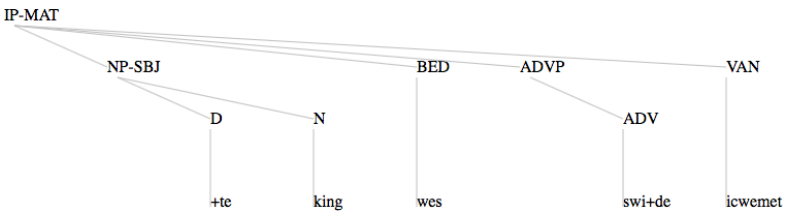


Figure 1: Example of a syntactic tree produced by PENN2SVG

4 Conclusion

The resources and tools introduced in the present paper allow for the lemmatisation of the lexical verbs of a ME corpus, and the subsequent exploitation of the newly inserted annotation in said corpus. The advantages are not limited to queries of isolated verb lemmata, but also entire verb groups, e.g. semantic verb classes, or native versus French-based verbs. As such, the platform is highly relevant for any study investigating semantic verb classes in ME or language contact between ME and OF/AN. A summary of the resources and tools presented is given in Figure 2.

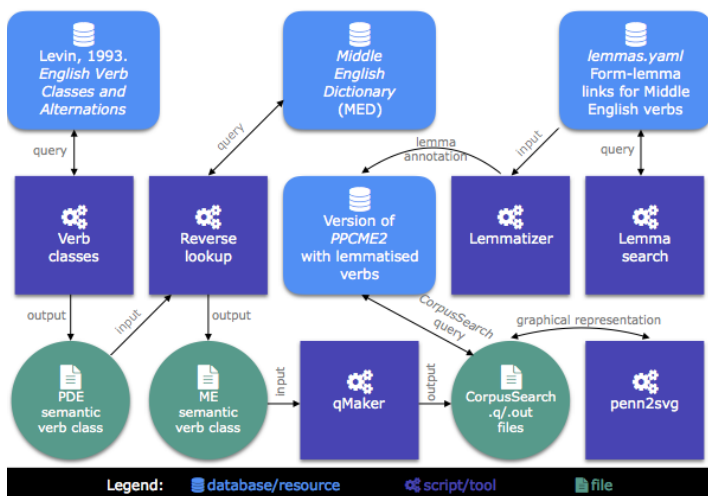


Figure 2: Summary of resources and methods available on the web platform

## 5 Outlook

A planned update of the *BASICS Toolkit* aims to simplify the application of the lemmatisation process for owners of a PPCME2 copy. In addition to the currently available option of downloading the form-lemma inventory and the *Python* script to insert the lemma information in the corpus, a second option allowing users to upload their copy of the corpus and download the lemmatised output is planned. Further, the lemmatisation of ME data in plain text is also envisaged in order to allow users to apply the process to any ME text, rather than being limited to texts featured in the PPCME2.

Besides optimising and updating existing resources and tools, a further desideratum of the *BASICS Toolkit* is to include a repository for queries and results from studies undertaken in the context of the BASICS project. The aim is not only to render the methodology transparent and results reproducible, but also to offer a dynamic database of verb valency for ME and OF/AN.

## Acknowledgements

Funding from the *Deutsche Forschungsgemeinschaft* (grant TR555/6-1) is gratefully acknowledged. I thank Carola Trips, Achim Stein, Yela Schauwecker, and Richard Ingham for their cooperation in the BASICS project. For their work on the form-lemma correspondences, I thank Lena Kaltenbach, Natascha Schultheiß, Lisa Seidel, and Jonas Stork. I would also like to thank the three anonymous reviewers

for their useful suggestions.

## References

- [1] Cynthia L. Allen. *Case Marking and Reanalysis: Grammatical Relations from Old to Early Modern English*. Oxford University Press, Oxford, 1995.
- [2] Bernard Bou. grammarscope. <http://grammarscope.sourceforge.net>, 2017.
- [3] Thomas Krause and Amir Zeldes. ANNIS3: A new architecture for generic corpus query and visualization. *Digital Scholarship in the Humanities*, 31(1):118–139, 2014.
- [4] Anthony Kroch and Ann Taylor. *The Penn-Helsinki Parsed Corpus of Middle English, Second Edition (PPCME2), release 3*. University of Pennsylvania, Philadelphia, 2000.
- [5] Beth Levin. *English verb classes and alternations: a preliminary investigation*. University of Chicago Press, Chicago, 1993.
- [6] Wolfgang Lezius, Hannes Biesinger, and Ciprian Gerstenberger. *Tigersearch manual*. Institut für Maschinelle Sprachverarbeitung (IMS), University of Stuttgart, Stuttgart, 2002.
- [7] Frances McSparran, Paul Schaffner, John Latta, Alan Pagliere, Christina Powell, and Matt Stoeffler. *Middle English Dictionary*. University of Michigan, Ann Arbor, 2001.
- [8] Michael Percillier. Verb lemmatization and semantic verb classes in a Middle English corpus. *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, pages 209–214, 2016.
- [9] Michael Proffitt, editor. *Oxford English Dictionary*. Oxford University Press, Oxford, 3<sup>rd</sup> edition, 2015.
- [10] Beth Randall. CorpusSearch. <http://corpussearch.sourceforge.net>, 2010.
- [11] Carola Trips and Achim Stein. Contact-induced changes in the argument structure of Middle English verbs on the model of Old French. *Journal of Language Contact*, 12, forthcoming.
- [12] David Trotter. *Anglo-Norman Dictionary 2 Online edition*. Modern Humanities Research Association, London, 2006.

# Word-level and higher level annotation of the Sardinian Medieval Corpus

Nicoletta Puddu and Achim Stein\*

University of Cagliari / University of Stuttgart

E-mail: n.puddu@unica.it / achim.stein@ling.uni-stuttgart.de

## Abstract

This paper is about the Sardinian Medieval Corpus (SMC), the first linguistically annotated digital resource of Medieval Sardinian. The first part presents the textual and linguistic characteristics and discusses them in the light of the problems they pose for both manual and automatic annotation. The second part describes the development of the first computational tools for the analysis of Medieval Sardinian, on the word level (lemmatization and part-of-speech tagging) and on the syntactic level (dependency parsing). It is shown how the manual and the automatic approach can be combined to build an annotated database efficiently, even for medieval texts.

## 1 Introduction

In this paper we will report the ongoing processes of coding and annotating the Sardinian Medieval Corpus (SMC). The SMC comprises different legal texts ranging from the 11th to the 15th century. This time frame starts from the first attestation of Sardinian in the 11th century as the official language of the “Giudicati”, the four independent kingdoms of Medieval Sardinia, to the arrival of the Catalan reign in Sardinia, which marked the end of the independence of the Giudicati. During this period Sardinia was characterized by a high level of plurilingualism [10, 246–248]. Latin was the language of the Church, but by the end of the 11th century it came to be used for international relations. In the 12th and 13th century Italian progressively assumed a significant role together with the increasing influence of Pisa and Genoa. Finally, from the 14th century on, when Sardinia became part of the kingdom of Aragona, Catalan progressively substituted Sardinian in official documents.

---

\*Authors’ note. The structure of the paper was developed by both authors. Nicoletta Puddu is responsible for Section 2; Section 3 describes joint work of both authors and was written by Achim Stein; Sections 1 and 4 were co-authored.

All these elements emerge, at different levels, in written documents and pose a challenge for creating an annotated corpus. Section 2 describes the compilation of the digital edition of the SMC, the choices concerning the textual XML encoding and the manual annotation at word level. In section 3 we present the applications of automatic and semi-automatic computational tools for tokenization, part-of-speech tagging and dependency parsing. Section 4 proposes some conclusions.

## 2 The Sardinian Medieval Corpus (SMC)

**The features of the SMC.** This section will concentrate on the processes of coding and annotation of the *Condaghes*, the oldest and most interesting Sardinian Medieval texts. Condaghes are documents recording acts of donation, transactions and income of churches and monasteries, often containing transcriptions of legal disputes, called *kertos*. An example of a *kertu* from the “condaghe di San Nicola di Trullas” (4, SNT, 164; cf. [7]) is given in the sentences (1) to (4). The symbol ‘=’ was inserted in the discussed examples to separate clitics in order to visualise the phenomenon (it does not occur in the manuscripts):

- (1) *osca mandaiti=mi donna Muscu et ego andai=bi=li ad Amendalas.*  
then sent=me lady Muscu and I went=there=to her to Amendalas.  
Then lady Muscu had me called and I went to her to Amendalas.
- (2) *et issa narraiti=mi ca: donnu, cussu kertu ki amus unpare,*  
and she told=me that lord this controversy that have together  
*pr'onore de Sanctu Nichola, si vos placet, campaniemusi=lu kene*  
for the honour of saint Nicolas if to you pleases settle=it without  
*iura vestra nen mea .*  
oaths yours nor mine .  
And she told me “Milord, for the honour of Saint Nicholas, let’s settle this controversy, if you like, without oaths, neither mine, nor yours.”
- (3) *et ego narai=li ca: donna, in benedictione! .*  
and I told=her that: lady in blessing  
And I told her: “Milady, bless you”
- (4) *et issa narraiti=mi: Prossa canpania date=mi su latus de Istefane Pira et*  
and she told=me for the settlement give=me the half of I. Pira and  
*levate vois sas filias: Furata intrega et anbos sos filios, et latus de Maria.*  
take you the daughters Furata complete and both the sons and half of Maria  
And she told me: “For the settlement, give me the half of Istefane Pira and take his daughters: the whole Furata and her sons and the half of Mary”

According to Schneider [16], court transcriptions are precisely those texts that most closely represent spoken language. They are typical of the Medieval period and especially important for historical linguistics as “they give access to perhaps the closest thing we have to authentic interaction between speakers from a cross section of society in the days before audio-recordings” [4, 221]. However, they cannot be considered as mere transcriptions of dialogues, since, on the one hand, they reflect literary models, especially biblical narratives [17] and, on the other

hand, as Kyto and Walker point out, “written records of a speech event are susceptible to interference—whether conscious or inadvertent—throughout the production process” [4, 221]. We will then consider our documents “faithful” in terms of Short, Semino and Wynne [18], i.e. keeping in due account context factors.

Our texts present all the difficulties common to historical corpora. Examples of the main characteristics of Medieval texts observed by Pinto [9, 268–9] follow:

1. A high degree of graphical variation. The personal name *Ytçoccor*, for instance, only in the Condaghe of San Nicola di Trullas, has eleven possible realizations, i.e. *Ytçoccor*, *Yçoccor*, *Itçoccor*, *Ythoccor*, *Ythoccor*, *Ithoccor*, *Ithoccor*, *’Çoccor*, *Itzoccor*, *Içoccor*, *Icthoccor*;
2. A limited and partial documentation. Texts are mainly legal and this determines the fact that the corpus can be considered neither representative nor balanced;
3. A high degree of linguistic heterogeneity. Code mixing with ecclesiastic Latin and vulgar Latin is common, especially in the beginnings and the closing of chartas.

In other words, Sardinian Medieval data, like all historical data, are “bad data” in Labov’s terms. In a corpus linguistics perspective, the Sardinian Medieval Corpus shares with other historical corpora some fundamental problems like limited size, lack of representativeness, lack of balance and lack of standardization [6].

**Text encoding and manual annotation.** Coding and annotating corpora for historical languages is a difficult task. Consequently, as Viana et al. [20] underline, this may often lead to the creation of text archives rather than corpora. Indeed, the available resources for Medieval Sardinian have contextual information but lack linguistic annotation [2]. Lass [5], in his seminal paper on editions, corpora and witnesshood, lists three inviolable desiderata for a proper historical corpus: (i) Maximal information preservation; (ii) No irreversible editorial intervention; (iii) Maximal flexibility. In order to comply with Lass’s desiderata, we decided to adopt the TEI P5 guidelines, one of the most widely accepted standards of annotation for historical corpora (see Puddu [12]). This choice allows us, on the one hand, to preserve the editorial choices in the text at the philological level (like variants, supplied texts or expansions of abbreviations), and on the other hand, to include contextual information in the TEI header.

The use of TEI P5 made it possible to tackle the problem of spelling variation, since we did not consider normalization as a reliable option for our corpus. We solved the problem tagging every word with a tag <w> which has two attributes: @type and @lemma. The tag @type conveys grammatical information, while the tag @lemma refers to the form chosen as reference form. In the case of the aforementioned *Yçoccor*, we chose a reference form (*Ytçoccor*) and used it as the value of the attribute @lemma. In this way, one can perform a search on the exact form *Yçoccor*, which will return only the identical graphic varieties. Moreover, one can make a search by lemma (*Ytçoccor*), which returns all the attested forms of *Ytçoccor* in its different orthographic variants. The “normalized” forms are the equivalent of

“glosses” in the philological tradition, and they facilitate corpus queries.

The use of TEI P5 also allowed us to relate contextual information to the text. We decided to mark the *kertos* in the Condaghes with specific tags, in order to identify the peculiar features which could mirror spoken language in these “conversations”. The participants in such “interactions” are identified by their name, sex and, in some cases, by their profession and their place of origin. We coded relevant contextual information in the TEI header for both text and participant description. This makes the SMC searchable by relevant sociolinguistic features which, as Vazquez and Marques Aguado [23] point out, must be kept in due consideration when studying Medieval texts. Every “participant” is identified by means of an `xml:id` which allows us to connect the participant to text sections “pronounced” by him. To identify these sections we used the tag `<q>` associated to an attribute `@who` as in the following example:

- (5) Petru de Thori sued me because «Why are you taking Sardinia away, who is mine?»  
`<p n="1">Certait mecu Petru de Thori ka.`  
`<q who="#a050">Procetiu mi la levas a Sardinia, ca es mea?</q></p>`

By doing this, we hope to relate language features with sociolinguistic and contextual factors like gender, profession and place of origin. It is particularly remarkable that we can identify female speakers, given the absence of “female voices” in historical corpora, especially in the Middle Ages.

### 3 Computational tools

**Applying tools to medieval texts.** The application of computational tools to medieval texts is notoriously difficult. In most cases, the elaboration of digital editions based on transcribed manuscripts precedes the linguistic annotation, as for example in the case of the Penn parsed corpora of historical English<sup>1</sup>. In the case of the SMC, the digitalization and annotation is work in progress, which we tried to facilitate by integrating computational tools. Consequently our focus is not a quantitative evaluation of the accuracy the tools achieve on the various levels of annotation. Rather, we show how digitalization and annotation can be linked in a productive process that leads to a faster and more coherent creation of a linguistically relevant resource, even with medieval texts.

Another typical and problematic issue of medieval corpora is the multilingualism of our corpus. As we saw before, Latin alternates with Sardinian, especially in the beginnings and the closing of chartas, and this is problematic for lemmatization, since we do not want to include Latin words in our lemma list. This contrasts with the interesting methodological challenge for the annotation on the level of syntax, where the Latin words often have to be treated on a par with native words to avoid interruptions of syntactic structures.

In the following paragraphs we discuss selected issues of tokenization, annotation at word level, and syntactic annotation.

---

<sup>1</sup><https://www.ling.upenn.edu/hist-corpora/>



**Clitics: a problem for tokenization.** Compared to the treatment of other (e.g. Germanic) languages, Romance languages present the additional problem of clitics, i.e., pronouns that attach to lexical forms (“hosts”, most often verbs), in pre-lexical (proclitic) or post-lexical (enclitic) position. Just like in Modern Italian, enclitic pronouns are attached to the host, and are thus problematic for tokenization. Although POS taggers can be trained on texts where the clitics remain attached to their hosts, this option increases the size of the tag set, since all the tags for inflected forms that can be potential hosts for clitics need to have variants for the cliticised form, in order to preserve this information. Since Medieval Sardinian has more potential clitic hosts than Modern Italian, adopting this option would result in a large tag set and decrease tagging accuracy.

Further problems arise on higher levels of annotation. For example we expect that it is harder for syntactic parsers (see section 3 below) to predict the arguments of verbs correctly if the clitic arguments are not tokenised, i.e., if they remain attached to the host. Likewise, semantic annotation (e.g. the resolution of anaphora) is also facilitated if the pronoun is a separate token.

As we said above, clitics are current in other Romance languages or varieties, e.g. in spoken French, but they are a still more characteristic feature of Medieval Sardinian. The frequency of enclitic pronouns is increased by the fact that the verb frequently occurs in sentence-initial position, from which clitics are banned according to the Tobler-Mussafia effect. Virdis [21] describes the use of resumptive clitics in the *Condaghe of San Pietro di Silki* as potential markers for non-subject arguments (direct objects, indirect objects, adverbials) that occur regularly when the lexical argument is thematic. Example (6) shows the Tobler-Mussafia effect, after the sentence-initial verb, and the first line of example (1) in section 2 above shows an enclitic cluster formed by a locative (*bi*) and a pronominal (*li*) element.

- (6) *Comporai=li a cComita de Bosobe et assos frates su saltu de serra de*  
 I.bought=to.him to Comita de Bosobe and to.the brothers the land of serra de  
*Iugale*  
 Iugale  
 ‘I bought the land in Serra de Iugale from Comita de Bosobe and his brothers’.  
 (SNDT 17)

The importance of the phenomenon for the description of Medieval Sardinian syntax, as well as our wish to provide a syntactically annotated corpus in which clitic arguments should fill a position of their own, led us to tokenise clitics separately, i.e., we detached them from their host. Clitic detachment was performed semi-automatically by using a Perl script that identifies potential host-clitic combinations in a new text. The script optionally uses a pre-defined list of potential clitic forms (previously extracted from manually annotated texts) or it starts from scratch, with an empty list. Whenever the script encounters a potential host-clitic combination, it proposes either to split the form or to override the analysis: users may either correct the predicted morpheme boundary, indicate a second boundary in the case of enclitic doubling (e.g. *deibili* → *dei-bi-li*), include a ‘memo’ code for undecidable cases, or reject the suggestion and optionally store the form in a stop

list in order to prevent future suggestions for this form.

This procedure allowed us to identify and detach the clitic forms in *Bonarcado* quickly and reliably. The link between clitic(s) and host was preserved in the attribute *rend* of the XML encoding. Example (7) shows the temporary annotation where the attribute of the clitic host has a 'log value' indicating the type of intervention, e.g. manual detachment of the last three characters *llu* (other word attributes were omitted here for the sake of readability). The temporary information can be removed later, but the value *aggl* remains, allowing users to reconstruct the original form at any time. The last word, *anchillas*, is an example of a word that was manually added to the stop word list. If encountered again, the script would silently skip it inserting an *auto-stop-cont* code.

```
(7) <w ... xml:id="w_621" rend="H:manu-yes3">partindo</w>
    <w ... xml:id="w_622" rend="aggl">llu</w>
    <w ... xml:id="w_623">ladus</w>
    <w ... xml:id="w_624">a</w>
    <w ... xml:id="w_625">pare</w>
    <w ... xml:id="w_626">cun</w>
    <w ... xml:id="w_627">clesia</w>
    <w ... xml:id="w_628">,</w>
    <w ... xml:id="w_629">cum</w>
    <w ... xml:id="w_630">serbos</w>
    <w ... xml:id="w_631">et</w>
    <w ... xml:id="w_632">cun</w>
    <w ... xml:id="w_633" rend="H:manu-stop-cont">anchillas</w>
```

**Word-level annotation.** In this paragraph we will present how tokenization, part-of-speech (POS) tagging, and lemmatization were carried out. With respect to POS tagging and lemmatization, we faced the well-known problems posed by medieval texts: rich orthographic variation, lack of authoritative sources for choosing the lemma etc. Both lemmatization and POS tagging of the SMC are ongoing, and will be refined in an iterative process each time new texts are added to the database.

No POS tag set was available for Sardinian (and even less so for Modern Sardinian) and the reflection on part-of-speech was very complex. We decided to adapt an existing tag set for Italian [15] and used it for the manual annotation. We then trained different POS taggers (TreeTagger [14] and Marmot [8]) on the *Trullas* corpus that had been manually pre-annotated (about 30.000 tokens, see [12]). Marmot was slightly more accurate, but we decided to use TreeTagger because it uses an editable lexicon and thus integrates better into an iterative training routine. The accuracy score was 94.6%. We applied this model to the *Condaghe di Santa Maria di Bonarcado*, edited by Viridis [22]. The resulting list of words unknown to the tagger were added to the tagger lexicon (3705 types, 10429 tokens). This method allowed us to perform an automatic pre-analysis of *Bonarcado* with an acceptable prediction of POS and lemmas. On smaller-sized corpora like these accuracy scores are less important given that even an analysis requiring some corrections can considerably speed up the process of annotating new texts.

**Syntactic parsing.** Since the Condaghes seem to have relatively few and rather fixed word order patterns, we hoped to be able to achieve good results by applying a method analogous to POS tagging for the syntactic analysis, using machine learning tools. Building on previous experiments with the *Syntactic Reference Corpus of Old French* (SRCMF) [11], we used a dependency parser of the *mate tools*, the joint transition-based parser (JTP) [1].

First, we performed a manual dependency analysis of a subset of 100 sentences taken from *Trullas*, where we applied the grammar model of SRCMF<sup>2</sup>, adapted to comply with the principles of Universal Dependencies<sup>3</sup>. We used the *Arborator* online tool [3] for annotating the corpus manually and also for correcting the parser output later. Then we trained the JTP parser on these 100 sentences, and applied it to 100 new sentences, taken again from *Trullas*. Evaluation on such a small test set is not meaningful, but the first results were quite encouraging. The fact that the *Trullas* corpus is rather monotonous from a syntactic point of view makes the task of the parser easier. Then we applied the trained model to the *Libellus Iudicum Turritanorum* [13], which is a late copy of a chronicle dating back presumably to the 13th c. Due to revisions in the process of its textual tradition, it shows a larger variety of word orders than *Trullas*, where more sentences are verb-initial (cf. example (6)).

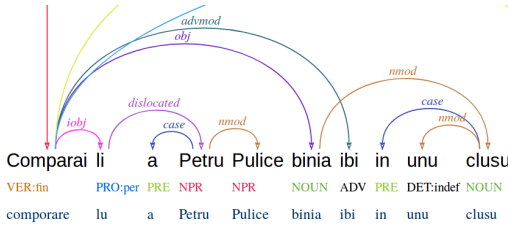
Experiments with Old French texts [19] showed that even with text types displaying strong syntactic differences (e.g. verse as opposed to prose in the SRCMF) it is possible to train a general parser model that successfully parses both types without significant decrease of accuracy. This was also applied to Sardinian. After manual correction of 53 sentences, we added these to the training corpus, re-trained the parser, and parsed *Libellus* again, with a much better result. Again, the lack of a gold standard annotation does not allow us to indicate accuracy scores, but for the treatment of Sardinian it is important that the parser copes well with both the verb-initial structures of the Condaghes and the more variable verb order of *Libellus*.

Due to their frequency, the parser also quite accurately predicts the dependencies between clitic and host, and between clitics and their lexical referents in the case of dislocations, where the clitic resumes the dislocated constituent. The partial graph in Figure 1 (example taken from *Trullas*) shows how these relations are represented in our grammar model.

---

<sup>2</sup>The SRCMF documentation is available at <http://srcmf.org>

<sup>3</sup><http://universaldependencies.org>



The verb governs the clitic *li* as an indirect object (*iobj*). The clitic in turn governs its referent, the prepositional phrase *a Petru Pulice* (*dislocated*).

(Graphical representation by Arborator [3])

Fig. 1: Dislocated elements and resumptive clitic pronouns

In order to avoid the multiplication of annotated versions of the corpus, we projected the parser output back into the original TEI corpus file. We are not aware of a TEI encoding standard for syntactic dependencies. Grammatical information as specified by the TEI guidelines only refers to morpho-syntactic features of lexical items. We represented the columns *ID HEAD* and *DEPREL* of the CoNLL shared task format as attributes (*conllid*, *head*, *deprel*) in the TEI-XML file. Example (8) shows the XML annotation for the partial sentence shown in Figure 1.

```
(8) <w ... conllid="1" head="0" deprel="root" xml:id="w_40">Comparai</w>
    <w ... conllid="2" head="1" deprel="iobj" xml:id="w_41" rend="aggl">li</w>
    <w ... conllid="3" head="4" deprel="case" xml:id="w_42">a</w>
    <w ... conllid="4" head="2" deprel="dislocated" xml:id="w_43">Petru</w>
    <w ... conllid="5" head="4" deprel="nmod" xml:id="w_44">Pulice</w>
    <w ... conllid="6" head="1" deprel="obj" xml:id="w_45">binia</w>
    <w ... conllid="7" head="1" deprel="advmod" xml:id="w_46">ibi</w>
    <w ... conllid="8" head="10" deprel="case" xml:id="w_47">in</w>
    <w ... conllid="9" head="10" deprel="nmod" xml:id="w_48">unu</w>
    <w ... conllid="10" head="6" deprel="nmod" xml:id="w_49">clusu</w>
```

The attributes *head* and *word* express the attachment of the nodes. For example, *Comparai* (*word=1*) is marked as the root node of the dependency graph (*head=0*), the clitic *li* (*word=2*) is marked as dependent of the root (*head=1*) and this relation is labelled as ‘indirect object’ (*iobj*). This kind of markup allows us to switch seamlessly between the more philological TEI markup and the CoNLL format that is needed for dependency parsing.

## 4 Conclusion

We presented the Sardinian Medieval Corpus (SMC) and showed how we faced the challenge of annotating a Medieval language with strong graphical and syntactic variation. The interplay between manual annotation (POS and lemmata) and NLP tools for word-level and syntactic annotation helps to build a medieval database in an efficient way. We created the first models for automatic part-of-speech annotation and dependency parsing of Medieval Sardinian. All the information that we add to the corpus, be it manually or automatically, is encoded in a TEI-XML compatible format.

## References

- [1] Bernd Bohnet, Joakim Nivre, Igor Boguslavsky, Richárd Farkas, Filip Ginter, and Jan Hajic. Joint morphological and syntactic analysis for richly inflected languages. In *TACL 1*, pages 415–428, 2013.
- [2] Maria Fortunato and Sara Ravani. L’informatica al servizio della filologia e della linguistica sarda: il corpus ATLiSO<sub>r</sub> (archivio testuale della lingua sarda delle origini). *Bollettino di studi sardi*, pages 53–90, 2015.
- [3] Kim Gerdes. Collaborative dependency annotation. *Dependency Linguistics*, 88, 2013.
- [4] Merja Kytö and Terry Walker. The linguistic study of early modern english speech- related texts. how ‘bad’ can ‘bad’ data be? *Journal of English Linguistics*, 31(3):221–224, 2003.
- [5] Roger Lass. ‘ut custodiant litteras’: Editions, corpora and witnesshodd. In Marina Dossena and Roger Lass, editors, *Methods and Data in English Historical Dialectology*, pages 21–50. Peter Lang, Bern, 2004.
- [6] Francesco Mambrini, Marco Passarotti, and Caroline Sporleder. Preface. 26(2):7, 2011.
- [7] Paolo Merci, editor. *Il condaghe di San Nicola di Trullas*. Delfino, Sassari, 1992.
- [8] Thomas Müller, Helmut Schmid, and Hinrich Schütze. Efficient higher-order CRFs for morphological tagging. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 322–332, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.
- [9] Immacolata Pinto. La derivazione in sardo medievale: una prima analisi in prospettiva sociolinguistica. In Piera Molinelli and Ignazio Putzu, editors, *Modelli epistemologici, metodologie della ricerca e qualità del dato. Dalla linguistica storica alla sociolinguistica storica*, pages 264–281. FrancoAngeli, Milano, 2015.
- [10] Immacolata Pinto, Paulis Giulio, and Ignazio Putzu. Morphological productivity in medieval sardinian. sociolinguistic correlates. action nouns and adverbs of manner. In Piera Molinelli, editor, *Language and identity in multi-lingual Mediterranean settings*, pages 245–268. Mouton de Gruyter, Berlin, New York, 2017.
- [11] Sophie Prévost and Achim Stein, editors. *Syntactic Reference Corpus of Medieval French (SRCMF)*. ENS de Lyon; Lattice, Paris; Universität Stuttgart, Lyon/Stuttgart, 2013. <http://srcmf.org>
- [12] Nicoletta Puddu. Costituzione del sardinian medieval corpus: prime proposte per la codifica e l’annotazione. In Piera Molinelli and Ignazio Putzu, editors, *Modelli epistemologici, metodologie della ricerca e qualità del dato. Dalla linguistica storica alla sociolinguistica storica*. FrancoAngeli, Milano, 2015.

- [13] A. Sanna. *Libellus Iudicum Turritanorum*. S'Ischiglia, Sassari, 1957.
- [14] Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. In Daniel Jones, editor, *Proceedings of the International Conference on New Methods in Language Processing (NeMLaP'94)*, Manchester September 1994, pages 44–49, Manchester, 1994. UMIST.
- [15] Helmut Schmid, Marco Baroni, Eros Zanchetta, and Achim Stein. Il sistema "treetagger arricchito" - the enriched treetagger system. In Bernardo Magnini and Amedeo Cappelletti, editors, *EVALITA 2007: Evaluation of NLP Tools for Italian*, pages 22–23, <http://www.evalita.it/2007/proceedings>, 2007.
- [16] Edgar Schneider. Investigating historical variation and change in written documents: New perspectives. In J. K. Chambers and N. Schilling, editors, *The Handbook of Language Variation and Change*, pages 57–82. Blackwell, Oxford, 2013.
- [17] Patrizia Maria Serra. Alle origini della scrittura letteraria in sardegna. In Patrizia Maria Serra, editor, *Questioni di letteratura sarda. Un paradigma da definire*, pages 19–60. FrancoAngeli, Milano, 2012.
- [18] M. H. Short, Elena Semino, and M. Wynne. Revisiting the notion of faithfulness in discourse presentation using a corpus approach. *Language and Literature*, 11(4):325–355, 2002.
- [19] Achim Stein. Parsing heterogeneous corpora with a rich dependency grammar. In Nicoletta Calzolari et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 26.-31.5.2014, Reykjavik, Iceland, 2014. European Language Resources Association (ELRA).
- [20] Vander Viana, Sonia Zyngier, and Geoff Barnbrook. Synchronic and diachronic use of corpora. In Vander Viana, Sonia Zyngier, and Geoff Barnbrook, editors, *Perspectives on Corpus linguistics*, pages 63–80. John Benjamins, Amsterdam, New York, 2011.
- [21] Maurizio Viridis. Note di sintassi sarda medievale. In Dieter Kremer and Heinz Jürgen Wolf, editors, *Studia ex hilaritate: mélanges de linguistique et d'onomastique sardes et romanes ; offerts à Monsieur Heinz Jürgen Wolf*, pages 507–526. Klincksieck, Strasbourg, 1996.
- [22] Maurizio Viridis. *Il Condaghe di Santa Maria di Bonarcado*. Centro Studi Filologici Sardi, Sassari, 2002.
- [23] Nila Vázquez and Teresa Marqués-Aguado. Editing the medieval manuscript in its social context. In Juan Manuel Hernández-Campoy and Juan Camilo Conde-Silvestre, editors, *The Handbook of Historical Sociolinguistics*, pages 123–139. John Wiley, 2012.
- [24] Sam Wolfe. The old sardinian condaghes: A syntactic study. *Transactions of the Philological Society*, 113(2):137–205, 2015.

# Marking Poetic Time: Building and Annotating a Hindi-Urdu Poetry Corpus for Computational Humanities Research

A. Sean Pue and Scott J. Nelson

Department of Linguistics and Languages  
Michigan State University  
E-mail: [pue@msu.edu](mailto:pue@msu.edu), [nelso672@msu.edu](mailto:nelso672@msu.edu)

## Abstract

This paper presents the methodology and design principles used to create and annotate a corpus of texts and performances of poetry in Hindi and Urdu, language(s) that as of yet lack a robust natural language processing toolkit. The aim of this annotation is to allow for large-scale analysis for computational humanities research. One goal of the design is to use computational techniques for annotation whenever possible. The poems are encoded in a transliteration scheme that maintains rigorous phonemic, orthographic, and morphological detail, lacking in the original Hindi (Devanagari) and Urdu (Nastaliq) scripts. That encoding is then used to facilitate the linguistic annotation of word and phonemic boundaries on acoustic recordings using a phonemic forced aligner, one originally developed for English. These boundaries are then adjusted by a human reviewer. Computational prosody detection allows the extraction of a metrical layer marking the timings of the durational metrical units of poetry in Hindi and Urdu. The resulting annotated corpus of texts and performances allows for new possibilities of algorithmic criticism, formal analysis of acoustic features across Hindi and Urdu poetries, pattern recognition over sound and text, evolutionary studies of poetic form, and enhanced data visualizations of poetry for use in computational humanities research.

## 1 Introduction

Poetry in Hindi and Urdu remains, as it has for centuries, a vibrant means of artistic, cultural, personal, and political expression. A vast textual archive of Hindi/Urdu poetry is rapidly developing on the Internet, carefully and rigorously typed in Unicode—a letter to number system, working across

fonts, that is the de facto standard of textual encoding. For computational humanities research, however, that encoding is not enough. A more rigorous form of annotation is necessary to capture phonemic, orthographic, and morphological details of Hindi-Urdu, arguably a single language that spans both multiple scripts and national boundaries. This paper presents a pilot method for developing a corpus of Hindi-Urdu poetic texts and performances for use in computational humanities research.

## 1.1 Hindi and Urdu: A Linguistic Overview

Hindi and Urdu are New Indo-Aryan languages that originated in the area of Delhi in India. Despite not being identified as separate languages until the late nineteenth century, they became, with the Partition of British India, nationally/officially separate languages, as Urdu became the primary national language of Pakistan and Hindi a primary language of India. Yet they are, in a linguistic sense, one and the same. They can be accurately described as “different *literary* styles based on the *same* linguistically defined subdialect” (Masica [4, p. 27]). Hindi is preferably written in the left-to-right Devanagari script. As a style, it involves more words that are *tatsam*, or taken from Sanskrit with minimal alteration. Urdu is best written right-to-left in the Nastaliq script, derived from that used for Persian/Farsi. As a literary style, Urdu generally involves more phrases from Persian and Arabic than Modern Standard Hindi. Devanagari texts can be transcribed into Nastaliq, and vice versa, with minor losses of information. On digital media, both are frequently written using various methods of Romanization.

## 1.2 Hindi and Urdu Prosody

While Hindi and Urdu are largely identical at the grammatical, syntactical, and phonological levels, there are distinct differences in their literary styles and especially in their poetic prosody. Both prosodies are durational rather than accentual, as the metrical units involve time rather than stress. Urdu poetry draws its understandings of metrical prosody from Persian, which was once the lingua franca of the Indian Subcontinent. That, in turn, grew out of the understandings of Arabic prosody formulated in the eighth century by Al-Ḳhalīl Ibn Aḥmad of Basra. Based on orthography, it distinguishes certain letters as “moving” (*mutaḥarrik*) or “resting” (*sākin*). The system is especially complex in its categorization, as all of the meters (*baḥr*) are described as deriving, through processes of metrical catalexis (*zaḥāfāt*), or the removal or modification of syllables, from a fixed set of “basic” *sālīm* meters. The system can alternatively be thought of as a combination of long and short metrical units, usually broken into particular feet (Pybus [7], Thiesen [11], Pritchett and Khaliq [6], Nagasaki [5]).

The prosody (*chhanda*) used for modern Hindi poetry is derived from



that used for both Sanskrit and earlier Braj Bhasha poetry. That system, also quite rigorous in its classificatory systems, has at its base an idea of heavy (*guru*) and light (*laghu*) syllables (*akshara*) in particular arrangements. Heavy syllables take longer to pronounce than light ones. These can be either in a fixed sequential pattern or in a form that requires a total weight (with heavy counting as two *mātrā* and light as one), often also involving a caesura, or pause between words in a particular location. The composition of these heavy and light syllables differs from the Perso-Arabic system used for most Urdu poetry (Kellogg [3], Snell [9], Nagasaki [5]).

## 2 Corpus Building Methodology and Design Principles

The explicit goal of this corpus of Hindi-Urdu poetry as text and performance is to allow for computational humanities research of both the texts and their acoustic realization. In addition to rigorous textual details, it therefore also requires annotation of performances at the word, phonemic, and metrical levels. An additional aim of this corpus is to permit the annotation of poetry that breaks with conventional poetic form.

### 2.1 Textual Encoding for Humanities Research and Linguistic Annotation

Though generally phonetic, neither the Devanagari preferred for Hindi nor the Nastaliq script preferred for Urdu contains sufficient information to allow for computational humanities research. Hindi, as usually spoken, does not correspond exactly to its orthographic notation. Urdu script, as typically written, does not usually mark short vowels, and the reader must therefore rely on prior knowledge and context to determine the exact pronunciation and meaning of a word.


For these reasons, computational humanities research, like natural language processing, must adopt additional forms of transliteration. The transliteration used in this project captures the orthography of texts, adds missing vowel markings and morphological boundaries, and allows for a realization of the text in Devanagari, Nastaliq, scholarly transliteration, and the International Phonetic Alphabet (IPA). The transliterated form *kiyaa* can then be realized in Devanagari as *किया*, in Nastaliq as *کِیا*, in scholarly transliteration (preferred by humanists) as *kiyā*, and in IPA, favored by linguists, as */kʲja/*.

Poetic texts are organized structurally in a way that can then be displayed using the conventions of the Textual Encoding Initiative [10] but that in their more basic input forms correspond to the principles of “tidy data”: where each variable is a column, each observation—in this case a word or

word segment—is a row, and each text is a table (Wickam [12]). The existing treebanks for Hindi-Urdu, also follow this methodology to some extent (Bhatt et al. [1]).

Producing this table usually involves three steps. First, a visual image of a book’s page is transcribed into Unicode, where each line corresponds to the lines of the printed text. Some minor information is also encoded, namely page numbers, titles, and subtitles. Stanza breaks are indicated by a blank line in the transcription. This text is then edited to add an underscore between words forming a phrase, especially in Urdu text. These phrases include future forms of words, where the suffix is written as a separate word, compound noun phrases, and also longer phrases involving the originally Persian *izāfat*, a particle equivalent to the English “of” that allows nouns to be modified by following nouns and adjectives.

Second, this annotated text is converted into a series of tables, one for each poem. The initial table consists of the following columns: “section,” “l” (for line), “seg” (for segment), “w” (for word), “text” (for original transcribed text), and “trans” (for transliteration). Section can contain either “title,” “sub” (for subtitle), or “lg” (for line group or stanza). Their presence indicate the beginning of a section, and can effectively be thought of or represented as spanning until the next section indicator. The “l” column only contains the tag “l,” indicating the start of a line. The column “seg,” for segment, contains either “phr” for phrase or “pc” for punctuation. The column “w” contains “w” to mark words. The “text” field represents the text of a particular line, or title, and also stores the text for a phrase or particular word. The “trans” column is only filled for corresponding word fields. Other columns can follow, such as notes, lemma, and sentence spans. This table can be easily translated into valid Textual Encoding Initiative XML and provides a convenient and tidy method for processing and editing.<sup>1</sup>

Adding the correct transliteration to particular phrases is a difficult process, as there are many cases of ambiguity. For example, the orthographic word  can be read as either *kiyā*, meaning “he/she/it did,” or *kyā*, meaning “what,” depending on its meaning and context. While many cases can be resolved using a dictionary lookup, some manual editing and creative use of probabilistic n-gram models to predict best transliteration is also necessary.

Once the transliteration is correctly set, an additional column representing the phonetic transcription (“ipa”) in International Phonetic Alphabet is also generated. This data will be used in acoustic annotation that follows. A sample of the resulting table is found in Table 1.

---

<sup>1</sup>The final table layout here, specifically in the use of multiline element spans, benefitted from a similar approach using GitDox introduced by Amir Zeldes at Linguistics Institute 2017. (See Zhang and Zeldes [15]).

section	l	seg	w	text	trans	ipa
title				ترقی پسند ادب		
		phr		ترقی پسند		
			w	ترقی	taraqqii-	ʈəɾəq̤q̤i
			w	پسند	pasand	pəsənd̪
		phr	w	ادب	adab	əd̪əb
lg	1			اِس کو ہاتھ لگایا ہوگا ہاتھ لگانے والے نے،		
		phr	w	اِس	is	is
		phr	w	کو	ko	ko
		phr	w	ہاتھ	haath	haʈʰ
		phr	w	لگایا	lagaayaa	ləgaʈa
		phr	w	ہوگا	hogaa	hoga
		phr	w	ہاتھ	haath	haʈʰ
		phr	w	لگانے	lagaane	ləgaːne
			w	والے	vaale	vale
		phr	w	نے	ne	ne
		pc		،	,	

Table 1: Sample table representation of the title and first line of an Urdu poem. During data entry, this table would be edited using a spreadsheet program. The first column “section” indicates whether the text is part of a “title” or “lg” (line group, or stanza). The next “l” column indicates lines; “segment” (for element) indicates whether there is a phrase (“phr”) or punctuation (“pc”); “w” marks words. The “text” column represents the transcribed text. Note that phrase and line text can be represented on the initial line to aid in transcription and also to capture any orthographic variations. The “trans” column represents transliteration of the text, capturing orthography and pronunciation. Finally, the “ipa” column, generated from the transliteration, represents phonemic transcription of the word into the International Phonetic Alphabet.

## 2.2 Acoustic Annotation

As an aim of this corpus is to supplement textual analysis, at both the orthographic and phonemic levels, with data from actual poetic performances, the poetic texts must therefore be made mapped onto audio recordings. To do so, we use the go-to software of phonetics research, Praat (Boersma [2]), invoking it using command-line scripts to be more efficient.

## 2.3 Marking Lines in Audio Files

Annotating lines in audio recordings of poetry is the first step in linking the text to the acoustic record. In order to automate it, we run a shell script (`marklines`) that loads the audio file and a Praat script that enables for the detection of pauses. The default values (minimum pitch of 70 Hz, silence threshold of -35 db, minimum silence interval of .25 seconds, and minimum sounding interval of .1 seconds) can be adjusted depending on the recording features. The script then produces a Praat TextGrid, an easily processed text file marking pause boundaries.

A human annotator then must label the lines, remove any errors (marking the segment as “cut”), and remove extraneous or add missing line breaks. These labels correspond to the line and line group labels of the text, which are generated from the tabular poem data. This notation allows also for the repetition of lines, as frequently occurs in poetic recitation. Once the labels are complete, another shell script (`cutlines`) calls a Praat script that reads the TextGrid and splits the individual lines out into separate audio files to be fed to the forced aligner, which will attempt to mark phonemic boundaries. It also reassembles a “cut” version of the file with extraneous pauses and errors removed.

## 2.4 Phoneme Detection for Hindi-Urdu Using an English-language Acoustic Model

Because Hindi-Urdu do not yet have a publicly available acoustic model, we have elected to align the Hindi-Urdu phonemes by substituting them for their closest English-language phonemes and then aligning those using the acoustic model provided by the Hidden Markov Model Toolkit (Young [13]), accessed through a modernized adaptation of the Penn Phonetics Lab Forced Aligner Toolkit (P2FA) [14].<sup>2</sup> This process is done using a shell script (`alignlines`). It maps the Hindi/Urdu phonemes to their closest approximate in the English phonemic system and then proceeds to generate a dictionary and run the forced aligner on the line-level recording segments.

---

<sup>2</sup>We are grateful to our colleague Karthik Durvasala for suggesting this approach.

## 2.5 Adjusting Phonemic Boundaries

A human annotator then adjusts the phonemic boundaries to correctly match the text. This “true gold” data can then be used, once sufficiently collected, to train the phoneme mapping to best match the English phonemes to the Hindi-Urdu ones. A feedback loop is also opened at this stage to allow for phonemic corrections. A shell script (`checkalignments`) loads the cut files and TextGrids into Praat and assists the annotator in navigating through the files. These text grids are then added to the cut audio file TextGrid by calling the shell script `reassemblelines`.

## 2.6 Computationally Adding Word-Level Annotation to Acoustic Files

Once the phoneme-level annotations have been corrected and any anomalies located, a computer program (`markwords`) adds an additional annotation level to the acoustic records that marks the word boundaries. It does so by simultaneously consuming marked phonemes in the audio file and the tabular text data to mark the interval of words in the TextGrid.

## 2.7 Computationally Adding Metrical Layers to Acoustic and Text Files

With the phonemic boundaries properly mapped, a metrical layer of annotation is also added. This process involves using a computer program (`markmetrical`) to distinguish the meter of the original text and, in reference to poem-level metadata, to map the original poem phonemes to the appropriate metrical unit. These units are durational, and they indicate phonemes that may span syllabic boundaries. If appropriate, metrical feet are also noted at this stage.

## 2.8 Final Audio Annotation

The final annotation of the cut audio files thus includes the following tiers: line, word/word-segment, phonemes, metrical units, and (optionally) metrical feet. A sample is depicted in Figure 1.

# 3 Applications in Computational Humanities Research

The corpus annotation procedures described above result in a corpus of texts and multiple performances annotated for computational humanities research. Texts are mapped to performances, allowing for research that involves both textual and auditory signals. The corpus aspires to be the

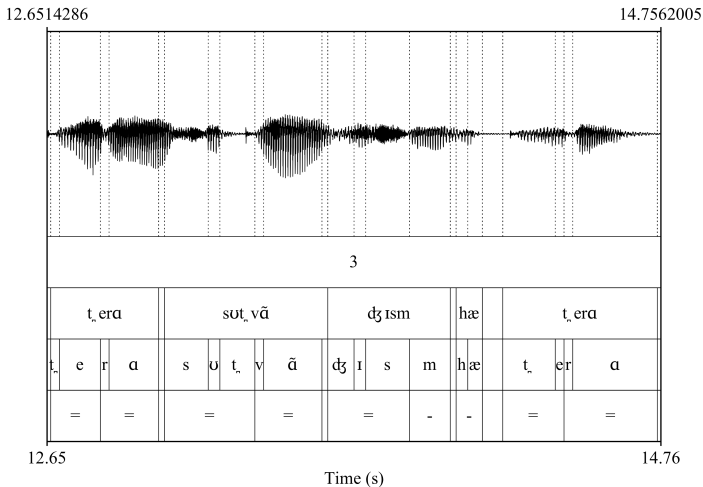


Figure 1: Sample audio annotation. The annotation tiers beneath the waveform are: line, word (represented in International Phonetic Alphabet), phonemes in IPA, and metrical units (where = is long and - is short).

building block for future computational research across the lingual divide between Hindi and Urdu, allowing for algorithmic criticism across languages (Ramsay [8]). A careful attention to both literary form and acoustic realization also enables pattern recognition over sound and text, evolutionary studies of poetic form, and enhanced data visualizations of poetry.

## 4 Acknowledgements

This research was supported by an Andrew W. Mellon Foundation New Directions Fellowship (Grant Number 11600613), awarded to the first author, and by matching funds provided by the College of Arts and Letters, Michigan State University.

## References

- [1] Rajesh Bhatt, Bhuvana Narasimhan, Martha Palmer, Owen Rambow, Dipti Misra Sharma, and Fei Xia. A multi-representational and multi-layered treebank for hindi/urdu. In *Proceedings of the Third Linguistic Annotation Workshop, ACL-IJCNLP '09*, pages 186–189, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

- [2] Paul Boersma. Praat, a system for doing phonetics by computer. *Glott International*, 5(9/10):341–345, 2001.
- [3] S. H Kellogg. *A Grammar of the Hindi Language; in which are treated the Standard Hindí, Braj, and the Eastern Hindí of the Rámáyan of Tulsí Dás, also the Colloquial Dialects of Marwar, Kumaon, Avadh, Baghelkand, Bhojpur, etc., with copious Philological Notes: Printed at the Am. Pres. Mission Press, Allahabad.* (Thacker, Spink & Co.), (Trübner), Calcutta, London, 1876. OCLC: 162639645.
- [4] Colin P. Masica. *The Indo-Aryan languages*. Cambridge language surveys; Variation: Cambridge language surveys. Cambridge University Press, Cambridge ; New York, 1991.
- [5] Hiroko Nagasaki and Ronald I. Kim, editors. *Indian and Persian prosody and recitation*. Saujanya Publications, Delhi, 2012. OCLC: 794973852.
- [6] Frances W Pritchett and Khaliq Ahmad Khaliq. *Urdu meter: a practical handbook*. South Asian Studies, University of Wisconsin, Madison, 1987.
- [7] G. D. Pybus. *A text-book of Urdu prosody and rhetoric*. Rama Krishna, Lahore, 1924.
- [8] Stephen Ramsay. *Reading machines: toward an algorithmic criticism*. University of Illinois Press ; Combined Academic [distributor, Urbana, Ill.; Chesham, 2012. OCLC: 756279615.
- [9] Rupert Snell. *The Hindi classical tradition: a Braj Bhāṣā reader*. School of Oriental and African studies, London, 1991. OCLC: 24794163.
- [10] TEI Consortium, Lou Burnard, and Syd Bauman. *P5: guidelines for electronic text encoding and interchange*. Text Encoding Initiative, Charlottesville, VA., 2008. OCLC: 244394645.
- [11] Finn Thiesen. *A manual of classical Persian prosody: with chapters on Urdu, Karakhanidic, and Ottoman prosody*. O. Harrassowitz, Wiesbaden, 1982.
- [12] Hadley Wickham. Tidy data. *The Journal of Statistical Software*, 59, 2014.
- [13] S.J. Young. The HTK Hidden Markov Model Toolkit: Design and philosophy. *Entropic Cambridge Research Laboratory, Ltd*, 2:2–44, 1994.
- [14] Jiahong Yuan and Mark Liberman. Speaker identification on the SCOTUS corpus. *The Journal of the Acoustical Society of America*, 123(5):3878–3878, 2008.

- [15] Shuo Zhang and Amir Zeldes. Gitdox: A linked version controlled online XML editor for manuscript transcription. In Vasile Rus and Zdravko Markov, editors, *Proceedings of the Thirtieth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2017, Marco Island, Florida, USA, May 22-24, 2017.*, pages 619–623. AAAI Press, 2017.



# Who *came riding* first? Le chevalier or the knight?

## A multiple corpus analysis investigating historical language contact

Yela Schauwecker and Carola Trips

University of Stuttgart, University of Mannheim

yela.schauwecker@ling.uni-stuttgart.de, ctrips@mail.uni-mannheim.de

### Abstract

This article investigates contact-induced change in the historical contact situation between Old French/Anglo French and Middle English and demonstrates how annotated corpora, digital resources, and new tools especially designed for historical linguistics, provide an empirical basis to gain new insights into this type of change which also have a bearing on linguistic theory. The phenomenon under scrutiny is the direct motion construction which, according to Talmy (1985), is typologically constrained: in verb-framed languages like (older stages of) French the Goal of Motion cannot be expressed syntactically in the same way as in satellite-framed languages like (older stages of) English. We show that Anglo French, the contact variety which developed between 1066 and 1500 on the British island, exhibits properties of a satellite-framed language that we interpret as instances of contact-induced change. Further, we show that linguistic influence may have been reciprocal, i.e. the direct motion construction in Middle English gained “French properties”. We discuss our findings in the context of theories of contact and change, and in the light of findings from research in language acquisition.

Between 1066 and 1500 Medieval England experienced a situation of intense language contact. Knowledge of Old French (OF), the prestigious foreign model code [11] which had been originally brought to England by William I and his army, rapidly spread, so that by around 1250 the population of Medieval England had changed from adult second language (L2) learners to child L2-learners and even simultaneous bilinguals [10]<sup>1</sup>. Furthermore, the contact variety Anglo-French (AF) arose which had the status of a language used in speech and writing [10]. Ingham (2012b) has shown that AF was used as vehicle language to learn Latin in so-called ‘song’ schools which shows that considerable competence of AF must have been normal for those who received an education. Bilingual code switching in accounts, lay subsidy rolls and leases has provided empirical support for a picture of effective

---

<sup>1</sup>Note that Ingham only talks about bilinguals and does not make the more fine-grained distinction made here by the authors of the article.

spoken language competence not only in education but also in French in day-to-day professional contexts, on the part of mother-tongue English speakers at this time ([9]).

Recent studies in the field of language acquisition have shown that in such bilingual acquisition scenarios, transfer of (syntactic) structures is likely to occur not only from the speakers' first language (L1) to their L2, but also from their L2 to their L1 (see e.g. [7]). In this study, we want to shed light on one such complex case of reciprocal linguistic influence which has hitherto not been discussed at all in the literature, namely that of the directed motion construction in Middle English (ME) and AF, the two languages under scrutiny in the historical contact situation mentioned above.

The directed motion construction is a perfect test case for acquisition and change because its availability is constrained by typological boundaries [16]. In the Germanic languages such as English, the Goal of Motion is always expressed outside the verb by adverbials and particles: a "satellite" in Talmy's terms [17, 486]. Languages, which exhibit such satellites (bound affixes, free prepositions, adverbs, etc.), are thus called "satellite-framed" (s-framed). In contrast, in verb-framed (v-framed) languages which do not exhibit "satellites", the Goal of Motion is expressed by other syntactic means, i.e. adverbial subclauses (e.g. gerundive constructions). Most of the Romance languages such as French are v-framed.

Thus, in an English sentence like (1a.) the Goal of Motion is expressed by the PP *to Paris* (1a.'). In French, the same construction can only have a locative reading ('the knight rides around in Paris').<sup>2</sup> In order to gain a directed motion reading, French resorts to the kind of constructions exemplified in (1b.')

- (1) a. *The knight rode to Paris.*  
 a.' *\*Le chevalier chevaucha à Paris.*  
 b. *The knight came to Paris by riding.*  
 b.' *Le chevalier entra Paris en chevauchant.*  
 c. *The knight came riding to Paris.*  
 c.' *Le chevalier allait à Paris en chevauchant.*

It becomes clear from these examples that, in general, v-framed constructions such as (1b.) and (1c.) are not excluded from s-framed languages. The Talmyan classification refers to the preferred syntactic option as it is used in everyday language and over a wide range of semantic notions [16, 62]. In other words, s-framed languages have a larger set of constructions that express directed motion, which implies that v-framed constructions are a subset of the constructions available in s-framed languages. This observation is crucial in the context of our investigation since bilinguals, on the one hand, tend to prefer constructions that are licensed in both of their languages [6], and, on the other, selective attention blinds the acquisition system to some degree to aspects in which the L2-sample exceeds the

<sup>2</sup>The reason for this is that French does not have simple lexical elements that convey Path ([21]). *Jusqu'à* has been analyzed as part of a complex prepositional phrase by [23].

L1-system [5, 92]. By taking into account research on acquisition and language typology, we predict to find that adult L2-speakers of AF

- a) take over some s-framed patterns from their dominant language into their v-framed second language even though they would not be included in the continental OF inventory.
- b) significantly increase their use of otherwise dispreferred v-framed constructions in their dominant language (ME) under the influence of their v-framed second language.

Our investigation provides new insights into the field of historical linguistics and language contact by

- a) using digital resources for historical corpora, syntactically annotated corpora and newly created tools and annotations<sup>3</sup>
- b) dealing with a contact variety that has not gained a lot of attention
- c) applying models of linguistic theory and results from language acquisition to contact-induced historical change

## 1 Directed motion in Old French and in Anglo-French

This section gives a brief overview of the constructions expressing directed motion in AF and OF. Further, it investigates the difference between OF and AF gerundive-constructions in more detail across verbs (prediction a)).

### 1.1 Directed motion in OF

Since French has been classified as a typically v-framed language, we do not expect to find s-framed patterns. If we do, these exceptions need to be explained. In continental OF, the expression of directed motion essentially relies on v-framed patterns. The thirteen s-framed examples that have been listed on various occasions in the literature (see [20, 111f]; [3, 42f]<sup>4</sup>), which seem to contradict this observation, date in fact from 1333 to the beginning of the sixteenth century, that is, from the Middle French (MF) period.<sup>5</sup> A sample-search across continental texts in the *Base de Français Médiéval* database (BFM, [2]) for *chevauchier* revealed virtually no further examples from OF. Instead, the Goal of Motion is most frequently expressed using either a durative *tant ... que* sub-clause as in (2), or a gerundive-construction as in (3):

---

<sup>3</sup>The BASICS tool kit, cf. <http://terrano.philosophie.uni-stuttgart.de/BASICS toolkit/>.

<sup>4</sup>From these, we would have to exclude 1400-1410, Baye, Journal 48: *qui dansoient par la ville* and 1369, Machaut, *Prise d'Alexandrie*, 107: *Il monta tantost à cheval, [...] Et chevaucha dedens la ville*, because contexts prove that they describe atelic movement.

<sup>5</sup>Researchers rely on lexical and syntactic criteria in order to establish the period division between Old and Middle French, which they conventionally situate at around 1350 (Tobler-Lommatzsch and Dictionnaire étymologique de l'ancien français).

- (2) **Tant** ont chevauchié ... **Qu'il** sont a Valedon venu.  
 so much have ridden ... that they are to Valedon come  
 'They rode until they gained Valedon.' (ca. 1200, Renard de Beaujeu)  
 (ca. 1200, Renard de Beaujeu)
- (3) **Arondel** y **va** **chevauchant**  
 Arondel there goes riding  
 'Arondel rides there' (pic., 1st q. c13, Galeran)  
 (pic., 1st q. c13, Galeran)

Overall, the OF data from the BFM confirm typological predictions as far as the absence of the s-framed type of Goal of Motion construction as well as the use of the v-framed construction is concerned. Things change, however, when we look at AF data.

## 1.2 Directed motion in AF

Based on the assumption that AF is a contact variety of a s-framed Germanic and a v-framed Romance language, hypothesis a) predicts more s-framed patterns in AF than in OF. This is indeed borne out, as we find significantly more Goal of Motion constructions in AF than in continental OF: for our sample-verb *chevauchier* we find five unambiguously resultative constructions with *a* and *sus* as in (4) in the BFM, in five AF texts from the second third of the twelfth century to the middle of the thirteenth century as in (5) and (6):

- (4) *Si vus a Leircestre peussiez chevalchier*  
 If you to Leicester could ride  
 'Could you ride to Leicester (agn., ca. 1174 Fantosme)  
 (agn., ca. 1174 Fantosme)
- (5) *Puis prist un batel et nagea au Roi de France*  
 Then took[he] a boat and swam to the King of France  
 'He then took a boat and sailed over to the King of France'  
 (c1245, Life of Edward the Confessor, anh)
- (6) *Tost cort a l'us u a fenestre;*  
 soon run[he] to the door or to window  
 'She keeps running from door to window'  
 (mid c12, Proverbes Salomon, anh)

Five occurrences of a total of 117 AF hits for *chevauchier* amount to 4.2 % and are contrasted with one occurrence of a total of 444 hits in continental texts (0.2 %). While the limits of this article do not allow us to give detailed analyses for the other manner-of-motion verbs as well, a quick search in the Tobler-Lommatsch (TL) dictionary [19] and in the Anglo-Norman Hub database (ANH [1]) revealed additional examples with unambiguously resultative goal-constituents for the verbs previously cited in the literature [3], [20]: two with *voler* (mid twelfth century and late thirteenth century), two with *courir*<sup>6</sup>(mid twelfth century and 1174), two

<sup>6</sup>The verb seems to have been used largely as a synonym for *aller*, and pronominal *i* does repeatedly occur with it on both sides of the Channel, but full fledged PPs like the two we refer to here are comparatively rare.

with *sauter* (ca. 1190, beginning of the fourteenth century), one with *navier* (first quarter of the twelfth century), and one with *nager* (1174). All of them are AF, and they all antedate the Middle French examples previously known in the literature.

On the other hand, the durative sub-clause given in (2), which is found with 4 % of the occurrences of *chevauchier* in OF, does not occur with *chevauchier* in the AF data at all. Thus, the data so far confirm that AF differs from OF in the expression of directed motion in the sense that AF is considerably more tolerant towards s-framed constructions than OF. Moreover, by their quantitative and chronological distribution, these results very clearly indicate an AF origin of this construction, even if we do not yet see clearly why and how it was subsequently taken over into Middle French<sup>7</sup>.

The contrast between OF and AF becomes even more obvious once we look at the gerundive constructions which are at the center of this contribution. As to *chevauchier*-gerunds, even bare numbers indicate a significant difference between AF and OF: in OF there are 18 gerundives of a total of 444 occurrences of *chevauchier* (4%), compared to 18 of a total of 117 AF occurrences of *chevauchier* (15.4%). Moreover, OF *chevauchier*-gerunds are built with *aller* in ten out of thirteen cases (76%), whereas in AF *aller* seems to be dispreferred (3 out of 13, 23%). These findings suggest that additional factors are at play. For this reason we took a closer look at the gerundive construction.

Across verbs we find a significant qualitative difference between AF and OF *venir*-gerundives. In AF, in contrast to OF,

1. *venir* seems to be largely desemantized
2. *venir* gerundives are extended to achievement verbs
3. *venir* gerunds can be used to convey a final sense

In OF, *venir* gerunds seem to be restricted to the modal case, as 78% of *venir*-gerundives take Manner of Motion-gerundives. This suggests that *venir* has maintained its original Motion semantics in this construction. 78% in OF compare to only 63% of modal gerundives in AF. Additional evidence for AF *venir* being more desemantized in this construction is provided by the *Life of Saint Francis* (from 1275), where the *venir* is combined with the static verb *agenoier* ‘to kneel’. We found no comparable case in the OF material.

- (7) *Devant les pez le roy vint agenoillant*  
 In front of the feet [of] the King came[he] kneeling  
 ‘To the king’s feet he came down kneeling’  
 (1275, stfrancis, 5)

Moreover, as seen above, in AF as well as in OF the gerundive construction is used with Manner-of-Motion verbs in order to introduce a telicizing Goal-constituent. But crucially, in AF it is also used with punctual accomplishment verbs such as *acoster* ‘to bring in (to)’, *atteindre* ‘to reach (sb.)’, *acontrer/encontrer* ‘to meet (sb.)’, *retourner* ‘to turn around’, *survenir* ‘to stumble upon (sb.)’. The only

<sup>7</sup>We appeal in this context to the “precocity of the Anglo-Norman literature” ([14] and [13]).

author who applies *venir* as an aspectual modifier of a given verb is Chrestien de Troyes who uses it with *atteindre* in two instances in two of his texts.

Finally, there are three instances where the *venir* gerundive is used to convey a final meaning, instead of the infinitive construction we would typically expect in OF. This is a strong indicator in favour of influence from ME, since only English but not OF can use the participle to express a final meaning. It is probably an instance of global copying, in the sense that the AF final gerundives reflect the ambiguity of the English *-ing*-form (modal vs. final, see below).

- (8) a. *Gudmod vint maneçant kar fel iert e tiran.*  
 Gudmod(Akk.) came [to] threaten because mean was he and [a] tyrant  
 ‘He came to threaten Gudmod, because he was a mean and tyrannical man’  
 (ca1170, Horn)
- b. *Si me mand cum ad nun e quei il vient querant*  
 So to me confide how has name and what he comes [to] ask  
 ‘So tell me what your name is and what you came to ask me for’  
 (ca1170, Horn)
- c. *Iceaux del front vindrent envaissant*  
 These from the front came [to] attack  
 ‘They came to attack at the front side’  
 (4th q. c12, Thomas de Kent)

## 2 The *come riding* construction in the history of English

This section briefly discusses possible s-framed and v-framed constructions in Old English (OE) and Middle English (ME) which were the systems that the OF speakers got in contact with. Further, we will investigate assumption b), which are expected frequency effects in ME in terms of the dispreferred v-framed construction *come + riding*.

### 2.1 Old English

Since OE is a Germanic language, we expect to find s-framed constructions where Manner or Cause are conflated in the Motion verb and Path is expressed by an adverbial. This prediction is borne out (see example (9)), it is the unmarked (most frequent) case<sup>8</sup>.

- (9) *Ða rad se æþeling Eadmund to Norðhymbran to Vhtrede eorl.*  
 then rode the atheling Edmund to Northumbria to Uhtred eorl  
 ‘Then the nobleman Edmund rode to Earl Uhtred to Northumbria.’  
 (ChronE\_[Plummer]:1016.17.1949)

<sup>8</sup>In our small-scale study of *ridan* this construction occurs 140 (90%) times of a total of 155 cases. 15 instances (9.7%) show the present participle *ridende* of which 5 instances (33.3%) show the *cuman + ridende* construction.

Two other constructions are possible: (i) the Motion verb *cuman* is followed by a bare infinitive expressing the Manner of Motion of the finite verb (see (10a.); (ii) *cuman* is followed by a present participle expressing the Manner of Motion of the finite verb (see (10b.)), cf. [4], [15].

- (10) a. *þa com þærto ridan sum cristen man sona ...*  
 then came thereto ride some Christian man soon ...  
 ‘Then soon some Christian man came riding to this place ...’  
 (ÆLS\_[Maurice]:90.5734)
- b. & *þær com þa fleogende Godes engel scinende swa swa sunne.*  
 and there came then flying God’s angel shining like the sun  
 ‘and there God’s angel came flying shining like the sun.’  
 (cathom1,ÆCHom\_I,\_31:445.177.6221)

In the third volume of [22] the construction *cuman* + *present participle* is found among the constructions of slight subordination with verbs of inchoation (§1793 (d) verbs of motion). Visser notes that it occurs very frequently in all stages of English and he provides a number of Old English (OE) and Middle English (ME) examples. By taking a closer look at the examples extracted from the *York-Toronto-Helsinki Parsed Corpus of Old English Prose* (YCOE), however, we found that (i) this construction does not occur very frequently (9.7% ), (ii) it predominantly occurs with an interpretation that is not the interpretation of the AF v-framed construction *venir* + *participle*.

Concerning the latter observation, two types of construction exist: one type where the participle does not modify the Motion verb (final), and one type where it does (modal). The meaning of these two constructions can be paraphrased as ‘come in order to do sth.’ and ‘come by doing sth.’, respectively.

- (11) a. *and siððan þæs on mergen com to Basilie biddende fulluhtes.*  
 ‘and afterwards in the morning (he) came to Basil to ask to be baptised.’ (ÆLS\_[Basil]:163.560)
- b. *him com ða ridende to sum arwurðe ridda sittende on snawhwitum horse.*  
 ‘Some honourable horseman came riding to him sitting on a snow-white horse.’ (ÆCHom\_II\_10:82.30.1631)

The example in (11a.) does not mean that the Agent of the sentence comes by baptising but that the Agent comes in order to ask for being baptised. In (11b.) and (11c.), however, this is exactly the meaning that is conveyed: the horseman comes by riding (not in order to ride).

The final type ‘come in order to do sth.’ has a token frequency of 33 and a type frequency of 22. The modal type ‘come by doing sth.’ has a token frequency of 26 and a type frequency of 10. We can conclude that in OE the *cuman* + *present participle* construction is not very frequent and it is more likely to occur with a participle that does not modify the Motion verb *come*.

## 2.2 Middle English

Exploiting the lemmatised version of the *Penn-Helsinki Parsed Corpus of Middle English* (PPCME2 [12]) qualitative and quantitative changes can be observed: (i) the construction with a bare infinitive is basically lost; (ii) the construction where the Motion verb *comen* is modified by a following verb of Manner of Motion is increasingly found (47 occurrences of this type vs. 28 occurrences of the type ‘come to do sth.’, overall 4.1% in the whole corpus). Relating these observations to OF and AF, the loss of the bare infinitive may be connected to contact with French since this type of construction did not exist (i.e. we may assume a priming effect towards other possibilities; but also note that at this time the *to*-infinitive develops in ME, see [15]). Concerning the construction *comen* + *Manner of Motion verb* we saw that although it occurs both in OF and AF, it is only in AF that it occurs with a wider range of verbs, also denoting accomplishment. This is exactly what we find in ME. In our small-case study of constructions with the Manner of Motion verb *riden* 26 occurrences of the present participle were found 17 instances (65.4%) of which occurred in the *comen* + *riding* construction. Here an increase from 33.3% in OE can be observed. Interestingly, about a fourth of these instances can be attributed to one text, *Malory's Morte Darthur* (1469) which is based on OF/AF sources.

- (12) *Thenne afore hym he sawe come rydyng oute of a castel a knyght,*  
‘Then before him he saw how a knight came riding out of a castle.’  
(MALORY,68.2300)

In recent studies a number of authors have investigated interference effects in contexts of language contact through translation [18],[8]. Since translators activate their competences in two languages, this process can be seen as a specific case of language contact. Haerberli (to appear) shows in two studies of the placement of object pronouns in ME texts and their OF/AF bases that statistical interference effects occur which may even lead to syntactic innovation. In our case, we may assume that the ME writers increasingly used the (modal) *come* + *verb of Motion* construction because (i) they recast the OF/AF text, and (ii) the construction was already one option to express Motion in ME. It is likely that these writers also used this construction even in non-translated texts more frequently because of syntactic priming effects known from language acquisition [7]. In this way, OF/AF may have had an impact on the frequency of this v-framed construction.

## 3 Conclusion

Concerning our findings, we can say that when speakers/writers of French got in contact with ME, a language that exhibited v-framed constructions as part of their s-framed grammar (OE and ME), they came across a construction that they were familiar with from their French inventory, namely the gerundive construction. What the speakers/writers added to their grammar was a new way to express Motion, i.e.



the construction with *venir*. Speakers/writers of continental OF were less exposed to ME and therefore they did not develop these constructions in the same way. In ME qualitative and quantitative effects in favour of the AF *venir+participle* construction can be observed which may be due to language contact through translations. So we may interpret our findings as reciprocal contact effects: from ME to AF and from OF/AF to ME. Our study suggests that whenever a v-framed and a s-framed language get in contact, it will be the v-framed properties of the s-framed language that will be copied and strengthened in the verb-framed language. Further studies are, of course, needed to corroborate this hypothesis.

## References

- [1] *Anglo-Norman On-line Hub*. Universities of Aberystwyth and Swansea, <http://www.anglo-norman.net/>, 2001.
- [2] *Base de Français Médiéval*. Céline Guillot-Barbance, ENS de Lyon, Laboratoire IHRIM, 2016.
- [3] Heather Burnett and Michelle Troberg. On the diachronic semantics of resultative constructions in French. In Christopher Piñón, editor, *Empirical Issues in Syntax and Semantics 10*, pages 37–54. Paris, 2014.
- [4] M. Callaway. *The Infinitive in Anglo-Saxon*. Carnegie Institution of Washington, Washington, DC, 1913.
- [5] Nick Ellis. Frequency-based grammar and the acquisition of tense and aspect in L2-learning. In M. Rafael Salaberry and Llorenç Comajoan, editors, *Research design and methodology in studies on L2 tense and aspect*, pages 90–117. de Gruyter Mouton, Berlin, 2013.
- [6] Helen Engemann. Learning to think for speaking about space in child bilingualism. *Revue française de linguistique Appliquée, Dossier: Dictionnaires*, XXI(2):49–64, 2016.
- [7] Eva Fernández, Ricardo Augusto de Souza, and Agustina Carando. Bilingual innovations: Experimental evidence offers clues regarding the psycholinguistics of language change. *Bilingualism: Language and Cognition*, (19):1–18, 2016.
- [8] Eric Haeberli. Syntactic effects of contact in translations: evidence from object pronoun placement in Middle English. *English Language and Linguistics*, Special issue, to appear.
- [9] Richard Ingham. Mixing language on the manor. *Medium Aenum*, (1):107–124, 2009.
- [10] Richard Ingham. *The transmission of Anglo-Norman: language history and language acquisition*, volume 9 of *Language faculty and beyond*. John Benjamins, Amsterdam, Philadelphia, 2012.

- [11] Lars Johanson. Contact-induced change in a code-copying framework. In Mari C. Jones and Edith Esch, editors, *Language change: the interplay of internal, external and extra-linguistic factors*, pages 285–313. de Gruyter, Berlin, 2002.
- [12] Anthony Kroch and Ann Taylor, editors. *The Penn-Helsinki Parsed Corpus of Middle English, Second Edition (PPCME2)*. University of Pennsylvania, Philadelphia, 2000.
- [13] Françoise Laurent. *La précocité de l'écriture hagiographique et l'identité normande: les vies de saints composées par Wace*. Publications numériques du CÉRÉdI, Rouen, 2013.
- [14] Dominica Legge. La précocité de la littérature Anglo-Normande. *Cahiers de civilisation médiévale*, (8):327–349, 1965.
- [15] Bettelou Los. *The rise of the To-Infinitive*. Oxford University Press, Oxford, 2005.
- [16] Leonard Talmy. Lexicalization patterns: Semantic structure in lexical forms. In T. Shopen, editor, *Language typology and syntactic description. Vol. 3: Grammatical categories and the lexicon*. Cambridge University Press, New York, 1985.
- [17] Leonard Talmy. Path to realization: A typology of event conflation. *Berkely Working Papers in Linguistics*, pages 480–519, 1991.
- [18] Ann Taylor. Contact effects of translation: Distinguishing two kinds of influence in old english. *Language Variation and Change*, (20):341–365, 2008.
- [19] Adolf Tobler and Erhard Lommatzsch. *Altfranzösisches Wörterbuch*. Weidmann, Berlin u.a., 1925ff.
- [20] Michelle Troberg and Heather Burnett. From latin to modern french: a punctuated shift. In Eric Mathieu and Robert Truswell, editors, *Micro-change and macro-change in diachronic syntax*, pages 104–124. Oxford University Press, Oxford, 2017.
- [21] Claude Vandeloise. *Spatial prepositions: a case study from French*. Univ. of Chicago Press, Chicago, 1991.
- [22] Fredericus Theodorus Visser. *An Historical Syntax of the English Language*. E. J. Brill, Leiden, 1963.
- [23] Maria Luisa Zubizarreta and Eunjeong Oh. *On the syntactic composition of manner and motion*, volume 48 of *Linguistic inquiry monographs*. MIT Press, Cambridge, MA, 2007.

# The Opacity of Modal Verbs in German: An ‘Optimal’ Answer to a Difficult Question

Elisabeth Scherr

Department of German Studies  
Karl-Franzens-University of Graz

E-Mail: [elisabeth.scherr@uni-graz.at](mailto:elisabeth.scherr@uni-graz.at)

## Abstract

This paper aims at showing how it is possible – via theoretical foundation, corpus analysis and statistical methods – to explain and motivate (up to now purely subjective) interpretations of modal verbs in German. The foundation of the research design is theoretic and based on the assumption that – in contrast to other modal verb interpretations – sentences with epistemic modal verbs are characterized by a specific deictic relation. It will be shown that this premise allows for the (motivated) definition of a number of properties accessible on the linguistic surface, concerning for example the subject / the agent role, mode or the temporal-aspectual features of the infinitive. Through the methodology of statistics and Optimality Theory it will be shown – exemplified by the German modal verb *dürfen* – that some of the defined properties can actually (i.e. statistically significant) indicate a specific interpretation of the inflected verb that does not depend on purely subjective evaluations.

## 1 Introduction

The primary motivation for the following research question was originally situated in the field of Variationist Linguistics. In non-professional contexts and in the media it is often said that ‘Austrians’ are characterized by communicative ‘indirectness’ and excessive politeness. Especially in comparison to speakers from Germany they would withhold strong personal perspectives in formal communication. Accordingly, the so-called *List of Learning Goals concerning Politeness and Intercultural Differences in the German-speaking Areas*<sup>1</sup> of the Institute of ‘Austrian German’ assumes the following: “Speakers of Austrian German, in comparison to speakers from Germany, act more indirectly in formal communication [...]. [...] Considerable self-promotion is rather avoided in Austria”<sup>2</sup> (Muhr [16]).

---

<sup>1</sup> *Lernzielliste zu Höflichkeitskonventionen und interkulturellen Unterschieden im deutschsprachigen Raum* [translation E.S.]

<sup>2</sup> „Österreichische Sprecher sind im Vergleich zu deutschen Sprechern in der öffentlichen Kommunikation eher indirekter [...]. Die starke persönliche Selbstdarstellung und das Herausstellen eigener Leistungen wird in Österreich eher vermieden.“ [translation E.S.]

Those tendencies would be reflected inter alia in the variable use of modal verbs (cf. *ibid.*).

From an empirical perspective, these assumptions severely lack data-driven support. Amongst other things this is due to the methodical difficulties that appear as soon as empirical proof for such questions is intended. Modal verbs are characterized by a vast functional complexity: They do not always convey indirectness or subjective perspective. In fact, within the scope of politeness or indirectness only the epistemic interpretational variant must be taken under consideration: Only epistemic modal verbs label propositions as being unclear in their factual status (cf. Kotin [11]) or as personal inference (Arrese [5]) (1), their deontic interpretations do not (2):

(1) epistemic:

*Für die Mieter muss es ein Schock gewesen sein.* (Nürnberger Nachrichten)  
,It must have been a shock for the tenant.'

(2) deontic:

*Ihren Führerschein muss die 43-Jährige nun abgeben.* (Schwäbische Zeitung)  
,The 43-year-old must hand in her driving licence now.'

Taken into consideration only the linguistic surface, one cannot tell the difference between epistemic and deontic interpretations of modal verbs: The verb is inflected and demands a bare infinitive in both cases. Thus, as it is with other opaque elements, their interpretation is highly subjective and at first glance empirically very hard to come by. The lack of significant evidence, on the other hand, leads to highly doubtful assumptions as the ones mentioned above and subsequently to prejudices with socially relevant implications.

But also in the majority of linguistic publications dealing with modal verbs too little or no attention is paid to their difficult classification.<sup>3</sup> Valuable contributions to the disambiguation of the different readings of modal verbs, however, came from the field of computational linguistics. The majority of these studies formulate features specific to a certain modal verb interpretation which are based on or subject to automatized corpus analyses. Most of this research has English as language of investigation. Ruppenhofer/Rehbein [17] for example ask annotators to make decisions regarding the interpretation of modal verbs in a sample of sentences and then try to elicit the characteristics of these sentences by means of a large-scale, automated corpus analysis. Their aim is to automatically generate

---

<sup>3</sup> Notable exceptions would be (inter alia) Abraham [1]/[2], Abraham/Leiss [3], Diewald [6], Heine [8]/[9], Kotin [10]/[11], Leiss [13].

interpretations via the application of algorithms. The authors themselves note, however, that their study is only a first tentative attempt to automatize the evaluation of modal verbs, portrayed as a complex challenge. Marasović / Frank [15] take a similar approach as they compile an automatically generated list of properties. As indications for epistemic interpretations (only for *must* and *can*) predicates of attitude (*believe*, *not know*, *fear*) as well as specific characteristics of the subject are mentioned. The results of this study can partly be parallelized with the present study (see below). A German-focused investigation is Zhou et al. [19], also attempting to determine the semantic properties of the different modal verb uses. In contrast to Ruppenhofer/Rehbein the authors take a priori characteristics which they consider as being relevant to the interpretation of modal verbs. In all these investigations sooner or later the question arises if the chosen methodology can be seen as reliable since the automated search does not always produce unequivocal results. Especially semantic features such as the *Aktionsart* properties of the infinitive, temporal relations or the aspectual properties are difficult to grasp for automatized studies and automatized annotations of such features seem to produce high error rates. The following aims at showing how an adequate, more objective evaluation of epistemic modal verbs can be executed and why it is crucial for this task to avoid unidimensional approaches and instead combine theoretically based assumptions with empirical analyses.

## 2 What is epistemic modality – deictically speaking?

The idea is that specific deictic relations of epistemic modal verbs allow for the (motivated) definition of a number of properties typical for epistemic interpretations. But what exactly are those ‘specific deictic relations’ and how do they serve to distinguish the different interpretation of modal verbs?

It is claimed here that the basic difference between deontic and epistemic interpretations of modal verbs is grounded on their deictic relation. Deontic modal verbs combine their lexical content with a temporal-deictic component, pointing towards future events (expressed by the infinitive). In other word: Something that is obligatory, necessary etc. is always future-projecting from a *hic et nunc* perspective, regardless of the time of reference:

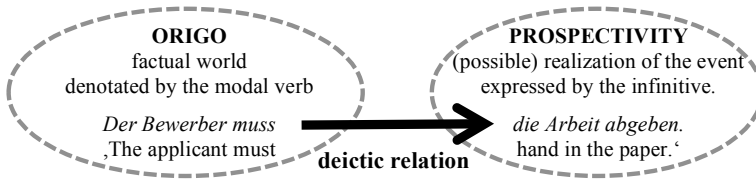


Figure 1: Deictic relation expressed by deontic modal verbs (Scherr [18])

This situation is totally different with regard to the epistemic interpretation: The lexical content of epistemic modal verbs is almost vanished up to the point where it just signalizes the strength of an inference (*dürfen, können, mögen, müssen*) or the source of a questionable statement (*sollen, wollen*). The deictic relation, however, is by no means future-projecting; instead it always indicates the reference to an event, which, in its factual status, is concurrent to the time of reference. In other words: The event evaluated in terms of its factual status is never located in a consecutive sequence but always concurrent or preceding with regard to the time of reference: The event expressed by the infinitive is (or is not) already realized. This information, however, is not accessible from the origos perspective; it is part of a *possible world* (Kratzer [12]).

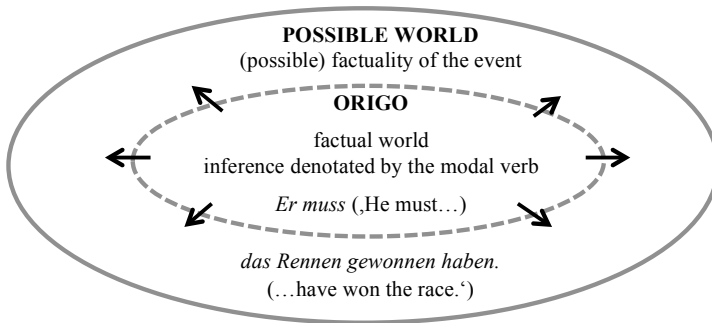


Figure 2: Deictic relation expressed by epistemic modal verbs (Scherr [18])

The question is: Why should one assume such a detailed and complicated basic interpretation of modal verbs, what is its explanatory value? It is the central characteristic, that epistemic modal verbs do not situate the event realized in the infinitive in a consecutive time period, which explains certain features accessible on the linguistic surface. The following list illustrates

some of them in exemplary selection<sup>4</sup>, followed by the tag used for annotation.

- *aktionsart* properties of the infinitive (INFZUSTAND)  
Non-telic, non-agentive infinitives such as *sein* ('to be') or *haben* ('to have') tend to support a non-future-projecting deictic relation. Conversely, telic or strongly agentive verbs such as *abgeben* ('to hand in') or *weggehen* ('to leave') tend to support a deontic reading.
- Aspectuality of the infinitive (ASPINF)  
Progressive forms such as the *Verlaufsform* stand in the way of future-projecting readings hence support the epistemic interpretation, c.f. *Er muss am Schlafen sein.* ('He must be sleeping.')
- Temporality and voice (INFKONSTREPI; \*INFVORPASS)  
Complex infinitives with perfective temporality are inherently stative, cf.: *Das muss weh getan haben.* ('This must have hurt'); Dynamic passive voice on the other hand delivers change-of-state-meaning with tendency towards a deontic interpretation: *Die Arbeit muss abgegeben werden.* ('The paper must be handed in.')
- Subject (\*SUBJPERS1/2; \*SUBJANI)  
Combinations between modal verb and first or second person (singular/plural) subjects tend to evoke a directive meaning (*you have to, I can, you must not* etc.). Epistemic modal verbs with first or second person subject are accordingly rare. It is further assumed that inanimate subjects tend to restrict a directive, future-projecting deictic relation as they usually do not serve as agents in the narrow sense.
- Grammatical mode (KONJII)  
Despite the functional complexity of the subjunctive, the coding of irrealis can be seen as one of its central purpose (cf. Fabricius-Hansen [7]). In other words: It prevents the implicature of an event being based on the actual, factual world (cf. Lötscher [14]). It is hence to be expected that inflected subjunctives tend to facilitate the epistemic interpretation.

---

<sup>4</sup> Other features concern the (in)definiteness-properties or adverbial properties. For reasons of space I decided to disregard them for this paper.

### 3 Database and Methods

The crucial questions are a) whether the defined properties can actually (i.e. statistically significant) indicate a specific interpretation and b) to which degree they indicate a particular interpretation of the inflected (modal) verb, in other words: Whether they can be brought into a relevance hierarchy illustrating the degree of significance for an epistemic interpretation. For answering these questions a total of seven subcorpora (for each canonical modal verb *dürfen*, *können*, *mögen*, *müssen*, *sollen*, *wollen* and *werden* in its epistemic variant) was drawn from a vast project corpus containing over 600 million words (tokens) illustrating the German *Gebrauchsstandard* in all of the German-speaking areas (press releases from Austria, Belgium, Germany, Liechtenstein, Luxembourg, South Tyrol and Switzerland). The seven subcorpora contain 300 epistemic and 300 non-epistemic examples each.<sup>5</sup> The validity of this selection was checked by *Annotator Agreement* (Fleiss' Kappa with the result of  $\kappa = 0.787536551$ ). As a next step, the previously defined properties were annotated manually in the seven subcorpora. An automated annotation was planned beforehand, due to high error rates, however, rejected. The significance of the occurrence of the single features was checked by the Chi-Square-Test and by means of (multivariate) Correlation Analysis: Only those features were taken into consideration which showed a significant frequency in the corpora (for each verb,  $p = 0.1$ ) and which showed a correlation coefficient higher than  $r \geq 0.4$  (substantial to very high correlation, cf. Albert / Marx [4]). The results of these calculations were determined separately for each verb and allowed for a ranking of significant features in accordance with their correlation coefficient. This coefficient displays the correlation of feature x with the epistemic interpretations of the respective sentences. For illustrative purposes the following table displays the results of the verb *dürfen*<sup>6</sup> (features with  $r \geq 0.4$  in bold):

---

<sup>5</sup> The consideration of epistemic and non-epistemic examples as well as the homogeneity of the corpora is crucial with regard to correlation analysis.

<sup>6</sup> Table 1 is a simplified presentation of the results due to the limited features considered for the purpose of this paper.



verb	feature	<i>r</i>
<i>dürfen</i>	KONJII	<b>0.98019606</b>
	*SUBJANI	<b>0.54671527</b>
	*SUBJPERS1/2	0.21282896
	INFZUSTAND	<b>0.42734305</b>
	INFKONSTREPI	<b>0.39857814</b>
	*INFVORPASS	0.08471352

Table 1: Correlation coefficients of (some) significant features of *dürfen*

It seems that for the verb *dürfen* the information concerning grammatical mode (KONJII) is highly indicative for epistemic interpretation: With the modal verb in the subjunctive mode the interpretation is epistemic with a probability of 98%, in the clearest cases supported by (some of) the other relevant features such as the animacy of the subject (\*SUBJANI) or the semantic or structural features of the infinitive (INFZUSTAND; INFKONSTREPI) (cf. (3)). It is just in cases where the subjunctive is used and all of the other features, however, speak in total against an epistemic interpretation, where the dominant feature may be overruled (cf. (4)). Both situations can very well be depicted with Optimality Theoretic methods:

- (3) *Der herannahende Zug dürfte die Postlerin zu spät gesehen haben [...].* (Kronen Zeitung, Steiermark und Kärnten).

‘The approaching train may have noticed the postwoman too late.’

	KONJII	*SUBJANI	INFZUSTAND	INFKONSTREPI
E $\frac{1}{2}$				
*E	*!	*	*	*

- (4) *Gerade in der Karnevalszeit trinken Jugendliche häufig hochprozentigen Alkohol, der ihnen nach dem Jugendschutzgesetz gar nicht verkauft werden dürfte.* (Rheinische Post).

‘It is especially during carnival season when high-proof alcohol is sold to adolescents which should not be sold to them according to the Protection of Young Persons Act.’

	KONJII	*SUBJANI	INFZUSTAND	INFKONSTREPI
E $\frac{1}{2}$		*	*	*
*E	*!			

Optimality Theory not only allows for the influence of differently weighted (principally violable) properties (*constraints*)<sup>7</sup> but also the generation of so-called *rankings*, which illustrate the influencing factors, their varying relevance, intend to allow combinations of features and *reranking* (cf. Scherr [18]).

Furthermore, with the aid of regression analysis it was tested which predictive force the features have for the interpretation of the modal verb. For *dürfen*, the regression analysis shows a coefficient of determination of 0.96, thus it can be assumed that the defined features can predict an interpretation with a probability of 96%. This in turn was tested against the evaluation of competent speakers. To that means, a questionnaire survey was carried out to check whether the calculated interpretation of the modal verb can be paralleled with the test person's assessments. This supportive pilot study in case of *dürfen* produced a result of  $\kappa = 0.91$ , which can be rated as excellent agreement between mathematical calculation and personal evaluation.

## 4 Some general findings

In a broader sense the illustrated approach may as well be applied to other semantic, pragmatic and/or functional questions. Within automated data analysis and computational linguistics it is especially these fields of interest which still imply major challenges. It should be pointed out that the theoretical motivation of defined features seems to be crucial: Only when the motivation of the selection of special features is clear, the specific challenges are obvious: For example the relevance of the infinitive can be located at a semantic (*Aktionsart*) as well as on a syntactic level (*features of voice/temporality*), as both strategies show more or less affiliation towards a specific deictic relation. By defining features of the linguistic surface which point to a qualitative assessment, the boundary between quantitative and qualitative research is certainly abolished (to a certain extent).

The results of this study, however, are also relevant for theoretical assumptions, which are time and again questioned in the course of the analysis. It appears, for example, that sometimes put forward features such as scope or other purely semantic features are not (yet) to operationalize. In this sense the collaboration between Computational Linguistics and the Humanities is essential. Up to now it seems that the purely automatized annotation is not leading to the strived results, at least for the aims of this study. Problems arise amongst others with features that concern purely

---

<sup>7</sup> Illustrated by the succession of features in the first row with the first one being the most influential.

semantic characteristics such as animacy of the subject or telicity of the verb. Those features have manifold realizations on the linguistic surface which make an automatized annotation hardly reliable. Last but not least it should be noted that results should always be interpreted with high caution and the awareness that one is dealing with only interpretational preferences here, not with absolute, objective decisions. In addition to that, it should be noted that the findings suggest that each modal verb has very differentiated outcomes; thus it seems hardly possible to proceed on assumptions concerning *the* modal verbs and their interpretational variants.

## References

- [1] Abraham, Werner (2009) Die Urmasse von Modalität und ihre Ausgliederung. Modalität anhand von Modalverben, Modalpartikeln und Modus. Was ist das Gemeinsame, das Trennende, und was steckt dahinter? In Abraham, Werner and Leiss, Elisabeth (eds.) *Modalität. Epistemik und Evidentialität bei Modalverb, Adverb, Modalpartikel und Modus*. Tübingen: Narr, pp. 215–302.
- [2] Abraham, Werner (2004) Modalität und Modalverben. Wohin führt uns die Syntax – inwieweit brauchen wir die Pragmatik? In Lindemann, Beate and Lentens, Ole (eds.) *Diathese, Modalität, Deutsch als Fremdsprache*. Tübingen: Stauffenburg, pp. 3–14.
- [3] Abraham, Werner and Leiss, Elisabeth (eds.) (2014) *Modes of Modality. Modality, Typology, and Universal Grammar*, Amsterdam and Philadelphia: Benjamins.
- [4] Albert, Ruth / Marx, Nicole (2016) *Empirisches Arbeiten in Linguistik und Sprachlehrforschung*. Tübingen: Narr.
- [5] Arrese, Juana (2011) Effective vs. epistemic stance and subjectivity in political discourse. In Hart, Christopher (ed.), *Critical Discourse Studies in Context and Cognition*, Amsterdam and Philadelphia: Benjamins, pp. 193–224.
- [6] Diewald, Gabriele (1999) *Die Modalverben im Deutschen. Grammatikalisierung und Polyfunktionalität*. Tübingen: Niemeyer.
- [7] Fabricius-Hansen, Cathrine (1997) Der Konjunktiv als Problem des Deutschen als Fremd-sprache. In Debus, Friedhelm and Leirbukt, Oddleif (eds.) *Studien zu Deutsch als Fremd-sprache. Aspekte der Modalität im Deutschen – auch in kontrastiver Sicht*. Hildesheim, Zürich and New York: Olms, pp. 13–36.
- [8] Heine, Bernd (1997) *Cognitive Foundations of Grammar*. New York and Oxford: Oxford University Press.

- [9] Heine, Bernd (1993) *Auxiliaries. Cognitive Forces and Grammaticalization*. New York: Oxford University Press.
- [10] Kotin, Michail (2007) *Die Sprache in statu movendi. Sprachentwicklung zwischen Kontinuität und Wandel*. Vol. 2: Kategorie – Prädikation – Diskurs. Heidelberg: Winter.
- [11] Kotin, Michail (2005) *Die Sprache in statu movendi. Sprachentwicklung zwischen Kontinuität und Wandel*. Vol. 1: Einführung – Nomination – Deixis. Heidelberg: Winter.
- [12] Kratzer, Angelika (1991) Modality. In Stechow, Arnim von and Wunderlich, Dieter (eds.) *Semantik. Ein internationales Handbuch zur zeitgenössischen Forschung*. Berlin and New York: de Gruyter, pp. 639–650.
- [13] Leiss, Elisabeth (2009) Drei Spielarten der Epistemizität, drei Spielarten der Evidentialität und drei Spielarten des Wissens. In Abraham, Werner and Leiss, Elisabeth (eds.) *Modalität. Epistemik und Evidentialität bei Modalverb, Adverb, Modalpartikel und Modus*. Tübingen: Narr, pp. 3–24.
- [14] Lötscher, Andreas (1997) Der Konjunktiv als pragmatischer Operator. In Heinz Vater (ed.) *Tempus und Modus im Deutschen*. Trier: Wissenschaftlicher Verlag, pp. 105–118.
- [15] Marasović, Ana / Frank, Anette (2016) Multilingual Modal Sense Classification using a Convolutional Neural Network. In: *Proceedings of the 1st Workshop on Representation Learning for NLP*. (URL: [https://www.aiphes.tu-darmstadt.de/fileadmin/user\\_upload/Group\\_AIPHES/publications/2016/repl4nlp.pdf](https://www.aiphes.tu-darmstadt.de/fileadmin/user_upload/Group_AIPHES/publications/2016/repl4nlp.pdf)).
- [16] Muhr, Rudolf (2000) Österreichisches Sprachdiplom Deutsch. Höflichkeitskonventionen. (URL: <http://www.oedeutsch.at/OESDCD/0INTRO/Gesamt-PDF/A06-hoeflichkeitskonventionen.PDF>).
- [17] Ruppenhofer, Josef / Rehbein, Ines (2012) Yes we can. Annotating the senses of English modal verbs. In Calzolari, Nicoletta et al. (eds.), *Eighth International Conference on Language Resources and Evaluation*. Paris: Language Resources Association, pp. 1538–1545.
- [18] Scherr, Elisabeth (2017) Die Opazität epistemischer Modalverben im Deutschen – Funktion, Form und empirische Fassbarkeit. Doctoral dissertation, University of Graz [in peer review].
- [19] Zhou, Megfei et al. (2015) Semantically Enriched Models for Modal Sense Classification. In *Proceedings of the EMNLP 2015 Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*. (URL: <http://www.aclweb.org/anthology/W15-2705>).

# The Poetic Corpus of Russian: Where the Poems are Written

Dmitri Sitchinava\* and Boris Orekhov<sup>+</sup>

<sup>\*\*</sup>School of Linguistics, Higher School of Economics, Moscow

<sup>\*</sup>Institute of the Russian language, Russian Academy of Sciences

Emails: <sup>\*</sup>mitrius@gmail.com, <sup>+</sup>nevmenandr@gmail.com

The paper discusses the marking of the composition location in the Poetic Corpus of Russian that enables customizing subcorpora by these locations and subsequent search by this parameter. The place names indicated by the authors are extracted, tagged and “normalized”, that is, all the different versions of names and minor locations are boiled down to a narrower range of locations put on an interactive map. This enables a study of lexical and other means used in the texts with regard to the location where the text is composed.

## 1. The Poetic Corpus of Russian: an introduction

The Poetic Corpus of Russian was launched in 2006 (see Grishina et al. [3]) and is available for online searching within the Russian National Corpus (ruscorpora.ru). It counts currently 11 million tokens, encompassing the period since the beginning of the 18<sup>th</sup> century until the end of the 20<sup>th</sup> century and has been further expanded every year. The texts include short and long poems, as well as drama in verse. The texts feature the word-by-word morphological markup common for the whole Russian National Corpus, to which another layer of information is added, viz. the poetic structure proper: its metric scheme (iambic, trochaic, free verse etc.), information on stanzas, the number of metric feet in each line, the type of rhyme and other parameters. A lemmatized word or a word combination is searchable within the subcorpus consisting of lines with customized poetic parameters, including the rhyming position. All the poems meeting a certain criterion can be also browsed without searching within their text a specific linguistic expression. The strong syllables (ictuses) in the traditional syllable-tonic or tonic (non-free) verse are systematically marked up, making the corpus a linguistic source in the history of the Russian stress in its own right (see eg. Grishina [3], Sitchinava [7] and others).

## 2. The text proper vs. Title-final Complex (TFC)

A poem as it is published by the author or in a critical edition often does not consist only of poetic (metric) lines, but includes also what is called sometimes Title-Final Complex (henceforth TFC; in Russian see eg the

Handbook of Poetry [1] on the topic). It may include all or some of the following elements: the title of the poem, subtitles, dedications, epigraphs (in the beginning of the text), the author's notes, the author's date of the composition of the poem, the place of its composition (in the end of the text) and possibly other elements. They are a part of the author's text and they are necessary for the analysis of a piece of poetry without instantiating verse in its proper meaning. The Corpus should provide the possibility to search key words within the TFC and the lines separately (see eg. Leibov [6]). For example, a search by the months' and seasons' names in Russian verse may not yield words like "summer" or "July" if they occur in the date in the end of the text (such as "July 23, 1856"). They are marked separately by the means of XML tags and they are separated from the metric lines. A regular search by Poetic corpus will not normally find them by default. The same should apply for the epigraphs (that are, in a vast majority of cases, quotations from other texts) and titles (the word usage in a poem's title can be a research topic in its own right apart from the study of the text of the poem's body, and the possibility of such a search query is to be provided; cf. Grishina [2]).

### 3. Marking up "composition locations"

The toponyms occurring in poetical texts have been already studied on corpora, for example by Leibov [5]: this author explores the "rhyming potential" of such place names as *Moskva* 'Moscow', *Varšava* 'Warsaw' and *Poltava* (Ukrainian town famous as the place of the 1709 battle fought by Peter the Great against the Swedish Caroline army and its Cossack allies); each of them has a range of political associations that are activated through rhyming words (such as *golova* 'head' or *slava* 'glory'). Different place names occur with different frequency in the rhyming position. A difficulty for this study was related to the fact that at that time the place markers included into the TFC were also marked up and searched alongside with the bulk of the text, which was the only search option and obfuscated the raw search results and statistics. In the work by Kuzmenko and Orekhov [4], the space of Russian poetry is analyzed from the point of view of toponyms (countries and cities) mentioned in the texts. Now, thanks to the markup of the TFC, we are able to compare the places mentioned in the texts with the location from the TFC. The map shows that in both cases Russian poetry is more European than American poetry. It has very few American toponyms. And the places mentioned by the authors and the places in which the poems were written are mostly the same.

A project of our team undertaken in 2016 consisted in separate marking of “composition locations” specified by the authors to make them a searchable field of the corpus’s database and to put them into the (modern) geographical map, showing the spatial domain of the Russia poetry. This parameter, however, could be marked if and only if it was specified in the known author’s text. Very often the poets failed to do so, either if they never did it (or neglected the TFC at all) or when a poem was composed in a “default”, unmarked location. It was unusual (although not unknown) for a poet living in Moscow to specify the city in all the texts created there; it was more natural for a denizen of Saint Petersburg or Odessa who visited the city. Alongside with this general challenge, some other problems occur.

The authors naturally specify the place names that were official or commonplace at the time of composition; these names could have graphic and other variants (*Sankt Peterburg* – *Petrograd* – *Leningrad*; *Peterburg*, *SPb.*, *Pburg* etc.), including even mistakes (for example, spelling of foreign place names with different order of letters or without some diacritics); all these variants were to be merged. These names could have been borrowed from different languages than it is the standard practice now (for examples, Estonian and Flemish place names were, before 1917, taken from German and French respectively, like *Hungerburg*, now *Narva-Jõesuu* in Estonia or names in *Saint-* instead of *Sint-* in Flanders). The authors often specified well-known or obscure microtoponyms beyond the city/town level (streets, neighborhoods, hospitals, hotels, restaurants etc.), whereas all the locations within the same city were to be kept together in order to make possible building of a subcorpus of this city. Some specifications within the TFC locations contained more than one location or additional information that was not toponymic or even altogether locative in nature (“train”, “asleep” etc.).

The corpus yielded 672 different locations that were later “normalized”, that is, for all of them a string of unified modern geographical names was proposed, with some localities included into other wider ones (for example: *Lubyanka IN Moscow*; *Boulevard Raspail IN Paris*). This boils down to about 400 “normalized” locations.

The metatextual markup of the Russian National Corpus was expanded by two additional fields: “location” and “normalized location”. The first one allows for an exact search as it was put but the author (*Leningrad* but not *Sankt-Peterburg*), whereas the second one consists of normalized locations that merge under one label all the alternative names.

#### 4. Search according to the geographical markup and case studies

The Russian National Corpus ([ruscorpora.ru/search-poetic.html](http://ruscorpora.ru/search-poetic.html)) provides now customization of subcorpora according to the both additional metatextual fields. The normalized location can be also specified with means of an interactive map based on Yandex Maps where the locations are marked according to the current Russian transliteration ([http://ruscorpora.ru/saas/poetry\\_map.html](http://ruscorpora.ru/saas/poetry_map.html)). It is possible to search within these customized subcorpora some lexical items that are characteristic for the texts created in a given location.

For example, it is possible to create a subcorpus of Russian poetic texts marked by Parisian locations; these poems are normally written by the Russian authors visiting Paris (not living there, as, for example, it was the case after the post-revolutionary emigration).

It is possible to use the subcorpora customized in such a way for studying some markers that correlate with the place where the text is created. The “Parisian texts” are predictably marked by a higher frequency of lexical markers that create the (stereotyped) image of the city: *bul'var* ‘boulevard’ (460 instances per million vs. 120 in the Muscovite corpus), *kaštan* ‘chestnut’ (250 instances per million in the Parisian corpus, not a single example in the Muscovite corpus), whereas *fonar* ‘lantern’ is not characteristic for either city and even is slightly more frequent (per million) in the Muscovite texts. Prosodic factors can also be statistically studied on these subcorpora, for example to define whether there is any statistically significant difference between the “Muscovite” and “Saint Petersburg” verse cultures (as this is sometimes claimed).

#### References

- [1] Azarova, Natalija, et al. (2016). *Poezija: Učebnik [Poetry: a handbook]*. Moscow: BSG Press.
- [2] Grishina, Elena (2005). Dva novyx proekta dlja Nacional'nogo korpusa: mul'timedijnyj podkorpus i podkorpus nazvanij [Two new projects for the Russian National Corpus: Multimedia Corpus and Titles' Corpus]. In: *Nacional'nyj korpus russkogo jazyka: 2003—2005*. Moscow: Indrik.



- [3] Grishina, Elena, Korchagin, Kirill, Plungian, Vladimir, Sitchinava, Dmitri (2009). Poeticheskij korpus v ramkax NKRJa: obščaja struktura i perspektivy ispol'zovanija [Poetic Corpus within the RNC: general structure and applications]. In: *Nacional'nyj korpus russkogo jazyka: 2006—2008. Novye rezul'taty i perspektivy*. SPb.: Nestor-Istorija.
- [4] Kuzmenko, Elizaveta, Orekhov, Boris (2016). Geography Of Russian Poetry: Countries And Cities Inside The Poetic World. In *Digital Humanities 2016: Conference Abstracts*. Jagiellonian University & Pedagogical University, Kraków. <http://dh2016.adho.org/abstracts/3>
- [5] Leibov, Roman (2012). Russkaja slava i pol'skaja stolica: k istorii odnogo rifmennogo kliše. [The Russian glory and the Polish capital: on the history of one rhyming cliché]. In: *Istorija literatury. Poëtika. Kino: sbornik v čest' Mariëtty Omarovny Čudakovej*. Moscow: NLO.
- [6] Leibov, Roman (2014). Neblagodarnyj pajščik: opyt korpusnogo analiza teksta [The Ungrateful Shareholder: an experience in corpus analysis of a text]. In: *Korpusnyj analiz russkogo stixa: vyp.2*. M.: Azbukovnik.
- [7] Sitchinava, Dmitri (2014). Akcentuacija glagola *byt'* v russkom stixe [Accentuation of the verb **byt'** in the Russian verse]. In: *Korpusnyj analiz russkogo stixa, vyp. 2*. M.: Azbukovnik.



# The Use of Language Corpora to Process Particles in a Monolingual Dictionary

Barbora Štěpánková

Czech Language Institute of the Czech Academy of Sciences

E-mail: stepankova@ujc.cas.cz

## Abstract

In comparison with its predecessors, the emerging monolingual dictionary of Czech provides more space for the description of particles. Due to the non-uniform treatment of particles in language handbooks, lexicographic work needs new methods. The new dictionary draws upon corpora of written Czech, but also corpora of spoken Czech and an annotated dependency corpus.

## 1 Introduction

This study is focused on the creation of dictionary entries for particles using corpora and corpus tools. It consists of three parts: 1. a summary of the corpora used for the new Czech monolingual dictionary; 2. a brief introduction of Czech particles, characterizing the class of particles and describing their position in Czech linguistics, especially in dictionaries and corpora; 3. methods for working with corpus data and possible applications of their results.

## 2 Dictionary and corpora

Since 2012, the Department of Contemporary Lexicology and Lexicography of the Czech Language Institute of the Czech Academy of Sciences has been working on the creation of a new monolingual dictionary of Czech with the working title *Akademický slovník současné češtiny* (The Academic Dictionary of Contemporary Czech) (Kochová et al. [7]), hereinafter the *ASSČ*. Its aim is to capture widespread contemporary Czech vocabulary used in public official and semi-official communication as well as in everyday (i.e. non-public, unofficial) communication. Unlike the previous Czech dictionaries, the *ASSČ* in preparation is not primarily based on targeted excerption, but rather, mainly on corpus resources. The essential material bases are the synchronic corpora of written texts from the Czech National Corpus, especially the reference representative corpora SYN 2010 and SYN 2015, and the SYN V5 versioned

corpus, the unification of all of the SYN-series of synchronic written corpora. The total size of the SYN version 5 is almost 4 billion words.<sup>1</sup>

When processing autosemantic words, a large corpus is an advantage, as automatic analysis software for collocations can be used, especially the corpus manager “sketch engine”. These tools are not suitable for processing synsemantic (function) words (including particles in our case), because a broader context than the words immediately surrounding them is needed to describe their meaning. Also, because they are not subject to inflection, their form cannot help to determine their meaning.

### 3 Particles

Czech particles as they are understood at present were first described by linguists at the end of the 1950s. Unlike basic word classes that have been described and defined rather uniformly in Czech grammars, particles still have not been properly delimited. However, according to grammars, they have a common feature – to express the speaker’s attitude. For example, the Polish linguist M. Grochowski, states “A particle is a unit from a metatextual level, it does not bring a comment to the world but rather to a speaker and his way of speaking about the world.” (Grochowski at al. [2], p. 397).

The main problems of automatic particle recognition are as follows:

1) there is little support for their definition in the contemporary dictionaries and grammars, which arises from their non-uniform processing;

- as indicated above, particles as a part of speech in their current form are a relatively recent phenomenon; handbooks concentrate on the typical representatives of each category primarily, and other particles are not often supported sufficiently by examples, so their usage and meaning are not always clear. Current dictionaries have two disadvantages: their age and somewhat outdated view (Slovník spisovného jazyka českého, Dictionary of the Standard Czech Language [12]) and limited scope (Slovník spisovné češtiny pro školu a veřejnost, Dictionary of Standard Czech for Schools and the Public [11]).

2) many particles are typical for spoken language, but have not been described in detail from this point of view in the literature;

3) homonymy and polysemy are common in these expressions.

Defining the breadth of homonymy and polysemy constitutes a specific linguistic problem that can be approached differently in each dictionary. The design of ASSČ proposes creating a separate headword/lemma for each part of speech a word is stably used as; it means part-of-speech difference is viewed as homonymy. Particles are most often homonymous with adverbs, or possibly conjunctions. The most important difference between particles and adverbs is their syntactic role: unlike adverbs, particles are not part of the syntactic structure. In conjunctions, their connecting function dominates,

---

<sup>1</sup> <http://wiki.korpus.cz/doku.php/en:cnk:syn:verze5> "3,836 billion words (tokens without punctuation)"

other functions (emphasizing, modifying) are limited in comparison with particles.

For example, the word *akorát* in the ASSČ will be divided into three lemmas (headwords):

**akorát I** - přísl.

v odpovídající, vhodné velikosti, času, stavu, množství ap.: *oblékl si kalhoty a byly **akorát***

(adverb: in a suitable, appropriate size, time, condition, quantity: *my pants fit **just right***)

**akorát II** - část.

1. omezuje platnost tvrzení na určitý prvek, skupinu, děj ap.: *k jídlu měli **akorát** kapustu*

2. zdůrazňuje měrové nebo časové určení: ***akorát** teď přišel*

(particle: 1. restrictive: *they had **only** cabbage to eat*, 2. emphasizing: *he came **just** now*)

**akorát III** spoj.

uvozuje větný člen nebo větu vyjadřující omezení nebo upřesnění předchozího tvrzení: *v normálních obchodech dostanete stejné zboží, **akorát** mnohem dražší*

(conjunction: restrictive, specifying: *in normal stores you get the same goods, **just/but** a lot more expensive*)

In some cases, the part of speech classification cannot be decided without knowledge of the broader context. For example: In the sentence *Petr je vážně nemocen*, the meaning of the word *vážně* can be interpreted in two ways: First – Petr is seriously ill; his illness is serious; Second – Petr is actually ill, it is not a joke. In spoken language, the difference can be recognized from the intonation and stress employed.

In corpora, a particle and its homonyms are often put under the same lemma and they share the same tags, cf. some expressions that were considered “typical particles” in grammars<sup>2</sup>, but that have been tagged in SYN corpora<sup>3</sup> in the following way:

*i* (and, too) - a conjunction

*ještě* (yet) – an adverb without gradation and negation;

*také* (too) – an adverb without gradation and negation;

Cf. In the last published dictionary, Slovník spisovné češtiny pro školu a veřejnost [11], these words are divided into two parts of speech: *ještě*, *také* an adverb and a particle, *i* a conjunction and a particle.

On the other hand, an expression with the same meaning can be tagged differently in the corpus, e.g. SYN 2015:

*Zvláště* (TT) *skladbu Zombie ocenili mohutným sborovým zpěvem.*

<sup>2</sup> Considering them “typical particles” in grammars does not exclude the possibility of classifying them as other parts of speech as well.

<sup>3</sup> Jelínek [5]

*They paid tribute **especially** (particle) to the song Zombie with a mighty choral rendition.*

*Z jeho neteře Raji se v útlém věku vyklubala vskutku zázračná kreslířka, **zvláště** (Db) koně zvládala suverénně.*

*His niece Raja turned out to be a miraculous illustrator at a tender age, she managed to draw **especially** (adverb) horses with great skill.*

Even though the Dictionary of Standard Czech for Schools and the Public separates the word **zvláště** into two parts of speech (adverb and particle), the meaning indicated in the last mentioned examples could be unambiguously rated as a particle.

## 4 Specialized corpora

In the Czech lexicographic tradition, the part of speech classification is a piece of information a user expects to find in the dictionary. At the same time, this information is closely connected to the description of functional-semantic roles, or the meanings of the words. The analysis of all corpus occurrences with a chosen lemma is beyond an individual's capabilities with regard to the high frequency of units. (See the table below.) Therefore, the verification of the expected semantic structure of a word was based on introspection and awareness of the lexicographer/author, the study of language handbooks, and surveying a random sample of corpus data. As indicated above, these resources are not always sufficient. Yet, to determine the functions of synsemantic words, we can use another corpus: the annotated Prague Dependency Treebank (PDT), created by the Institute of Formal and Applied Linguistics.<sup>4</sup>

The treebank contains a large amount of Czech texts with morphological, syntactic and complex semantic annotation. Semantic annotation, the so-called tectogrammatical layer, is particularly important for lexicographic processing, as it captures relations between words, not only their surroundings, and is represented by semantic functors that go across the surface parts of speech. Links between layers are another advantage, together with the software tool for corpus searches PML-TQ<sup>5</sup> (see Pajas [10]) it is possible to formulate queries across layers, e.g. one can search for the parts of speech which include the particle *také*, the functors assigned to it on the tectogrammatical layer, etc.<sup>6</sup>

4 Bejček et al. [1], Hajičová et al. [4], Mikulová et al. [9]

5 <http://lindat.mff.cuni.cz/services/pmltq/#!/treebank/pdt30/query/>

6 Some of the aforementioned methods were used in analyses of emphasizing particles (see B. Štěpánková [13]) and subsequently serve as a base for production/creating of dictionary entries in the ASSČ.

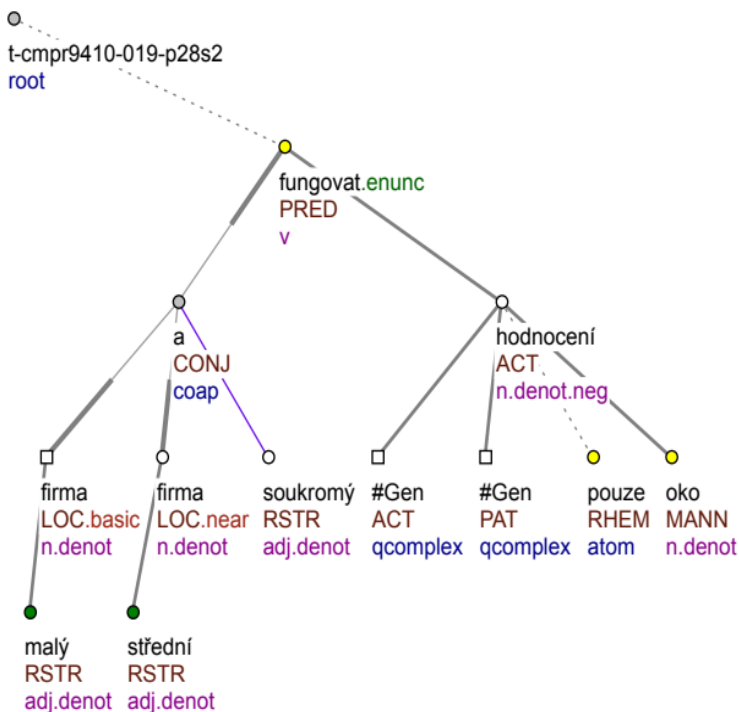


Figure 1.

*U malé a střední soukromé firmy funguje **pouze** hodnocení "od oka".*  
*For small and medium-sized private businesses, **only** ratings by guesstimate work.*

For some of the particle types, the presence of the annotation of topic-focus articulation is essential, as it gives us information about the so-called scope of the particle<sup>7</sup>. For example, it is possible to obtain a list of parts of speech or functors for which a close relation to the particle is indicated in the dependency trees, even though the words in the relation could be non-adjacent. (see Figure 1)

The PDT has some disadvantages for (non-computational) linguists: the work with the querying tools is more complex and the annotated data contain only journalistic texts. The examples for meanings verified in the PDT need to be looked up (based on similarity of construction) in the balanced corpora SYN 2010 and 2015.

SYN 2015 also incorporates a syntactic layer based on PDT, particularly the so-called analytical layer which captures surface syntactic

<sup>7</sup> Although particles are not considered a syntactic part of speech, scope is a specific relation that could be observed between a particle and a word or words that the particle emphasizes or influences (cf. Hajičová [3], Koktova [8], Štěpánková [13])

roles. The layer was created by an automated parser called “TurboParser”. According to T. Jelinek [6], its success rate is 80-85%, which is unsurprisingly less reliable than the morphological layer. Having the syntactic information is nonetheless an advantage. Expressions whose automatically assigned tags do not match the expected results (based on dictionaries, handbooks, and linguistic awareness) or differ statistically from the majority of results could (after filtering evident errors out) signify specific phenomena (idioms, multiword expressions, etc.) in some cases, e.g. *jen* (*only, just*) with an exceptional tag **TT-----/ExD** occurs in many examples with an elided verb:

*Já jen [= já jsem jen chtěla říct/dodat], že žádní hrdinové z filmů a z knížek ve skutečnosti neexistují.*

*I just [wanted to say/add] that no film or book heroes exist in reality.*

Particles typical for spoken language (e.g. filler particles) introduce a special problem: in the written language, fiction, and journalistic texts which are primarily available in the SYN corpora, their frequency is minimal and they are often marked (e.g. under the influence of a foreign language in translated texts). Instead, to verify their usage we can use some of the ORAL corpora also developed at the Institute of the Czech National Corpus.

## 5 Conclusion

As stated above, some information needed to process particles as dictionary entries can be obtained from, in addition to linguistic handbooks (grammars and dictionaries), specific types of corpora. The core of the lexicographic work with material still lies in corpora of the SYN type, which utilizes basic corpus tools. With regards to the frequencies of particles (see Table 1) we proceed from a random sample, similarly to more frequent expressions of other parts of speech, to determine a broad concept of their semantic structure. In our case, a lexicographer usually goes through a sample of 300 results manually.

An author writing a scholarly paper or handbook typically focuses on selected expressions, either common ones or, conversely, specialized ones. A lexicographer writing a dictionary, on the other hand, must deal with each word incorporated in the word list in detail. Examples illustrating particular meanings or specific grammatical constructions etc. (picked based on the PDT, for example) are deliberately looked up later.

In the end, the lexicographer investigates the degree to which the examples in the corpus sample correspond to the description based on tags. If the result reveals a great difference, changes the actual classification can be considered. Occasionally, it is possible to track down the cases and contexts in which a word behaves like a different part of speech. Lexicographic analysis and a dictionary based on a corpus can thus become tools to improve the user-oriented properties of the corpus itself.



	jen	ještě	prostě	také	už
SYN 2015	<b>228 725</b>	<b>163 636</b>	<b>27 669</b>	<b>148 948</b>	<b>286 827</b>
SYN V 5	<b>7 452 136</b>	<b>5 666 042</b>	<b>686 954</b>	<b>7 410 333</b>	<b>10 935 227</b>

Table 1.

Comparison of frequencies of selected particles in balanced corpus SYN 2015 and unified corpora SYN V 5

### **A sketch of corpus examples analysis based on tagging and syntax functions in SYN 2015:**

#### ***ještě*** (*yet/still*)

- 163 636 occurrences in total, all of them tagged as "Db" (an adverb without gradation and negation)
- random sample of 100 sentences/examples:  
afun (syntactic tag on the analytic layer):

AuxZ (emphasizing word)	52
Adv (adverbial)	47
Exd (actual/textual ellipsis)	1

Result of manual classification: most of the AuxZ examples should have been tagged as TT (a particle), as the analytical function suggests.

#### ***prostě*** (*simply*)

- celkem 27 669 výskytů, z toho:
- Dg (adverb with comparative degree) - 27 664x ,
- TT (particle) - 5x (unclear cases, probably mistakes)
- random sample 100 sentences/examples:
- all Dg (an adverb with the degree of comparison);
- afun (tag on the analytic layer):

Adv (adverbial)	66
AuxY (adverb or a particle)	26
Exd (actual/textual ellipsis)	7
Apos (appositon conjunction)	1

Result of manual classification: only 3 of 66 Dg/Adv are adverbs (fit to the characteristics of adverbs), the rest are particles; expressions tagged with

AuxY correspond to particles. In general, tagging *prostě* as TT would be more appropriate.

#### ***také (too)***

celkem 148 948 výskytů, všechny otagovány - Db (an adverb without gradation and negation)

- random sample 100 sentences/examples:

afun (tag on the analytic layer):

Adv (adverbial)	57
AuxZ (emphasizing word)	42
Exd (actual/textual ellipsis)	1

Result of manual classification: all the Adv expressions could be understood as TT, their contexts are very similar to those of the AuxZ.

#### ***akorát (just)***

- celkem 2827 výskytů, z toho:

Db (an adverb without gradation and negation) - 2 825x,

TT (a particle) - 2x

- random sample 100 sentences/examples:

afun (tag on the analytic layer):

Adv (adverbial)	83
AuxZ (emphasizing word)	9
ExD (actual/textual ellipsis)	7
Pnom (nominal predicate)	1

Result of manual classification: Only 21 of 83 words tagged with Adv could be considered adverbials (or part of a multiword adverbial units), the rest correspond to emphasizing particles and should be tagged with the TT part-of-speech tag and analytical function AuxZ.

## **Acknowledgement**

This work has been supported by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project CZ.02.1.01/0.0/0.0/16\_013/0001781).

## References

- [1] Bejček, Eduard et al. (2012) Prague Dependency Treebank 2.5 – a revisited version of PDT 2.0. In: *Proceedings of the 24th International Conference on Computational Linguistics (Coling 2012)*, pp. 231-246, Coling 2012 Organizing Committee, Mumbai, India.
- [2] Grochowski, Maciej, Kisiel, Anna, Żabowska, Magdalena (2014) *Słownik gniazdowy partykul polskich*. Krakow: Polska Akademia Umiejętności.
- [3] Hajičová, Eva (1995) Postavení rematizátorů v aktuálním členění věty. *Slovo a slovesnost* 56. pp. 241-251.
- [4] Hajičová, Eva, Kirschner, Zdeněk, Sgall, Petr: (1999) *A Manual for Analytical Layer Annotation of the Prague Dependency Treebank* (English translation) (html). **Available:** PDF, PS, BibTeX  
<https://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/a-layer/html/index.html>
- [5] Jelínek, Tomáš (2008) Nové značkování v Českém národním korpusu, In: *Naše řeč* 91. pp. 13-20.
- [6] Jelínek, Tomáš (2015) *Syntaktická analýza a syntaktické značkování*:  
[http://wiki.korpus.cz/doku.php/pojmy:syntakticka\\_analyza#vyhledavani\\_syntaktickych\\_struktur\\_v\\_kontextusyntakticke\\_atributy](http://wiki.korpus.cz/doku.php/pojmy:syntakticka_analyza#vyhledavani_syntaktickych_struktur_v_kontextusyntakticke_atributy) (2016/06/07).
- [7] Kochová, Pavla, Opavská, Zdeňka, Holcová Habrová, Martina (2014) At the Beginning of a Compilation of a New Monolingual Dictionary of Czech (A Report on a New Lexicographic Project). In: Abel, A., Vettori, C., Ralli, N., ed. *Proceedings of the XVI EURALEX International Congress: The User in Focus*. 15–19 July 2014. Bolzano/Bozen. pp. 1145–1151.
- [8] Koktova, Eva (1986) *Sentence adverbials in a functional description*, Amsterdam/Philadelphia.
- [9] Mikulová Marie et al. (2006) *Annotation on the tectogrammatical level in the Prague Dependency Treebank. Annotation manual*. Technical report no. 2006/30. Prague: ÚFAL MFF UK.
- [10] Pajas Petr, Štěpánek Jan (2008) Recent Advances in a Feature-Rich Framework for Treebank Annotation. In *The 22nd International Conference on Computational Linguistics - Proceedings of the Conference*. Manchester. pp. 673-680.
- [11] *Slovník spisovné češtiny pro školu a veřejnost* (1978) (Second, revised edition 1994; third, revised edition 2003.) Praha: Academia.

[12] *Slovník spisovného jazyka českého* (1960–1971). Praha: Academia.

[13] Štěpánková, Barbora (2014) *Aktualizátory ve výstavbě textu*. Praha: ÚFAL.

## Corpora

Kopřivová, M. - Lukeš, D. - Komrsková, Z. - Poukarová, P. - Wacławicová, M. - Benešová, L. – Křen, M.: *ORAL: korpus neformální mluvené češtiny, verze 1 z 2. 6. 2017*. Ústav Českého národního korpusu FF UK, Praha 2017. Dostupný z WWW: <http://www.korpus.cz>

Křen, M. – Bartoň, T. – Cvrček, V. – Hnátková, M. – Jelínek, T. – Kocěk, J. – Novotná, R. – Petkevič, V. – Procházka, P. – Schmiedtová, V. – Skoumalová, H.: *SYN2010: žánrově vyvážený korpus psané češtiny*. Ústav Českého národního korpusu FF UK, Praha 2010. Dostupný z WWW: <http://www.korpus.cz>

Křen, M. – Cvrček, V. – Čapka, T. – Čermáková, A. – Hnátková, M. – Chlumská, L. – Jelínek, T. – Kovářiková, D. – Petkevič, V. – Procházka, P. – Skoumalová, H. – Škrabal, M. – Truneček, P. – Vondříčka, P. – Zasina, A.: *SYN2015: reprezentativní korpus psané češtiny*. Ústav Českého národního korpusu FF UK, Praha 2015. Dostupný z WWW: <http://www.korpus.cz>

Křen, M. – Cvrček, V. – Čapka, T. – Čermáková, A. – Hnátková, M. – Chlumská, L. – Jelínek, T. – Kovářiková, D. – Petkevič, V. – Procházka, P. – Skoumalová, H. – Škrabal, M. – Truneček, P. – Vondříčka, P. – Zasina, A.: *Korpus SYN, verze 5 z 24. 4. 2017*. Ústav Českého národního korpusu FF UK, Praha 2017. Dostupný z WWW: <http://www.korpus.cz>

PDT 2.5 <http://lindat.mff.cuni.cz/services/pmltq/#!/treebank/>

# The European Union case law corpus (EUCLCORP) – a multilingual parallel and comparative corpus of EU court judgments

Aleksandar Trklja<sup>1</sup> and Karen McAuliffe<sup>2</sup>

<sup>1</sup>University of Vienna

E-Mail: [aleksandar.trklja@univie.ac.at](mailto:aleksandar.trklja@univie.ac.at)

<sup>2</sup>University of Birmingham

E-Mail: [k.mcauliffe@bham.ac.uk](mailto:k.mcauliffe@bham.ac.uk)

## Abstract

The empirical approach to the study of legal language has recently undergone profound development. Corpus linguistics study has, in particular, revealed previously unnoticed features of the legal language at both the lexico-grammatical and discourse level. Existing resources such as legal databases, however, do not contain functionalities that enable the application of corpus linguistics methodology. To address this gap in the context of EU law we developed a multilingual corpus of judgments that allows scholars and practitioners to investigate in a systematic way a range of issues such as the history of the meaning(s) of legal term, the migration of terms between legal systems, the use of binominals or the distribution of formulaic expressions in EU legal sub-languages. As well as being the first multilingual corpus of judgments it is also the largest legal multilingual corpus ever created. Since it contains case law from two sources (the Court of Justice of the European Union and EU national courts) it is also the largest comparable corpus of legal texts. The aim of the corpus is to contribute to the further development of the emerging field of language and law.

## 1 Introduction

The purpose of the present paper is to demonstrate how corpora and corpus linguistics methodology can contribute to the empirical study of the relationship between language and law. The European Union case law corpus (EUCLCORP) is a standardised, multidimensional and multilingual corpus of the case law of the Court of Justice of the European Union (CJEU) and of certain EU member states' constitutional/supreme courts. The project to produce a first 'proof of concept' version of the corpus was supported by a

European Research Council (ERC) Proof of Concept grant (based at the University of Birmingham, July 2016 - December 2017). The corpus was coded linguistically and with external metadata to enable users to contrast the meanings of terms and phrases across languages and legal systems, to compare translation options and to monitor the consistency of translation in EU case law. Furthermore, EUCLCORP allows users to track the migration of terms between legal systems and may be particularly useful for the creation of data-driven legal dictionaries and terminological databases. One important feature of legal terms is that they are semantically ambiguous. The complex search technique used in EUCLCORP allows users to identify differences in the meaning of terms by observing their typical patterns in CJEU judgments and the judgments of the relevant member state constitutional/supreme courts.

## **2 Existing resources**

Large online legal research services such as Westlaw and LexisNexis provide searchable databases of national and international case law. However, these databases do not contain corpus tools such as concordance lines, collocations, keywords or n-grams. Although the algorithms they use are both powerful and complex, users cannot compare the use of expressions in contexts across documents, or extract the most frequent words or multi-word expressions used in documents. Nor can users search within specific sections. Only general access to entire documents is available. Furthermore, Westlaw is accessible only in English and LexisNexis is available in only six EU languages and Russian. The EU's own online legal resource, EURLex, does allow users to compare different language versions of EU case law and legislation at the document level, but does not contain national case law. Other comparable databases include: N-Lex, JURIFAST, Caselex, and CODICES. The latter three do not provide any access to EU case law and are domain-specific: JURIFAST provides access to cases that deal with the application of EU law; Caselex is not yet completed and provides access to the 'most important' national cases that deal with certain areas of EU law (e.g. company law, competition law, consumer protection law); CODICES to a selection of important cases from constitutional courts. N-Lex is a work in progress and provides access to the most important legal acts in the majority of EU member states but not to case law. Parallel searches and searches of terms in contexts are not possible in any of the existing legal databases. These features are summarized in the table below.

Features/Database	Westlaw	LexisNexis	EURLex	N-Lex	JURIFAST	CODICES	Caselex	EUCLCORP
Multilingual		✓	✓	✓		✓	✓	✓
All EU languages			✓	✓		✓		✓
Parallel search								✓
Terms in contexts								✓
EU case law			✓				✓	✓
National case law	✓				✓	✓	✓	✓
Access to full texts	✓	✓	✓	✓	✓	✓	✓	✓
Non-domain specific	✓	✓	✓					✓

Table 1: Available functionalities in existing legal databases and EUCLCORP

EUCLCORP thus offers something that current legal databases do not and cannot. It is a multilingual, non-domain specific and comparable corpus containing case law from both the CJEU and national constitutional/supreme courts, which allows users to compare the use of legal terminology and concepts in context and to extract frequent expressions. The sub-corpus containing CJEU case law is also a truly multilingual translation corpus that will allow users to pull up different language versions of a particular term to investigate translation consistency and track the migration of legal terms through languages and case law.

### 3 Purpose of EUCLCORP

EUCLCORP will be of use to practitioners such as EU lawyers or translators and scholars of linguistics, translation studies or law. Legal linguistics is an emerging area of research (Tiersma, 1999; Mattila, 2013). As some recent studies have demonstrated (e.g. Biel, 2014; Trklja, 2017), a corpus linguistics approach can bring significant new insights to the study of the relationship between language and law. EUCLCORP has been created with the aim to foster the development of empirical legal linguistics studies. In the course of the research project ‘Law and Language at the European Court of Justice’ the authors of the present paper identified various interesting research questions that were difficult to investigate due to the lack of appropriate resources.

EUCLCORP makes it possible to address this gap by providing a resource that allows users of EU law to investigate in a systematic way:

- the history of the meaning(s) of a particular legal term;

- features that distinguish legal language from languages used in other registers;
- in the case of ambiguous terms – the senses in which they are most frequently and most typically used;
- the influence of national legal languages on EU case law (and vice versa);
- the impact of translation on the development of EU case law;
- discourse relations and argumentation patterns in judgments;
- the use of formulaic expressions in EU and national case law;

It should also be noted that EUCLCORP is innovative in terms of its size and richness. The present version of the EUCLCORP contains all CJEU judgments (in 23 languages) and eight member state constitutional/supreme court case law from 1952 onwards. The corpus will be ‘live’ in that, following its launch, it will be continually updated to contain new case law and case law from other types of member state courts. Thus, as well as being the first ever corpus of judgments it is also the largest legal multilingual corpus ever created. Since it contains case law from two sources (the CJEU and national courts) it is also the largest comparable corpus (a collection of ‘similar’ texts in different languages or in different varieties of a language).

The present (proof of concept) version of EUCLCORP contains all CJEU judgments (in 23 languages) and eight member state constitutional/supreme court case law from 1952 onwards. The aim (dependent on resources) is that EUCLCORP will be ‘live’ insofar as, following its launch it will be regularly updated with new case law and eventually with case law from ‘lower’ member state courts. Thus, as well as being the first ever corpus of judgments it will also be the largest legal multilingual corpus ever created. Since it contains case law from two sources (the CJEU and national courts) it is also the largest comparable corpus (a collection of ‘similar’ texts in different languages or in different varieties of a language).

## 4 Project development and annotation

The project has been developed in the following phases:

- Phase one: project application



- Phase two: data compilation
- Phase three: data annotation
- Phase four: web-interface
- Phase five: testing

The project is presently in its final stage.

The corpus has been annotated with linguistic and external metadata information. The annotation of the both types of information was carried out automatically. Linguistic information includes tokenization, lemmatization, parts-of-speech tags, sentence and paragraph boundaries and enumeration of sentences and paragraphs. The corpus was tokenized, lemmatized and POS-tagged by means of TreeTagger. In addition, several scripts were created in AWK to annotate the sentence and paragraph related information.

Non-linguistic metadata contains the following information for CJEU subcorpus: text sections (Summary, Parties, Grounds, Costs, Operative Part and Subject), language of the case, case name, case number, date and cellar number. The national courts do not contain section-related information because at the present stage it was not possible to identify such sections automatically. Non-linguistic metadata included to national judgments are language of the case, name of the court, date, case name and names of judges.

The present version of EUCLCORP contains all electronically available CJEU judgments in 23 languages and judgments from eight national courts. The number of cases was restricted by their availability and format. In the context of EU case law, differences between languages are due to the fact that not all judgments have been translated across all official languages. This is due to factors including the incremental addition of official languages with each enlargement/member state accession as well as changes in translation policy reflected in the rules of procedure of the Court of Justice itself.

EUCLCORP is available in two different formats: vertical and xml format. These two formats are used by most of existing corpus managers and corpus tools.

In the vertical format, each lexical item is described in terms of a token, parts-of-speech category and the lemma form. Other types of metadata (see above) are annotated in the form of XML tags. The purpose of creating the version of the corpus in the vertical format is because this format can be handled by Corpus Workbench tools (CWB)(Evert and Hardie, 2011). The complex search option available in the web-interface version of EUCLCORP

is based on CWB. CWB distinguishes between two types of XML tags: i) structural XML tags and ii) general XML tags. The first type is used for the annotation of sentence, paragraph and sections related information. Structural attributes allows the users to specify the regions within which the search should be performed. The second type of XML tags allows the users to display external metadata such as case number or date.

The conventional XML version of the corpus is TEI compliant and it has been created for the use with the corpus tools such as WordSmith (Scott, 2008) or AntConc (Anthony, 2014). All the sentences from judgments included in the CJEU sub-corpus were aligned at the sentence level to enable the search on parallel concordance lines.

Metadata allows users to limit their searches by focusing on specific periods, judgments or sections. For example, users can investigate the occurrence of specific terms only in the section called Grounds in specific periods. The search based on CWB allows users create complex search queries. For example the following search query identifies all occurrences of all word forms of the verb *increase* with any noun in Grounds in all judgments from 1980s. One useful feature of the CWB system is that it allows users to specify the number of words that might occur between search terms. In our example, `[] {0,2}` indicates that the number of items that might occur between *increase* and a noun ranges between zero and two. Some of the expressions identified by means of this query include *increased rate*, *increase the assets*, *increase the starting amount*.

```
[lemma="increase"           &                               tag="V.*"][] {0,2}[
tag="N.*"]::match.meta_date="1980.*" within grounds
```

Users do not have to provide as an input such a complex search query. As illustrated in Figure 1 the web interface offers a user-friendly set of options.

SearchFrequencyCollocationsN-grams

Query for English

Basic search

TokenincreaseV.\*

☐ begins with
☐ ends with
☐ case sensitive

Tokens in between: 02

TokenLexemeN.\*

☐ begins with
☐ ends with
☐ case sensitive

+ -

Search only in:

Summary

Parties

Grounds

Costs

Operative part

Subject

Metadata

case name

case number1980.\*

doc cellar

Figure 1: User-friendly interface for the search query [lemma="increase" & tag="V.\*"][] {0,2} [ tag="N.\*"]::match.meta\_date="1980.\*" within grounds

In addition, to the above described features the following functionalities have also been included in EUCLCORP: concordance lines (both monolingual and parallel) and collocation analysis and ngram analysis. The tool collocation analysis includes various statistical measurements (log-likelihood ratio, t-score, z-score, MI, MI3 and Dice coefficient).

These functionalities make it possible to address various research questions concerned with the study of language and law. For example, sentence and paragraph enumeration can help to identify expressions which are strongly associated with certain positions in discourse or text. Trklja and McAuliffe (forthcoming, 2018) demonstrate how paragraph initial metadiscursive (Hyland, 1998) formulaic expressions serve as indicators of discourse organization in ECJ judgments. The scope of the present paper does not allow an in-depth report of each aspect of those results. The results from this study indicate, for example, that the most frequent type of expressions signals Consideration-Conclusion relations in discourse. Such items signal connections between preceding and subsequent parts of

discourse. They indicate that the argumentation laid down in ECJ judgments proceeds from a discourse section concerned with discussion (Consideration) to the one which is concerned with providing a conclusion (Conclusion). Some of the expressions that signal this type of relations in discourse are *It follows from those considerations*, *It follows from the foregoing*, *It is apparent from the*, *It is clear from the*. Another important set of formulaic expressions associated with the paragraph-initial position are those items that signal questions and answers (e.g. *In its first question the*; *The questions referred to the*, *Is it of any significance*, *Does it make any difference*; *The answer must therefore be*; *The answer to the third*).

These findings illustrate how EUCLCORP can contribute to our understanding of the textual organization of ECJ judgments.

## 5 Conclusion

The purpose of the present paper was to present EUCLCORP as a new resource that can contribute to the development of empirical linguistic investigations of relations between language and law. Some additional societal benefits which are anticipated from the development of EUCLCORP include:

- Generation of new knowledge in relation to identifying the ‘EU law meaning’ of legal terms across 23 of the 24 Official EU languages;
- Increased opportunities for research projects using EUCLCORP as a tool with which to research in the fields of linguistics;
- Contribution to teaching materials available to students of law/legal linguistics and learners of legal languages: one problem in designing teaching materials for law students/those studying legal English/French etc. is the lack of examples available. EUCLCORP will serve as a source of examples for such teaching;
- Further development of the emerging field of law and language, with opportunities for interdisciplinary research projects using EUCLCORP as a tool;
- By adding to the big data currently available in legal databases, EUCLCORP has the potential to contribute to a better understanding of EU law and of the Europeanization of law as well as improved administration of justice.

The aim (dependent on resources is) to extend the present corpus in two ways: first, by including judgments from all European national courts; second, by annotating the corpus for legal concepts by using term extraction technique.

## Acknowledgments

The authors wish to thank, in particular: the European Research Council for funding the EUCLCORP project; colleagues at the Court of Justice of the European Union for their help in sourcing the relevant digital versions of judgments, in particular Mr Alfredo Calot Escobar and Andrew Paton; colleagues at the Court of Justice of the European Union and at the European Commission, who have been extremely helpful in the testing phase of this project, in particular Geoff Thomas, Dorothee Müller, Sandrine Umutoni, John Evans, John Kirby and Timothy Cooper. We would like to thank to Wim Peters, Michał Woźniak and Marek Medved who contributed to the development of EUCLCORP. We also wish to thank to anonymous reviewers for constructive and helpful comments on the earlier version of the paper. The usual disclaimers apply.

## References

- [1] Anthony, L. (2014) *AntConc (Version 3.4.3) [Computer Software]*. Tokyo: Waseda University.
- [2] Biel, L. (2014) ‘The textual fit of translated EU law: a corpus-based study of deontic modality’. *The Translator*, 20(3): 332-355.
- [3] Evert, S. and Hardie, A. (2011) ‘Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium.’ *Proceedings of the Corpus Linguistics 2011 conference*. Birmingham: University of Birmingham.
- [4] Gozdz-Roszkowski, S. (2011) *Patterns of linguistic variation in American legal English: A corpusbased study*. Frankfurt am Mein: Peter Lang.
- [5] Hyland, K. (1998) *Hedging in scientific research articles*. Amsterdam: John Benjamins

- [6] Mattila, H. (2013) *Comparative Legal Linguistics: Language of Law, Latin and Modern Lingua Francas*, 2nd ed. London: Routledge.
- [7] Scott, M. (2008) *WordSmith tools version 5*. Liverpool: Lexical Analysis Software, 122.
- [8] Tiersma, P. (1999) *Legal Language*. Chicago: University of Chicago Press.
- [9] Trklja, A. (2017) 'A corpus investigation of formulaicity and hybridity in legal language: a case of EU case law texts' In *Phraseology in legal and institutional settings: a corpus-based interdisciplinary perspective*, S. Gozdz-Roszkowski and G. Pontrandolfo (eds.), London: Routledge. 89-109.
- [10] Trklja and McAuliffe (forthcoming, 2018) 'Formulaic metadiscursive signalling devices in judgments of the Court of Justice of the European Union: A new corpus-based model for studying discourse relations of texts.' *English for Specific purposes*.

# NORMO: An Automatic Normalization Tool for Middle Hungarian

Noémi Vadász and Eszter Simon

Research Institute for Linguistics  
Hungarian Academy of Sciences

E-mail: {vadasz.noemi, simon.eszter}@nytud.mta.hu

## Abstract

The paper presents NORMO, an automatic normalization tool for Middle Hungarian texts with a memory-based and a rule-based module, which consists of character- and token level rewrite rules. The automatically normalized text eases and shortens the manual normalization work and results an edible output for further NLP tools. After exposing the modules of NORMO, we provide a thorough evaluation of the modules and the entire system, and we compare its performance to that of similar tools as well.

## 1 Introduction

The availability of annotated language resources is becoming an increasingly important factor in more and more domains of linguistic research: even outside the realms of computational linguistics high-quality linguistic databases can provide a fertile ground for investigations in theoretical linguistics. Historical corpora represent a rich source of data and phenomena from this perspective but only if relevant information is specified in a computationally interpretable and retrievable way. Several historical corpora have been built for Indo-European languages, such as the Penn-Helsinki Parsed Corpus of Middle English [6], or the Tycho Brahe Parsed Corpus of Historical Portuguese [5], and for non-Indo-European languages as well, such as the Old Hungarian Corpus [12].

For historical corpora, the development of several annotation layers, which is more or less prototypical in modern language corpora, requires a number of computational language processing tasks: i) sentence segmentation and tokenization; ii) normalization of tokens; iii) morphological analysis and morphosyntactic disambiguation. The normalization step is usually not required in the case of modern language corpora, but the situation is different in the case of historical texts. Because of the heterogeneity of the old orthographic system applied in these texts, this step, in which the original tokens are transcribed into their modern form, is required. This is a common step applied in most of the projects aiming at processing

historical linguistic material, e.g. [7, 8]. Normalization is inevitable and is obviously of critical importance: without normalization the performance of automatic annotation in later stages will suffer a dramatic decrease [11].

Since manual normalization is time-consuming and requires highly skilled and delicate work, application of automatic methods should be considered. Depending on the approach to be taken in searching for a solution for the normalization problem, several “metaphors” can be considered. A natural analogy to apply is approaching the normalization task as a translation from one representation language to another and therefore using machine translation models, e.g. [3, 10]. A successful class of current solutions are defined in the noisy channel paradigm [4, 9]. Most of the work on text normalization of historical documents is centered around a manually crafted or automatically induced set of correspondence rules [2] or some form of approximate matching based on a distance metric, often the Levenshtein distance, e.g. [1].

Here we present a normalization tool which has been developed for aiming the manual normalization of Middle Hungarian texts. It is fine-tuned for a Middle Hungarian Bible translation by Gáspár Károli, which is the first Hungarian translation of the entire Bible from 1590. However, it is not only a memory- and rule-based tool for the Károli Bible translation but an infrastructure which can be expanded with further dictionaries and rules for being able to normalize other texts from the history of the Hungarian language. There is no denying that a number of further techniques could be applied to find an efficient solution to the problem, however, given some practical constraints and requirements coming from the project (simple and quick to implement and integrate into a chain of modules responsible for different steps of normalization; perform well enough to be a significant aid in (the inevitable) manual correction process), we opted for an approach combining memory-based and rule-based methods.

## 2 Manual Normalization

Since the normalized text represents the basis of further text processing tasks, the highest accuracy is needed. Therefore, normalization is done by human annotators. However, since manual normalization is time-consuming and expensive, we want to cut the time needed for normalization with an automatic tool. Our experiences show that checking and correcting a normalized text is not only faster than normalizing the text manually but easier as well, because human annotators only need to correct some erroneous normalized words instead of typing all word forms.

One of the principal criteria of the normalization step is adherence to the original text – at least at the level of the morphosyntactic representation. Thus, we aimed for preserving all words and morphemes, even those which do not exist in Modern Hungarian. The second principle of normalization is consistency, thus orthographic variants of the same lexical item must be neutralized and converted into the same normalized version. We always followed the Modern Hungarian spelling



rules during the normalization process.

The tokenization step is also done manually during the normalization step. According to the Modern Hungarian spelling rules, some words have to be split up, while others have to be joined together. When a word in the original text belongs to different constituents, the word is split into the relevant parts. Words which are spelled apart in the original text, but constitute one word in Modern Hungarian, are joined. All these steps change the number of tokens, thus automatic tokenization rules have to operate across word boundaries (see Section 3.2.2). Each cross-boundary operation is marked, which is necessary for the evaluation of the token-level rewrite rules (see Section 4).

Since modern punctuation rules were created only in the 17th century, we cannot split the text into sentences based on the punctuation marks used in the original texts. For this reason, sentence splitting is also made manually during the normalization step. Some of the token-level rules applied in NORMO effect sentence segmentation as well, but – since it is a separate language processing task – it was not evaluated in this phase of the development.

### 3 Methodology

NORMO consists of two main modules: a memory-based and a rule-based module. The latter one contains character-level and token-level rules, which operate inside a token and across word boundaries, respectively. The order of the steps is important, since they depend on each other. First, memory-based replacements (Section 3.1) are done, whose result does not need further correction. The remaining tokens will undergo two kinds of rewrite rules: first, character-level rules run (Section 3.2.1), then second, token-level rules (Section 3.2.2) output the result of the whole process.

#### 3.1 Memory-based Normalization

Memory-based normalization uses a memory dictionary containing the most frequent original orthographic form–normalized form pairs. The dictionary is a `tsv` file in which the normalized counterpart is added to the original orthographic form. Since it was made by a human annotator, its advantage is precision. Once a token is found in the dictionary, it does not need further replacement or correction, consequently, they will serve as stop words for further normalization steps.

We have pointed out the high accuracy of the memory-based normalization. The recall depends on the size of the parallel corpus from which the memory dictionary was created. The parallel corpus contains the first five books of the New Testament of the Bible translation of Károli which were already normalized by human annotators. Since the manual normalization and the development of the automatic normalization tool have been done in parallel, the size of the memory dictionary is always growing. At the time of the evaluation, it contains 502 entries.

The disadvantage of memory-based normalization is that a full string match between the original word and its entry in the dictionary is needed. Since Hungarian is an agglutinative language, a lemma can get various inflections, but the memory dictionary contains only the most frequent forms. Hence it can occur that some variants are left out from the memory-based normalization. Not to mention that if the orthography of the original text is not consequent and a frequent word has two or more variants with at least one character difference they also can be left out. Before checking the word in the dictionary, we make the lower case version of each token to avoid the errors coming from the different capitalization of the word and its entry in the dictionary.

## 3.2 Rule-based Normalization

The rule-based module works with manually defined rewrite rules. These rules come from two sources: i) some of them were defined on the basis of known changes in the history of Hungarian, and ii) corpus-based observations could also help to define them.

The rewrite rules are context-dependent, where the context can be a word boundary or the neighboring characters. In the case of a match, a context-dependent rule changes the original word form to a new form in which the rewrite rules have been applied.

### 3.2.1 Character-level rules

Character-level rules operate inside a token. First, there are rules that handle the differences between the Middle and Modern Hungarian alphabet, as it converts the old characters which are not part of the Modern Hungarian alphabet (e.g. *Æ* and *ę*) to their modern counterparts (*e* and *é*, respectively). Second, character-level rules cover phonotactic regularities which are indicated in the Modern Hungarian spelling rules. For example, at the end of a word only *ó* or *ő* can occur, their short counterparts, *o* and *ö* never. According to this, NORMO converts word ending *o* and *ö* to *ó* and *ő*, respectively.

The advantage of using character-level rules during normalization is its disadvantage as well. On the one hand, character-level rules do not need full string match, they operate inside a word once the context matches. Accordingly, NORMO handles inflected variants of a lemma differently – in contrast with the memory-based approach, where full string match is needed. On the other hand, phoneme co-occurrences show different behavior inside of a morpheme and on morpheme boundary. For example, the rule that *i* in an intervocal position has to be converted to *j* is only true inside a morpheme, but it remains *i* on morpheme boundary. As Hungarian is an agglutinative language, we faced this problem in many cases. When setting up the rules, we had to take the number of the valid and invalid word forms generated by a rule into account. If the rule generates more invalid word forms than valid ones, we did not use the rule but left it to the human annotators.

The character-level rewrite rules are implemented as regular expressions. Since more than one rewrite rule can operate in one word and the output of a rule can trigger another one, the order of the rules is crucial. Thus, we have to take all existing rules into account when defining a new rule. For instance, the output of rules  $[tz \rightarrow c]$  and  $[(\wedge s)z(t) \rightarrow \backslash 1sz\backslash 2]$  depends on their order.

### 3.2.2 Token-level rules

Token-level rules operate across word boundaries. Token-level rules may also generate invalid word forms, but – in contrast with character-level rules – only one token-level rule is applied on a word form (after applying as many character-level rules as needed).

First, token-level rules can merge two tokens into one. For instance, Hungarian verbs often have verbal particles, which appear pre-verbally in neutral Hungarian sentences. In these cases, they are attached to the beginning of the verb, thus they constitute one token with the verb. However, this spelling rule applied in Modern Hungarian did not exist in Middle Hungarian, thus they had to be joined (*le megy*  $\rightarrow$  *lemegy* ‘he/she goes down’). Another example is the case of the demonstrative pronoun ‘az’ with the interrogative pronoun ‘ki’, which are actors in the story of the development of the modern relative pronoun (*az ki* ‘that who’  $\rightarrow$  *aki* ‘who’). Additionally, there are words commonly written in two tokens which are needed to be merged (*szent lélek* ‘holy spirit’  $\rightarrow$  *szentlélek*).

Second, token-level rules can split a token into two tokens. The aforementioned verbal particles appear post-verbally in imperative sentences, and they are spelled apart from the verb according to the Modern Hungarian spelling rules. However, they are often spelled together in Old and Middle Hungarian texts, thus they have to be separated (*menjle*  $\rightarrow$  *menj le* ‘go down!’). Another example is the case of *is* (‘also’) particle or *nélkül* (‘without’) postposition, which are always written together with the previous word, but in Modern Hungarian they form a separate token.

Token-level rules do not always output a valid Modern Hungarian word. Homographs cause problems for token-level rules, for example the word *meg* can be a verbal particle or a conjunction. In the first case – and if it is followed by a verb, which is another factor that should be taken into account –, it has to be joined to the next word, but in the second case it should not. Handling these erroneous outputs is left to the human annotators.

There are token-level rules affecting sentence tokenization, for instance a rule inserting a comma and an empty line indicating sentence boundary before ‘hogy’ conjunction and relative pronouns initiating a new clause. In this paper, we do not evaluate rules influencing sentence-level tokenization.

## 4 Evaluation

For evaluating the NORMO tool, we used the manually normalized books of the Károli Bible, namely the gospels of Matthew, Mark, Luke, and John and the Acts of the Apostles. The manually normalized version of the gospels served as the gold standard dataset in the evaluation. The number of the tokens based on the original form is 114,580 token, while based on the normalized version it is 109,289 tokens.

We used two metrics to evaluate the performance of NORMO. First, we calculated normalization accuracy, which is the percentage of tokens normalized identically to their gold standard counterpart [10]. This metrics shows the performance of the memory-based normalization and the character-level rules and indicates token-level accuracy. Second, we evaluated the performance of the token-level rules. Changes across word boundaries made by NORMO were divided into three main groups: a) false positive: NORMO joined or split words which are not joined or split in the gold standard; b) false negative: NORMO did not join or split words which are joined or split in the gold standard; and c) true positive: NORMO joined or split words which are joined or split in the gold standard too. Knowing the population of these subgroups, we were able to calculate precision, recall and F-measure. Table 1 shows the results.

	token-level accuracy (%)	above token-level		
		precision (%)	recall (%)	F-measure (%)
Matthew	81.04	87.90	69.40	77.56
Mark	80.62	86.18	65.80	74.62
Luke	81.58	84.56	67.50	75.07
John	83.10	94.20	70.60	77.03
Acts of the Apostles	79.95	90.72	67.13	77.16
avg	81.23	88.63	68.16	77.06

Table 1: The evaluation of NORMO on five parts of the Károli Bible.

The results reflects the usual outcome of rule-based systems: while the precision of the memory- and rule-based systems is fairly high, the recall is lower. Improvement of the recall can be obtained by extending the dictionary. At the time of the evaluation, we have used a dictionary with 502 entries. The ideal dictionary size were tested for reaching the highest recall with the least invested manual work. The performance of the memory-based module in NORMO were evaluated using dictionaries with different sizes – increased by 50. Figure 1 shows the results.

The performance of NORMO were compared with another automatic normalization tool called NORMA [1]. Similarly to NORMO, it is also a rule-based tool, but the character-level rewrite rules are learned from a manually normalized sam-

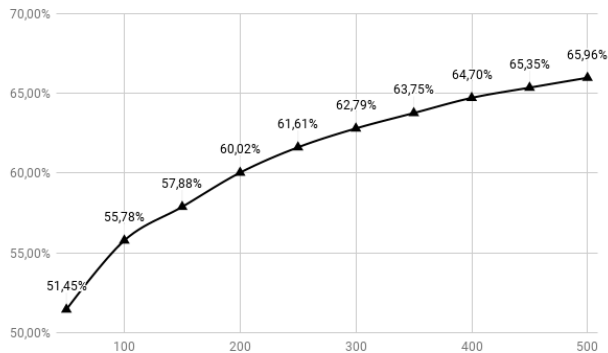


Figure 1: The improvement of the normalization accuracy according to the growing dictionary size. The size of the dictionary in words is on the horizontal axis, normalization accuracy is on the vertical axis.

ple text serving as the training data using a modified Levenshtein-algorithm. To compare the performance of the character-level rules, we only ran the rule-based module of NORMO. Table 2 shows the results. The results show that language history knowledge improve the performance of a rule-based tool in the case of a training data of this size.

	NORMO (%)	NORMA (%)
Matthew	68.29	61.58
Mark	66.43	61.70
Luke	67.03	61.31
John	68.59	59.14
Acts of the Apostles	66.98	60.18
avg	67.36	60.67

Table 2: Comparison of normalization accuracy of the character-level rules in NORMO and NORMA.

In Table 3, the token-level accuracy of the original texts shows how close the original spelling is to the modern one. In this case, token-level accuracy was counted before applying NORMO on the input text. The table also contains the number of the tokens which do not need any modification, since their old spelling version is the same as its modern counterpart ('good before normalization'). The last two columns show how many tokens were mistakenly modified by NORMO

(‘wrong after normalization’). Since the average error ratio produced by NORMO is quite low (3.72%), we can say that NORMO is highly precise in this sense too.

	token-level accuracy (%)	good before normalization (#)	wrong after normalization (#)	(%)
Matthew	41.92	8,033	330	4.11
Mark	41.81	5,110	235	4.60
Luke	41.94	8,734	398	4.56
John	45.08	7,062	181	2.56
Acts	41.78	8,237	240	2.91
sum/avg	42.44	37,176	1,384	3.72

Table 3: Token-level accuracy before normalization, the number of good tokens before normalization, the number of wrong tokens after applying NORMO, and its error ratio.

Another dimension of evaluating a tool which aims to facilitate and ease manual normalization is comparing the time devoted to the manual work to the time needed for inspecting and correcting the output of NORMO. Table 4 and Figure 2 show the results of this comparison. Here we take the token number of the original text form as the basis of the evaluation, because it serves as the input for both manual and automatic normalization. The estimated work time (100 token/hour) is based on previous experience when manually annotating the Old Hungarian Corpus. The actual work hours show how much time the manual correction of the output of NORMO took, and the last column contains the ratio of the estimated work time to the actual work time.

	token number (#)	est. work time (h)	act. work time (h)	ratio (%)
Matthew	28,520	285.2	400	140.25
Mark	18,150	181.5	116	63.91
Luke	30,805	308.05	96	31.16
John	23,435	234.35	80	34.13
Acts	28,631	286.31	100	34.92

Table 4: Comparison of the time needed for fully manual normalization to the time needed for manual correction of the output of NORMO.

The results show that the book of Matthew required the largest amount of manual correction. On the one hand, it is due to the fact that automatic normalization of the book of Matthew was done in an earlier phase of the development of NORMO.

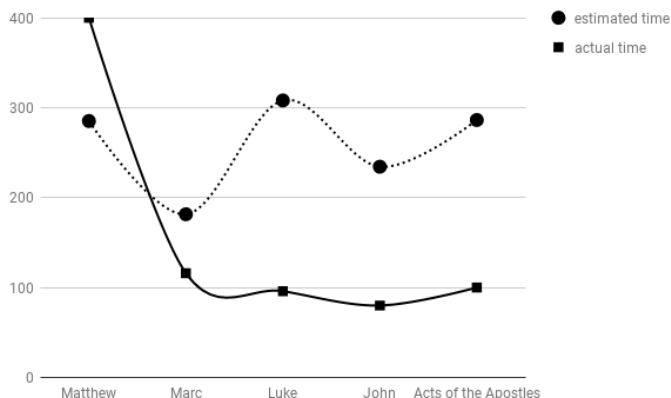


Figure 2: The time needed for correcting the output of NORMO. The texts are on the horizontal axis, the time (in hours) is on the vertical axis.

On the other hand, unlike previous practices, every word form gets a normalized counterpart even if it is not a valid word form in Hungarian. Therefore, the annotator needed more time for correcting these erroneous word forms. At a later stage of the development of NORMO, the higher the performance were, the fewer corrections and consequently the less time were needed.

## 5 Summary and Conclusions

In this paper, we presented NORMO, an automatic normalization tool with a memory dictionary and context-dependent rewrite rules which solves the task of automatic normalization of Middle Hungarian texts. The output of the tool makes the work of human annotators easier and faster, even though it has to be manually checked and corrected for the highest quality of the text which needs to be an edible input for further NLP tools.

In the current phase of the development, NORMO is adapted and fine-tuned to the language of the Károli Bible, but later it can be expanded with new memory dictionaries and rules which make it applicable for further Middle Hungarian texts. Our future plans include adapting it to other texts (e.g. to other Middle Hungarian Bibles) and implementing and evaluating other methods (e.g. Levenshtein-based normalization, machine learning approaches) as well.

## References

- [1] Marcel Bollmann. (Semi-)Automatic Normalization of Historical Texts using Distance Measures and the Norma tool. In *Proceedings of ACRH-2*, pages 3–14, Lisbon, Portugal, 2012.
- [2] Marcel Bollmann, Florian Petran, and Stefanie Dipper. Rule-based normalization of historical texts. In *Proceedings of LaTeCH-2011*, pages 34–42, Hissar, Bulgaria, 2011.
- [3] Marcel Bollmann and Anders Søgaard. Improving historical spelling normalization with bi-directional LSTMs and multi-task learning. In *Proceedings of COLING 2016: Technical Papers*, pages 131–139, Osaka, Japan, 2016.
- [4] Eric Brill and Robert C. Moore. An improved error model for noisy channel spelling correction. In *Proceedings of ACL 2000*, 2000.
- [5] Charlotte Galves and Helena Britto. The Tycho Brahe Corpus of Historical Portuguese, 2010. URL: <http://www.tycho.iel.unicamp.br/tycho/corpus/en/index.html>.
- [6] Anthony Kroch and Ann Taylor. The Penn-Helsinki Parsed Corpus of Middle English (PPCME2), 2000. URL: <http://www.ling.upenn.edu/ppche-release-2016/PPCME2-RELEASE-4>.
- [7] Tony McEnery and Andrew Hardie. *Lancaster Newsbooks Corpus*, 2003. URL: <http://www.lancs.ac.uk/fass/projects/newsbooks/default.htm>.
- [8] Attila Novák, Katalin Gugán, Mónika Varga, and Adrienne Dömötör. Creation of an annotated corpus of Old and Middle Hungarian court records and private correspondence. *Language Resources and Evaluation*, Jun 2017.
- [9] Csaba Oravecz, Bálint Sass, and Eszter Simon. Semi-automatic Normalization of Old Hungarian Codices. In *Proceedings of LaTeCH 2010*, pages 55–60, Lisbon, Portugal, 2010.
- [10] Eva Pettersson. *Spelling Normalisation and Linguistic Analysis of Historical Text for Information Extraction*. PhD thesis, Uppsala University, Department of Linguistics and Philology, 2016.
- [11] Paul Rayson, Dawn Archer, Alistair Baron, Jonathan Culpeper, and Nicholas Smith. Tagging the Bard: Evaluating the accuracy of a modern POS tagger on Early Modern English corpora. In *Proceedings of the Corpus Linguistics Conference (CL2007)*, UK, 2007.
- [12] Eszter Simon. Corpus building from Old Hungarian codices. In Katalin É. Kiss, editor, *The Evolution of Functional Left Peripheries in Hungarian Syntax*, pages 224–236. Oxford University Press, 2014.



# A Study on Appraisal Resources in Argumentative Essays by non-English Major Postgraduates

Wang Hongli and Chen Ying

School of Foreign Studies

Xi'an Jiaotong University

E-mail: hansen@mail.xjtu.edu.cn

## Abstract

Argumentative writing requires authors to analyze, discuss, and resolve controversies in a way that is clear, convincing, and considerate of diverse points of view. However, it is difficult for non-native English speaking students to construct interpersonal meaning in argumentative writing. Based on Appraisal theory, this study mainly investigates the employment of Appraisal resources in English argumentative essays written by Chinese non-English major postgraduates. To begin with, this study explores the overall distribution of Appraisal resources in argumentative essays and finds that Engagement resources, as the most frequently used ones, play a crucial role in achieving the rhetorical purpose of argumentative essays. The pattern of more Engagement than Attitude and Graduation could be regarded as the characteristic of the argumentative genre. Afterwards, this research compares and contrasts Appraisal resources across high-rated and low-rated argumentative essays and explores how Appraisal resources co-articulate to achieve the rhetorical purpose of argumentative essays. The findings of this study can contribute to successful argumentative essay writing and offer some pedagogical implications.

## 1 Introduction

Persuasive writing is a demanding task that requires the use of complex language to analyze, discuss, and resolve controversies in a way that is clear, convincing, and considerate of diverse points of view (Nippold, Wald-Lonergan, & Fanning, 2005, p. 125). Thus, argumentative writers should learn how to analyze and evaluate, how to position themselves and how to articulate their positions in a congruent manner. However, EFL/ESL students consider that argumentative writing is difficult and challenging (Hirose, 2003; Lee, 2006) and their essays are often criticized for the absence of critical stance and personal voice (Johns, 1997). Liu (2013) found that proper use of evaluative language in argumentative writing contributes to establishing personal voice and positioning of readers, thus resulting in more effective critical arguments.

Developed within SFL, Martin and White's Appraisal theory (2005) offers a comprehensive, systematic, and thorough typological framework for analyzing interpersonal meaning of discourse at the level of discourse

semantics. This framework comprises three major subsystems, namely, Attitude, Engagement and Graduation, each having a few subtypes.

Studies on evaluative language in argumentative essays on the basis of Appraisal theory have different foci on the three subsystems of its framework (Lee, 2008; Liu Dan, 2013; Swain, 2010) or have different dimensions of comparison (Lv, 2015; Liu & Thompson, 2009; D.P. Thomas *et al.*, 2015; Liu, 2013). However, most studies have paid little attention to how Chinese non-English major graduates employ Appraisal resources in argumentative essays. Moreover, many studies have generally analyzed evaluative language in case studies or in relatively small data projects. This paper investigates how Chinese non-English major postgraduates deploy Appraisal resources to construct interpersonal meaning in English argumentative essays.

## 2 Methodology

Drawing upon Martin and White's Appraisal theory (2005), this study first explored the overall distribution of Appraisal items in persuasive essays, and then compared high-rated and low-rated argumentative essays and summarized some patterns for articulating Appraisal resources.

### 2.1 Data collection

This study selected 160 argumentative essays written by first-year non-English major postgraduates in the course *Advanced English Writing* from a Chinese university over the past three years. The topic of the writing task was “*Should Financial gains be the most important factor in choosing a career?*” with the requirement of “Discuss the extent to which you agree or disagree with the opinion above. Support your views with reasons and/or examples from your own experience, observations, or reading.”

In order to explore the deployment of Appraisal resources in argumentative essays, an Argumentative Essays Corpus (AEC) was established, comprising 160 essays with 80610 tokens. This corpus consists of two sub-corpora: a High-rated Argumentative Essays Corpus (HAEC) and a Low-rated Argumentative Essays Corpus (LAEC), with 80 texts in each corpus. The former contains 42819 tokens with an average of 535 per text while the latter includes 37791 tokens with an average of 472 per text. Each essay was labeled according to its level (namely H or L), and was assigned a sequential number for ease of access as needed.

## 2.2 Rating and annotation

This paper employs the same scoring method and rating rubrics as the “Analyze an Issue” task in the analytical writing section of the GRE General Test. The GRE General Test adopts holistic scoring and assesses critical thinking skills along with the writer’s grammar and the mechanics of writing. All raters have extensive experience scoring tests and were all received training prior to evaluating the essays. Each essay was rated on a scale from 0 to 6 in half-point increments. Essays rated at 3.5 or higher were grouped in the HAEC and those below 3.5 were grouped in the LAEC.

Each essay was scored by two separate raters using the six-point holistic scale. In holistic scoring, raters separately assign scores on the basis of the overall quality of an essay in response to the assigned task. If the two scores differ by more than one point on the scale, the discrepancy is adjudicated by a third rater. Otherwise, the two scores are averaged and rounded to the nearest half-point interval on the 0–6 score scale as the final score on each essay.

The UAM CorpusTool (UAMCT) was used to annotate Appraisal items in each essay. The coding scheme, adapted from Martin and White’s (2005) Appraisal theory, is presented in Appendix A.

Given that annotating Appraisal options in argumentative essays is subjective in that it is an interpretive act, in order to ensure the validity of annotation, and following the recommendation of (Geng & Wharton, 2016; Xie, 2016), one external analyst, who is familiar with Appraisal theory, independently annotated all essays. The disagreements on any problematic items were recorded and discussed with a third analyst and then were settled through elaboration and clarification of the defining criteria for Appraisal typology.

## 2.3 Data analysis

As the texts in the argumentative essay corpus have different length, raw frequency counts of Appraisal options need to be normalized to an appropriate basis for meaningful comparison between the two sub-corpora. The normalized basis should be approximate to, or at least should not be higher than, the average length of text in a corpus, in order to avoid artificial inflation of the counts for rare features (Biber & Jones, 2009, p. 1299). As a result, the basis for normalization was set at 100 words given that each of the two HAEC and LAEC sub-corpora has an average of about 480 and 423 per text, respectively.

The study employed qualitative and quantitative methods to explore the Appraisal resources in the two sub-corpora. The data was analyzed using *the Pearson Chi-square test, the Mann-Whitney U test and the t-test*.

### 3 Results and Discussion

#### 3.1 Appraisal resources in the full corpus

Table 1 displays the percentages and frequencies of the three subsystems--Attitude, Engagement, and Graduation--in the Appraisal framework observed in the entire corpus. The entire corpus has a total of 7430 instantiations with Appraisal feature. Obviously, as the most frequently used option, the Engagement subsystem has more than twice as many appraisal features as either of the other subsystems, which indicates that Engagement items play a crucial role in achieving the rhetorical purpose of argumentative essays.

Subsystem in Appraisal framework	Percentage	Frequency
Attitude	23.89%	1775
Engagement	51.79%	3848
Graduation	24.32%	1807
Total	100%	7430

Table 1: Percentages and Frequencies of Appraisal Resources in the Full Corpus

Within the Attitude subsystem, Appreciation (n=1059, 59.66%), as the dominant subcategory, greatly outnumbers Affect (n=587, 33.07%), while Affect greatly outnumbers Judgement (n=129, 7.27%). More Appreciation options are likely to contribute to objectivity and persuasion in argumentative essays by reducing the subjectivity and personalization.

#### 3.2 Comparison of Appraisal between HAEC and LAEC

As shown in Table 2, both corpora show a strong preference for Engagement to impart evaluative meanings although the distributions of Appraisal in the two groups are somewhat different. In addition, the normalized frequencies (per 100 words) of Appraisal resources appear to reveal that writers of low-graded argumentative essays apply more Appraisal resources than those of high-graded ones. Both the Pearson Chi-square test ( $p<.05$ )and T t-test ( $p<.05$ ) results indicate that the two groups differ significantly in use of Appraisal resources, and further analysis of the results shows that the high

score group’s preference for Engagement and Attitude over Graduation is stronger than the low score group.

	HAEC			LAEC		
	Frequency	Normalized Frequency	Percentage	Frequency	Normalized Frequency	Percentage
Attitude	864	2.02	24.39%	911	2.41	23.44%
Engagement	1895	4.43	53.49%	1953	5.17	50.24%
Graduation	784	1.83	22.13%	1023	2.71	26.32%
Appraisal	3543	8.27	100%	3887	10.29	100%

Table 2: Appraisal Resources across HAEC and LAEC

3.3 Attitude across two sub-corpora

The procedure presented in section 3.2 was used to identify any significant difference in Attitude options across HAEC and LAEC. As the three subtypes of Attitude are not normally distributed, the Mann-Whitney U test instead of t-test was used to verify the accuracy of research results (Xiao & Bi, 2015).

Although both sub-corpora show identical distribution patterns, students in HAEC tended to use Appreciation realizations, while students in LAEC were more inclined to use Appreciation and Affect options (see Table 3). HAEC and LAEC only display a significant difference in the normalized frequency of Affect ( $U = 1671.00, p = .000$ ), revealing that students in HAEC adopted much fewer Affect resources than those in LAEC, thus exhibiting less emotional disclosure resulting in texts that appeared more objective, persuasive, authoritative instead of subjective and personal.

	HAEC			in	LAEC			in
	Frequency	Normalized Frequency	Percentages Attitude		Frequency	Normalized Frequency	Percentages Attitude	
Affect	207	0.48	23.96%		380	1.01	41.71%	
Judgement	76	0.18	8.80%		53	0.14	5.82%	
Appreciation	581	1.36	67.25%		478	1.27	52.47%	

Table 3: Attitude Resources across HAEC and LAEC

As for the subcategory of Affect orientation, both groups used more non-authorial than authorial evaluations, but in comparison with students in LAEC (authorial evaluations=37.37%; non-authorialevaluations=62.63%), students in HAEC (authorial evaluations=20.29%; non-authorial evaluations=79.71%) were more likely to use non-authorial evaluations, thus making the writer take less responsibility for the attitudinal value assessment and resulting in a more objectively sounding text.

With regard to Affect type, writers in HAEC adopted a greater variety, and had more balanced use of the subcategories of Affect, indicating highly sophisticated deployment of Affect resources and a high level of writing and language skills in HAEC. On the contrary, writers in LAEC manifested an unbalanced use of the subtypes of Affect, as the Un/happiness (51.05%) items account for more than half the total percent of Affect.

In addition, the Affect resources across HAEC and LAEC are mostly realized by behavioral surge or surge of feelings, such as *like, enjoy, happy, boring, interested, want, hope, and complaint*, which is consistent with the findings of Liu (2013) and Liu and Thompson (2009). The appraisal agents in these surges are presented and foregrounded and this kind of foregrounded Affect has the potential to position readers attitudinally, and provokes their emotional response, thus achieving emotional resonance between readers and writers, which in turn may well enhance persuasion. However, a close analysis of Affect realizations reveals the richness of the HAEC essays and the paucity of the LAEC essays. The two groups also show a significant difference in lexical richness. The top eight frequently used Affect realizations (want, interested, like, love, interest/s, enjoy, happy, boring) in HAEC are much lower than those in LAEC. Moreover, the top eight most frequently used words take up 57% of all Affect realizations in LAEC, while those same words only occupy 37% in HAEC, suggesting that students in HAEC deploy more varied vocabulary to express their emotions. In addition, writers in HAEC sometimes employ nominalized items (e.g. *boredom, anxiety, depression*) to express Affect, which obscures the agent of emotion and makes the text impersonal.

In HAEC, students employed Affect resources to prosodically interact with other Attitude resources so as to provide support for arguments and strengthen persuasive power. This interplay of Affect, Judgement and Appreciation was also identified in previous research (Lee 2008; Liu, 2013; Liu & Thompson, 2009). Example 1 in HAEC and Example 2 in LAEC both place emphasis on the importance of interest in job. However, the author in Example 1 employs the interplay of Affect and Appreciation or Affect and Judgement to state the significance of interest in job, thus achieving a strong sense of persuasion. Compared with Example 1, the writer in Example 2 first personally states the significance of interest and then illustrates the advantage through separate use of Affect and Appreciation, thus imparting a subjective and unconvincing assertion that is unsubstantiated with supporting evidence.

#### *Example 1*

First of all, a person's interest (Affect: Satisfaction) in an area is a

crucial (Appreciation: Valuation) factor that helps (Appreciation: Valuation) to build his career and self-actualization. If a person likes (Affect: Happiness) the job he is doing, he will be completely immersed in his career with enthusiasm (Affect: Satisfaction) and perseverance (Judgement: Tenacity). Then, it is much easier (Appreciation: Composition) for him to grow rapidly and achieve (Judgement: Capacity) his dream. (H03)

*Example 2*

...I even think that interest is more important (Appreciation: Valuation). If we are interested (Affect: Satisfaction) in our work, we will use greater passion (Affect: Happiness) to do it. So our work efficiency will make greatly improve. At the same time our work performance will be very good (Appreciation: Reaction). We also will be very happy (Affect: Happiness). (L13)

It seems that both sub-corpora demonstrate the same distribution pattern of Judgement items since more Social Esteem options are applied than Social Sanction ones. Despite the identical distribution pattern, however, writers' preference for the four subcategories is significantly different. The predominance of Capacity, Normality and Propriety and irrelevance of Tenacity and Veracity in the high score group were found to be similar to previous studies (Wu & Allison, 2003; Lee 2008; Liu, 2013; Liu & Thompson, 2009), suggesting writers' inclination to evaluate people's capacity and behaviors in HAEC. In contrast, the low score group tend to employ Capacity and Tenacity, demonstrating writers' inclination to evaluate people's capacity and psychological disposition in LAEC.

The Judgement instantiations are generally realized by adjectives and adverbs and the targets of Judgement resources are usually explicitly identified as human beings across the two sub-corpora, which is inconsistent with the findings of Lee's (2008) and Liu's (2013) research. These researchers insisted that nominalized expressions of Judgement items without explicit targets are deemed as characterizing successful argumentative academic writing which tend to sound impersonal and thus maintain a certain level of formality. This discrepancy indicates that the effective use of Judgement should receive considerable attention in the teaching of argumentative writing so as to establish an impersonal and formal tone.

The three subcategories of Appreciation resources across the two sub-corpora show the same distribution pattern, with Valuation as the dominant option in Appreciation for both groups. This reveals that the two groups are more inclined to explicitly evaluate the worthiness, usefulness and importance of things or events and this probably results from the topic of the writing task---the most important factor in choosing a career---in the present

study.

It is interesting to find that despite the predominance of Valuation in both sub-corpora, students in HAEC deployed a greater diversity of realizations of Valuation to illustrate the significance of things or events, such as *important* (196), *significant* (25), *vital* (12), *crucial* (10), *necessary* (8), *essential* (7), *indispensable* (6), *primary* (6). However, writers in the LAEC tended to concentrate on the adjective *important* to express the Valuation as this word accounts for nearly 77% of the total frequency of Valuation in the LAEC while it is only 49% of the total frequency of Valuation in the HAEC. As a result, students in the LAEC sometimes found it difficult to make their points clear and to appropriately validate their positions due to insufficient lexical proficiency, which in the end resulted in a weakened argumentative potential and lack of clarity.

### **3.4 Attitude mode and Attitude polarity across two sub-corpora**

With regard to the Attitude mode, most resources were encoded explicitly and directly rather than implicitly and indirectly and even no Evoked Attitude was detected. However, Evoked Attitude subtly exerts an impact on the reader's position, as Evoked Attitude is a primary mechanism by which a text insinuates itself into a reader's attitudes (Macken-Horarik, 2003, p. 299). The dominance of Inscribed Attitude in both corpora, on one hand, conforms to the nature of argumentative essays in that it requires writers to explicitly express their opinions and positions towards the topic. On the other hand, the overuse of Inscribed Attitude in both corpora may result from over-adaption in that Chinese students fully realize the stereotypical view of indirectness or vagueness in expressing ideas and might, therefore, deliberately and excessively adopt explicit Appraisal (Xie, 2016). More importantly, the underuse of Evoked Attitude may reflect the authors' insufficient language competence, thus making the text sound unduly bold and direct.

As to Attitude Polarity, it appears that both corpora have a similar distribution pattern as the proportion of Positive Attitude is much higher than that of Negative Attitude across HAEC and LAEC. For one thing, predominant Positive Attitude could be a strategy to align with readers in order not to provoke potential queries and disagreements. For another thing, Chinese culture favors the dominant Positive Attitude under the influence of traditional Confucian thought; Chinese students generally prefer collectivism over individualism and accord great respect to the ideas of superiors (Tylor & Chen, 1991; Carson, 1992). Nevertheless, preference for Negative Attitude in



the HAEC (20.37%) is significantly greater than that in the LAEC (12.84%). Moreover, compared with the LAEC, even though there is no statistically significant difference between the two groups, students in the HAEC applied more Negative Attitude values, which is consistent with the findings of Lv's (2015) research. More Negative Attitude values suggest that students in the HAEC introduce negative aspects to bring out more argumentative effect and make the focal point more prominent, thus exhibiting a critical way of thinking by means of instantiating opposing standpoints (Lv, 2015).

## 4 Conclusion

On the basis of Appraisal theory, this paper explores the deployment of Appraisal resources in English argumentative essays written by Chinese non-English major postgraduates. The findings of this study are intended as a contribution to the proper use of evaluative language in argumentative writing and to improving the argumentative writing skills of students. Moreover, this paper provides some pedagogical implications. Explicit instruction of Appraisal demonstration and the proper use of Appraisal resources in accordance with Appraisal theory should be included in the teaching of argumentative writing.

Despite our endeavor to explore the use of Appraisal resources in English argumentative essays, this study still has some limitations: small sample size, relative subjectivity in coding Appraisal resources and insufficient investigation of instantiations of Appraisal resources, especially how Appraisal items co-articulate with each other. Further study should make a more comprehensive and thorough study on deployment of Appraisal resources, especially the prosodic patterns and features of Appraisal resources in a larger sample of English argumentative essays or the deployment of Appraisal resources in argumentative essays of different topics.

## References

- [1] Biber, D., & Jones, J. (2009). Quantitative methods in corpus linguistics. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics* (Vol. 2): An international handbook (pp. 1286-1304). Berlin: Walter de Gruyter.
- [2] Geng, Y., & Wharton, S. (2016). Evaluative language in discussion sections of doctoral theses: Similarities and differences between L1 Chinese and L1 English writers. *Journal of English for Academic*

- Purposes*, 22, 80-91.
- [3] Johns, A. M. (1997). *Text, role, and context: Developing academic literacies*. Cambridge: Cambridge University Press.
  - [4] Lee, S. H. (2008). Attitude in undergraduate persuasive essays. *Prospect*, 23(3), 43–58.
  - [5] Liu, D. (2013). Intercultural Contrast Study of Engagement system in English and Chinese Argumentative Writing. *Foreign Language Research*, 3, 31-35.
  - [6] Liu, X. (2013). Evaluation in Chinese university EFL students' English argumentative writing: An appraisal study. *Electronic Journal of Foreign Language Teaching*, 10(1), 40-53.
  - [7] Liu, X., & Thompson, P. (2009). Attitude in students' argumentative writing: A contrastive perspective. *Language*, 1, 3-15.
  - [8] Lv, G. (2015). Appraisal Patterns in Chinese EFL Argumentative Essays. *Theory and Practice in Language Studies*, 5(4), 818.
  - [9] Martin, J. R., & White, P. R. R. (2005). *The language of evaluation: Appraisal in English*. New York, NY: Palgrave Macmillan.
  - [10] Nippold, M. A., Ward-Lonergan, J. M., & Fanning, J. L. (2005). Persuasive writing in children, adolescents, and adults: A study of syntactic, semantic, and pragmatic development. *Language, Speech and Hearing Services in Schools*, 36, 125–138.
  - [11] Swain, E. (2010). Getting engaged: Dialogistic positioning in novice academic discussion writing. In E. Swain (Ed.), *Thresholds and potentialities of systemic functional linguistics: Multilingual, multimodal and other specialised discourses* (pp. 291–317). Trieste: Edizioni Università di Trieste.
  - [12] Thomas, D. P., Thomas, A. A., & Moltow, D. T. (2015). Evaluative stance in high achieving Year 3 persuasive texts. *Linguistics and Education*, 30, 26-41.
  - [13] Wu, S. M., & Allison, D. (2003). Exploring appraisal in claims of student writers in argumentative essays. *Prospect*, 18(3), 71-91.
  - [14] Xie, J. (2016). Direct or indirect? Critical or uncritical? Evaluation in Chinese English-major MA thesis literature reviews. *Journal of English for Academic Purposes*, 23, 1-15.