Proceedings of the Workshop on Corpus-Based Research in the Humanities (CRH)

10 December 2015
Warsaw, Poland

Editors:
Francesco Mambrini
Marco Passarotti
Caroline Sporleder

Sponsors

CLARIN-PL
Common Language Resources and Technology Infrastructure

IPI PAN

Zespół Inżynierii Lingwistycznej

ISBN: 978-83-63159-19-1

Institute of Computer Science
Polish Academy of Sciences
ul. Jana Kazimierza 5
01-248 Warszawa, Poland

http://ipipan.eu/

Preface

The workshop on *Corpus-Based Research in the Humanities* (CRH) is a direct descendant of the workshop on *Annotation of Corpora for Research in the Humanities* (ACRH), which was held three times: in Heidelberg (5.1.2012), Lisbon (29.11.2012), and Sofia (12.12.2013).
All three editions were co-located with the international workshop on *Treebanks and Linguistic Theories* (TLT), a tradition which we continue with CRH.

The new name was motivated by the wish to change the focus slightly, towards corpus-based research in the humanities in general. While the earlier editions focused on questions related to annotation and a number of papers in the current proceedings do so as well, we wanted to visibly broaden the scope of the workshop, as even the earlier editions of the workshop had attracted submissions that did not centre on the question of annotation. In fact, there are many scholars in the humanities who use textual corpora in their everyday work but are not interested in or just do not need to deal with annotation issues. This is partly due to the fact that many corpora still lack linguistic annotation at all, thus requiring scholars to use just the raw text for their research purposes. As our original motivation for initiating the ACRH workshop series was to bring together the often separate communities of (digital) humanities and computational linguistics and to foster communication and collaboration between them, we felt that the focus on annotation in the name of the workshop was undermining our intention by discouraging humanities researchers working with corpora to submit papers.

In addition to changing the name of the workshop, we made several smaller adjustments. First, we included several scholars from digital humanities in the programme committee. While such a mixed committee is not entirely without problems due to different reviewing cultures in digital humanities and computational linguistics, we still believe this a step in the right direction for bringing both communities closer together and assessing submissions from both areas fairly. Second, this year's call asked for long abstracts (up to six pages) rather than full papers. This reflects common practices in the digital humanities better and did help to attract more proposals. Finally, we decided to organise the workshop on a biannual basis instead of an annual one in order to reduce the workload of the organisers and reviewers and avoid competing with too many similar workshops too frequently.

In total we received 17 long abstracts by authors from 12 different countries in Europe and South and North America. Each submission was reviewed independently by three members of the programme committee in a double-blind fashion. After the reviewing process, we accepted 11 submissions. One further submission was moved from TLT to CRH because it was a better fit to the topics of CRH than those of TLT. The overall acceptance rate was 70.6%. This reflects the fact that the average quality of the abstracts was high and most of them received favourable reviews. Another positive observation is that a number of the workshop speakers are promising young scholars.

We hope you will enjoy the workshop and the proceedings and wish to thank all authors who submitted papers, the 19 members of the programme committee, Reinhard Förtsch, who kindly agreed to give the invited talk, and last but not least the local and non-local organisers of TLT-14 and in particular the chair of the local organisation committee, Adam Przepiórkowski.

The CRH Co-Chairs and Organisers
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:
Monica Berti (Germany)
Federico Boschetti (Italy)
David Bouvier (Switzerland)
Neil Coffee (USA)
Lonneke van der Plas (Malta)
Dag Haug (Norway)
Neven Jovanovic (Croatia)
Mike Kestemont (Belgium)
John Lee (Hong Kong)
Alexander Mehler (Germany)
Roland Meyer (Germany)
Willard McCarty (UK)
John Nerbonne (The Netherlands)
Bruce Robertson (Canada)
Neel Smith (USA)
Uwe Springmann (Germany)
Melissa Terras (UK)
Sara Tonelli (Italy)
Martin Wynne (UK)

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:
Adam Przepiórkowski (chair)
Michał Ciesiołka
Konrad Gołuchowski
Mateusz Kopeć
Katarzyna Krasnowska
Agnieszka Patejuk
Marcin Woliński
Alina Wróblewska

Contents

Four elements to make the fifth
Integrating multiple disciplines in the iDAI.world

Reinhard Förtsch
Deutsches Archäologisches Institut, Berlin, Germany
E-mail: reinhard.foertsch@dainst.de

Abstract

The focus of this keynote is the larger picture of how different disciplines of classics and archaeologies from analogue times, when taken onto the digital level, change. And what that change looks like, when focused on the often misunderstood role of computer linguists inside the data ecosystem of classics and archaeologies.

In a digital age disciplines, at least when they share some common goals like in classics and archaeologies, are no more credible if they don't change their usual routines. Idf they don't change the status of splendid isolation they acquired - in different intensity - as a result of more than a centuries development in specialisation. The disciplines at least are called up - through new horizons of technological possibilities - to change the way they interact or at least restart interdisciplinary interaction they once had. It has long been observed that, beside the usual common intellectual goals of the disciplines, new hubs of interdisciplinarity are being created by shared technologies or at least technological paradigms of digital data-processing. I use »technological« not in the wrongful sense of being the opposite of »intellectual«, the technological will not only have to be digestes and integrated intellectually, but its inherent und implicit intellectual power will have to be unlocked. A common goal could be, lets say, the analysis of the rise of the Athenian "empire" in the 5th Century BCE (known in Greek as "Pentekontaetia", the "fifty-year period"), for which many different sources exist, like literary texts, inscriptions, monuments and objects, and in which Classical Philology, Epigraphy, Ancient History and last not least Classical Archaeology have their stakes. Some technological hubs could be GIS and Spatial Analysis Technologies for textual and material sources, Textmining, Datamining and Pattern Recognition, Statistics and the like.

I will look at this from the perspective of the German Archaeological Institute (DAI), an operation founded in 1829 in Rome that geographically was centered around the Cultures of the Mediterranean Zone, North European Prehistory, then grew into Egypt and the Orient, until it also touched parts of Asia, Mongolia, the

Northern and Middle Americas until Peru and Easter Islands and has a certain tendency into deeper Africa.

Organisationally this is performed in 20 different departments in about 17 different cities in 14 countries.

In terms of scientific disciplines the DAI covers a range of Archaeologies (Prehistory, Classical Archaeology, Chinese Archaeology, Sudan Archaeology) and Ancient History up to Epigraphy, Natural Sciences and IT from Object Database Technologies up to Remote Sensing.

The whole IT-architecture of such an operation has to reflect its integrity, and this is what the so-called idai.world, as we call it, is about. The idai.world is a modular research data environment, conceptualized as a generic digital infrastructure for the projects of the DAI. So the question I often hear is: »what for do we need researchers with background in digital philology or linguistics«, or, mostly from people outside the DAI, what for does DAI need Computer Linguistics at all? »Outisde the DAI« also covers the area of many kinds of reviewers for funding proposals. We sometimes experienced a deep understanding problem on their side, as the reviews show, and this also means for us that in the past we often also had a problem of making clearer cases for what we want to achieve and already did achieve. To show the role of language resources interacting with other areas and technological paradigms, I will look at some areas of that idai.world that do matter for understanding how language resources integrate with material object resources.

idai.vocab and idai. thesauri are for general language processing purposes. idai.vocab is condensing the multilinugual (up to Russian, Farsi and Arabic) domainspecific vocabularies of DAI's scientific activities. idai. thesauri is structuring the traditional and the new upcoming vocabularies of DAI's library services and information systems well down into projects specialised on certain excavations and research projects. It is not too surprising that the thesaurus paradigm is also playing into the way idai.vocabulary is organized. While we develop this, we find more an more methodological interplay also with idai.chronontology and the way chronological systems have been integrated also into thesauri like the Getty “Art and Architecture Thesaurus” (AAT). Behind all three, a thesaurus backbone has to be set up to which a mapping will have to take place, notwithstanding the representation of the data as Linked Open Data also.

Beside these infrastructural level, a unifying project like “Hellespont” can show how material and language resources as well as some of the mentioned paradigms do come together centered around Thucydides' account of the Pentekontaetia, the core description of what has be become the »Classical Age« in later tradition.

The intention is show how the sum of different elements can be more than their number, in the same way as, to quote a famous statement by Jimmy Page about the music of Led Zeppelin, four musicians playing at their best create an alchemy that makes for the “fifth element”.

Unsupervised regularization of historical texts for POS tagging

Fabian Barteld, Ingrid Schröder and Heike Zinsmeister

Institut für Germanistik
Universität Hamburg
E-mail: firstname.lastname@uni-hamburg.de

Abstract

This paper presents an unsupervised method to reduce spelling variation in historical texts in order to mitigate the problem of data sparsity. In contrast to common normalization techniques, the historical types are not mapped to corresponding types in a standardized target language (e.g. normalizing Early New High German to Modern German). Consequently, no additional resources such as manually normalized data, parallel texts or a dictionary of the target language are needed. Furthermore, our approach does not use any annotation and is thus not dependent on the existence of annotated data. We evaluate the usefulness of this approach using POS tagging.

1 Introduction

In the DFG-funded project *Reference Corpus Middle Low German/Low Rhenish (1200-1650)*, a corpus of Middle Low German (GML) and Low Rhenish texts annotated with fine-grained parts of speech and lemmas is created for linguistic and philological research.[1] The annotation must be of high quality for the corpus to be accepted as a reference corpus by relevant research communities. In order to speed up the annotation process and to ensure consistent annotation, automatic pre-annotations and error detection methods are used. However, the initial usefulness of statistical approaches is limited due to data sparsity. GML is a low-resourced historical dialect and it exhibits a great deal of spelling variation, as is common for historical texts; furthermore, the corpus contains texts from different time periods, dialect regions and domains, which amplifies the spelling variation in the overall corpus.

This paper contributes to a line of research that uses normalization methods to overcome the problem of data sparsity induced by spelling variations when training

[1] The project is a cooperation between the University of Hamburg (Ingrid Schröder) and the University of Münster (Robert Peters). More information about the project can be found at http://referenzkorpus-mnd-nrh.de.

a POS tagger on historical texts. We present an unsupervised technique to reduce spelling variation that requires no target language – unlike common normalization techniques used for historical texts (e.g. normalizing Early New High German to Modern German) – nor does it use any annotation. Consequently, this technique is not dependent on additional language resources such as manually normalized or annotated data, the existence of parallel texts or a dictionary of the target language.

2 Related work

Spelling variation is problematic for statistical NLP applications, as it increases data sparsity. This issue can either be addressed by adapting the NLP tools to handle the variation or by a pre-processing step that reduces the variation before an off-the-shelf implementation of a tool is applied to the data. Here, we concentrate on the second approach, often called *normalization*.

Most of the work on normalizing historical texts uses a standardized modern language as a target and treats normalization as some kind of "transformation of a historical form into its modern equivalent" [2]. One problem with normalization techniques of this kind – including unsupervised ones – is that they always require additional resources to define the target language. These may be parallel texts used to create training pairs of historical variant and target language form [3], an unannotated corpus and a lexicon of the target language [17] or, at the least, a lexicon that defines the target language [13]. This is unproblematic for historical variants that are closely related to a standardized modern language. In the case of GML, the problem is that there is no such standardized language. The modern version of GML – Low German – is still merely a group of spoken dialects without a standardized written form. For this reason, we are pursuing a different approach that we call *regularization*[2] to distinguish it from normalization with a target language.[3]

Regularization can be defined as the *conflation* of all spelling variants into one target form. Normalization as defined above is then a special instantiation of regularization in which the target forms are defined externally by a target language.[4]

In [14] and [16], two approaches are presented whereby possible spelling variants of one word are induced directly from a historical corpus in an unsupervised fashion. However, both of these approaches rely on existing annotations (lemmas [14] or POS tags [16]). As we are creating the first annotated corpus for GML, the amount of annotated data available is very limited. This motivates us to explore an approach toward regularization that uses tokenized, unlabelled texts as input without any additional resources for the selection of conflation candidates in the texts.

[2] We would like to thank the anonymous reviewers for their comments on the terminology.

[3] Note that there are similar approaches that do not make this terminological distinction, e.g. [16] refer to normalization without using a target language.

[4] However, in cases in which the target language exhibits spelling variation as well (e.g. the variation between *-s* and *-es* as the genitive marker in Modern German), normalization differs from regularization as spelling variants might be mapped to different target forms in the former approach.

Normalization is used not only to reduce the spelling variation in a corpus but also to enable the general usage of existing resources for the target language. This re-usage of existing resources is not possible when using regularization. However, there are many corpus-related tasks that do not depend on additional resources that can benefit from the reduction of spelling variation, e.g. keyword statistics [1] and automatic error-detection with systems like the one described in [6].

We evaluate our approach by testing the impact of regularization on the accuracy of POS tagging in GML texts. Related work has shown that the POS tagging of Middle High and Early New High German texts can be improved by reducing the spelling variation. In [7], Dipper compares the tagging of diplomatic transcriptions of Middle High German texts with that of normalized variants for which the normalization was realized with handwritten rules and manually corrected. She reports improvements in per-word accuracy for POS tagging between 3.75% and 4.81% for different parts of the corpus. In [16], Logačev et al. conflate possible spelling variants before training and applying a POS tagger on Early New High German texts. They report mixed results with a maximum of 1.6% improvement in per-word accuracy, including a decrease in accuracy for one text (-0.2%).

3 The data

In this study, we use four texts (see Table 1; full bibliographical information is given in the bibliography) from the GML Reference Corpus that have been manually corrected for tokenization and sentence boundaries. We employ a simplified version of the transcription in which abbreviations are expanded, among other aspects, thereby already reducing spelling variation. For the experiments, all tokens have been lowercased.

Name	Year	Domain	Type	Tokens	Types
Johannes	~1480	religious texts	manuscript	19645	2305
Griseldis	1502	literature	print	9062	2251
OldenbSSP	1336	law	manuscript	21800	2731
SaechsWeltchr	1st half 14th c.	arts	manuscript	18215	3255

Table 1: The texts used for the experiments

All texts are from the same dialect region (North Low Saxon). They only differ in the year of writing/printing, the textual domain and the medium (print or manuscript). As alluded to above, they exhibit a large number of spelling differences, e.g. the 3SG.PST.IND of the verb *blîven* '(to) stay' appears as *blef* (6x) in Johannes and as *bleff* (3x) in Griseldis (see Examples 1 and 2).

(1) vnde nemande bleff vngewenet .
and nobody.SG.NOM stay.3SG.PST.IND NEG-PTCP-cry-PTCP .
'and nobody could help but crying' (Griseldis)

(2) vnde blef vp eme
and stay.1SG.PST.IND upon he.3SG.M.DAT
'and stayed by him' (Johannes)

Such spelling differences are not limited to instances between texts. Spelling variation is also observed within individual texts, as can be seen with the example *anbegin* 'beginning', which appears as *anbegin* and as *anbegyn* in Griseldis and in a third version (*anbeginne*) in Johannes. In the current experiment, we ignore spelling variation that appears within one text and concentrate on regularization of the variation between texts.

For the evaluation of the POS tagging, Johannes and Griseldis (cf. Table 1) have been manually tagged with the POS tagset HiNTS, which is an adapted version of HiTS [8], a tagset created for the GML reference corpus that consists of 105 tags.[5] Both texts contain a comparable number of types, although Johannes is more than twice as long as Griseldis. The reason for this is that the gospel of Johannes is a rather repetitive and formulaic text. This is reflected in the better POS-tagging accuracy results for Johannes in a 20-fold cross-evaluation (with whole sentences) on each text individually using RFTagger[6] [19]. The accuracy is 90.0% ± 1.5 for Johannes and 83.9% ± 2.7 for Griseldis. When training the tagger only with "out-of-domain" data, (i.e. on the other text), we observe for both texts a drop in the mean accuracy of about 15% (cf. Table 2) on the same 20 parts.

	% correct	% correct	% correct	% unknown
		(known)	(unknown)	
Johannes	73.5 ± 1.9	83.2 ± 2.4	48.7 ± 3.5	28.3 ± 3.1
Griseldis	69.3 ± 3.3	80.6 ± 2.3	46.6 ± 6.6	33.0 ± 3.3

Table 2: Per-word tagging accuracy trained on out-of-domain data

The differences in performance can be easily explained by the amount of unknown words, which is higher when only "out-of-domain" data is used for training and which is also higher for Griseldis in general. As some of the unknown words are due to spelling variations, we expect that substituting an unknown type with a known spelling variant will improve the performance of the POS tagger.

[5] At the time of writing, the GML corpus is still under construction, and the POS annotation used for evaluation in our experiments is pre-final.

[6] RFTagger performed second-best in initial tests, slightly below HunPos [10]. Both taggers outperformed SVMTool [9] and CRFSuite [18] with standard POS-tagging features. RFTagger has been chosen over HunPos for the presented experiments, as morphological categories are added to the POS tags in further steps. However, our regularization technique improved POS-tagging accuracy for all of these taggers.

4 Regularization with string and context similarity

In this section, we present a language-independent approach to unsupervised regularization using string and context similarity.

For string similarity, we use the similarity measure *Proxinette* [11, 12]. Proxinette was designed to measure the morphological similarity of lexemes and to find morphologically related words in a lexicon. We use Proxinette in the variant described in [12], where character n-grams are the only features.

The intuition behind Proxinette is that the more character n-grams two types share, the more similar they are. The n-grams are weighted by their frequency in the corpus: More frequent n-grams contribute less to the similarity. In the original version of Proxinette, the character n-grams have a minimum length of three. However, this leaves spelling variants such as *yck* and *ic* 'I' unconnected. Therefore, we vary the minimal n-gram length as a parameter in our experiments (Ngram). Proxinette is computed based on a graph and is consequently efficient and scalable and can be employed to compare many types. Proxinette returns the similarity of two types as a number between 0 and 1: 0 means no similarity (i.e. no shared n-gram), while higher values denote a greater similarity.

We use Proxinette similarity to select the known types that are most similar (i.e. have the maximal Proxinette similarity of all known types) as conflation candidates for an unknown type. We restrict this choice by a threshold for the similarity, which is also varied as a parameter in our experiments (Prox). A higher threshold reduces the overall number of types for which conflation candidates are generated.

The conflation candidates are then filtered using Brown clusters [4], allowing only the candidates chosen by Proxinette that fall into the same cluster. The number of Brown clusters is varied as an additional parameter [5] (Brown). When more than one conflation candidate exists, one is chosen at random.

For Proxinette, we use our own implementation. The Brown clusters are computed using the implementation described in [15].[7] As a fourth parameter, we vary the amount of data utilized to compute the Proxinette similarity and the Brown clusters (Data): We either use only Johannes and Griseldis (base) or all texts presented in Table 1 (all).

5 Results

In this section, we present the results of a POS-tagging experiment using the regularization technique described in the previous section. The baseline is given by tagging the raw texts (cf. Table 2). As with the baseline, we tag the texts Johannes and Griseldis with RFTagger trained on the other text. In contrast to the baseline, the text to be tagged is first regularized by conflating unknown types with similar known types.

[7] The source code is available at `https://github.com/percyliang/brown-cluster`.

	Text	Ngram	Prox	Brown	Data	% correct
best	Johannes	1	0.000001	125	all	75.7 ± 1.6
	Griseldis	1	0.02	25	base	71.4 ± 3.0
combined	Johannes	2	0.000001	50	all	75.6 ± 1.7
best	Griseldis	2	0.000001	50	all	70.5 ± 3.1

Table 3: Best per-word tagging accuracies on regularized data

Table 3 shows the best results of the experiment. All improvements are significant.[8] The parameter values for the best results are divergent. Especially when only the base texts are used, the numbers of Brown clusters yielding the best results differ: For Griseldis, 25 and 50 are the best options; neither are among the three best values for Johannes. However, when using all texts, 50 Brown clusters lead to the second-best results for Johannes and Griseldis. For Johannes, this result is almost as good as the best result. For Griseldis, the difference between the optimal parameters and the combined best result is greater, but it still leads to an increase of about 1% for the tagging accuracy.

Inspecting substituted types and types that are not substituted more closely points toward further steps to improve the system. For instance, Griseldis has at least four variants of *ik* ‘I’: *ik*, *yck*, *yk* and *ick*. Johannes exhibits at least three variants: *ik*, *ic* and *jk*. All of these variants are in the same Brown cluster (50 clusters, all texts). However, when regularizing Griseldis, only *ick* gets substituted with *ic*. The reason for this is that string and context similarity are modeled separately, and only the most similar types according to Proxinette can be conflated with types in the same Brown cluster. For *yck* and *yk*, there are types that are more similar (according to Proxinette) than one of the actual spelling variants appearing in Johannes. By allowing types other than the most similar types as conflation candidates these variants could be conflated.

Table 4 presents conflated types in Johannes that appear more than 10 times in the text. An ‘x’ in the fourth column indicates whether the conflations are actually spelling variants. An examination of wrongly conflated types indicates that POS tagging can benefit from these conflations as well. The wrongly conflated types can be divided into four categories:

(1) Morphologically connected types that belong to the same part of speech (derivation or inflection) such as *loue* ‘believe, praise, promise’ (1SG.PRS.IND (among others)) – *louen* ‘believe, praise, promise’ (INF (among others)); (2) Types that belong to the same part of speech such as *ioden* ‘Jews’ – *boden* ‘messengers’ and *schare* ‘cohort’ – *hare* ‘hair’; (3) Morphologically connected types that belong to different parts of speech (derivation) such as *lose* ‘(to) loosen’ (but may also be: ‘replacement, redemption’) – *loszheyt* ‘flippancy, devilment’; and (4) Types that

[8] The significance was verified using a paired t-test with the tagging results on the 20 parts used for the cross-evaluation. The significance level was set to 0.05 with adjustment to 4 tests using Holm’s method.

Freq.	Type	Conflation	Spelling variant	Translation
11	hochtijt	hochtid	x	‘celebration’
12	hijr	hir	x	‘here’
12	scole	schole	x	‘shall’
13	scrift	schrifft	x	‘writing’
14	echter	echtes		‘again’ (diff. morphology)
15	ghecomen	komen		‘come’ (diff. inflection)
15	iiij	iii		Roman numerals
15	schare	hare		‘cohort’; ‘hair’
16	loue	louen		‘believe, praise’ (diff. inflection)
17	efte	eft	x	‘or’
17	sic	sick	x	‘herself, himself... ‘
18	neman	nemande	x	‘nobody’
21	uadere	vadere	x	‘father’
22	jk	jodoch		‘I’;‘but’
27	comen	komen	x	‘come’
30	lef	leff	x	‘beloved’
43	scolen	scholen	x	‘shall’
61	ioden	boden		‘Jews’; ‘messengers’
67	uader	vader	x	‘father’

Table 4: Conflations of unknown types in Johannes (2; 0.000001; 50; all)

belong to different parts of speech such as *jk* ‘I’ – *jodoch* ‘but’.

For POS tagging, only the third and fourth types are harmful, as the other conflations can still help the tagger to predict the right POS tag. Therefore, for other tasks – e.g. fine-grained POS tagging including morphological tags and lemmatization – the conflation needs to be stricter than for simple POS tagging.

6 Conclusion

In this paper, we have presented a language-independent, unsupervised regularization approach that utilizes string and context similarity and does not make use of any resources other than unlabelled, tokenized texts from the language to be regularized. Applying this approach to POS tagging, we were able to increase the per-word accuracy for two historical non-standardized GML texts by about 2% with the optimal parameters and 2% for one text and 1% for the other with parameters giving the combined best result. These are small but still statistically significant improvements.

An analysis of the conflations shows that the algorithm misses some spelling variants because only the most similar types according to Proxinette are considered as conflation candidates. To further improve the algorithm in this direction, we plan to directly integrate the syntactic context into the process of selecting conflation candidates instead of only using it as a filter.

In the study presented in this paper, we divided the data into two parts: one part that defined the target forms and another that was regularized toward these target forms. Such a division comes naturally when using regularization as a preprocessing step for supervised algorithms, as in our evaluation setup. However, for applications that are based on unsupervised algorithms or simple text statistics such as keyword analysis, it would be helpful to avoid such a division. We expect that this would also improve supervised approaches, as the training data would be regularized as well. We did not investigate this in the current study, but our future research will examine this issue.

Additionally, it should be noted that the usefulness of our approach is not limited to POS tagging: All applications that rely on consistent spellings will benefit from regularization. We will explore this in further experiments.

Resources

This paper is created reproducibly using org-mode (`http://orgmode.org`). The org-files, including all the scripts needed to reproduce the experiments, are available at github (`https://github.com/fab-bar/paper-CRH4`). This version also includes an appendix with additional data from the experiments.

Acknowledgements

The work of the first and second authors has been funded by the DFG. We would like to thank the anonymous reviewers for their helpful remarks, Sarah Ihden for help with the translations from Middle Low German and Claire Bacher for improving our English. All remaining errors are ours.

Primary data

Johannes *Buxtehuder Evangeliar*. GML manuscript from about 1480. Transcribed in the DFG-funded project “Referenzkorpus Mittelniederdeutsch / Niederrheinisch (1200-1650)”.

Griseldis *Griseldis / Sigismunda und Guiscardus*. GML print of two tales from 1502. Transcribed in the DFG-funded project “Referenzkorpus Mittelniederdeutsch / Niederrheinisch (1200-1650)”.

OldbSSP *Oldenburger Bilderhandschrift des Sachsenspiegels*. GML manuscript from 1336. Transcribed in the DFG-funded project “Referenzkorpus Mittelniederdeutsch / Niederrheinisch (1200-1650)”.

SaechsWeltchr *Sächsische Weltchronik*. GML manuscript from the first half of the 14th century. Bremer Hs. der Rezension B (Hs. 16). Transcribed in the

DFG-funded project “Referenzkorpus Mittelniederdeutsch / Niederrheinisch (1200-1650)”.

References

[1] Baron, Alistair, Rayson, Paul and Archer, Dawn (2009) Word Frequency and Key Word Statistics in Corpus Linguistics. *Anglistik: International Journal of English Studies*, 20, 41–67.

[2] Bollmann, Marcel, Dipper, Stefanie, Krasselt, Julia and Petran, Florian (2012) Manual and Semi-automatic Normalization of Historical Spelling – Case Studies from Early New High German. In *Proceedings of the 11th Conference on Natural Language Processing (KONVENS 2012), LThist 2012 workshop*, pp. 342–350, Vienna, Austria.

[3] Bollmann, Marcel, Petran, Florian and Dipper, Stefanie (2014) Applying Rule-based Normalization to Different Types of Historical Texts – An Evaluation. In Vetulani, Zygmunt and Mariani, Joseph (eds.) *Human Language Technology Challenges for Computer Science and Linguistics*, pp. 166–177, Heidelberg et al.: Springer International Publishing.

[4] Brown, Peter F., deSouza, Peter V., Mercer, Robert L., Della Pietra, Vincent J. and Lai, Jenifer C. (1992) Class-based N-gram Models of Natural Language. *Computational linguistics*, 18, 467–479.

[5] Derczynski, Leon, Chester, Sean and Bøgh, Kenneth S. (2015) Tune your Brown Clustering, Please. In *Proceedings of Recent Advances in Natural Language Processing (RANLP 2015)*, pp. 110–117, Hissar, Bulgaria.

[6] Dickinson, Markus and Meurers, Detmar (2003) Detecting Errors in Part-of-Speech Annotation. In Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL-03), pp. 107–114, Budapest, Hungary.

[7] Dipper, Stefanie (2011) Morphological and Part-of-Speech Tagging of Historical Language Data: A Comparison. *Journal for Language Technology and Computational Linguistics*, 26, 25–37.

[8] Dipper, Stefanie, Donhauser, Karin, Klein, Thomas, Linde, Sonja, Müller, Stefan and Wegera, Klaus-Peter (2013) HiTS: Ein Tagset für historische Sprachstufen des Deutschen. *Journal for Language Technology and Computational Linguistics*, 28, 85–137.

[9] Giménez, Jesús and Màrquez, Lluís (2004) SVMTool: A General POS Tagger Generator based on Support Vector Machines. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC’04)*, pp. 43–46, Lisbon, Portugal.

[10] Halácsy, Péter, Kornai, András and Oravecz, Csaba (2007) HunPos: An Open Source Trigram Tagger. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions (ACL '07)*, pp. 209–212, Stroudsburg, PA, USA.

[11] Hathout, Nabil (2009) Acquisition of Morphological Families and Derivational Series from a Machine Readable Dictionary. In Montermini, Fabio, Boyé, Gilles and Tseng, Jesse (eds.) *Selected Proceedings of the 6th Décembrettes*, pp. 166–180, Somerville, MA: Cascadilla Proceedings Project.

[12] Hathout, Nabil (2014) Phonotactics in Morphological Similarity Metrics. *Language Sciences*, 46, 71–83.

[13] Hauser, Andreas W. and Schulz, Klaus U. (2007) Unsupervised Learning of Edit Distance Weights for Retrieving Historical Spelling Variations. In Mihov, Stoyan and Schulz, Klaus U. (eds.) *Proceedings of the First Workshop on Finite-State Techniques and Approximate Search*, pp. 1–6, Borovets, Bulgaria.

[14] Kestemont, Mike, Daelemans, Walter and De Pauw, Guy (2010) Weigh your Words – Memory-based Lemmatization for Middle Dutch. *Literary and Linguistic Computing*, 25, 287–301.

[15] Percy Liang (2005) *Semi-supervised Learning for Natural Language*. MEng thesis, Massachusetts Institute of Technology.

[16] Logačev, Pavel, Goldschmidt, Katrin and Demske, Ulrike (2014) POS-tagging Historical Corpora: The Case of Early New High German. In Henrich, V., Hinrichs, E., de Kok, D., Osenova, P., and Przepiórkowski, A. (eds.) *Proceedings of the Thirteenth International Workshop on Treebanks and Linguistic Theories (TLT13)*, pp. 103–112, Tübingen, Germany.

[17] Mitankin, Petar, Gerdjikov, Stefan and Mihov, Stoyan (2014) An Approach to Unsupervised Historical Text Normalisation. In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage (DATeCH '14)*, pp. 29–34, New York, NY, USA.

[18] Naoaki Okazaki (2007) CRFsuite: A Fast Implementation of Conditional Random Fields (CRFs) (URL: http://www.chokkan.org/software/crfsuite).

[19] Schmid, Helmut and Laws, Florian (2008) Estimation of Conditional Probabilities with Decision Trees and an Application to Fine-grained POS Tagging. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1 (COLING '08)*, pp. 777–784, Manchester.

A semi-automatic method for thematic classification of documents in a large text corpus

Łukasz Borchmann[†], Filip Graliński[*],
Rafał Jaworski[*], Piotr Wierzchoń[†]

Adam Mickiewicz University
[*]Faculty of Mathematics and Computer Science
[†]Institute of Linguistics
[*]ul. Umultowska 87, 61-614 Poznań, Poland
[†]al. Niepodległości 4, 61-874 Poznań, Poland
borchmann@rainfox.org, {filipg,rjawor,wierzch}@amu.edu.pl

Abstract

One of the essential steps in the analysis of large document collections is the thematic classification of those documents. As technical capabilities for document acquisition and storage have increased significantly, text corpora have grown to sizes which make manual analysis by humans infeasible. For that reason, it is necessary to design processes of document analysis which take this problem into account. This paper describes a method for the thematic classification of a large collection of documents. If done manually, this process would require an enormous amount of human work. Instead, a semi-automatic method based on keyword classification is proposed and evaluated.

1 Introduction

For the making of a thematic classification of texts according to branch of knowledge, those texts were chosen which appear in the system of Polish digital libraries [8, 1]. This system is largely based on *dLibra* software, which was first deployed in 2002 by the Digital Library of Wielkopolska. *dLibra* enables the convenient cataloguing of scanned library items together with appropriate bibliographic description (with fields such as title, subject, description, format, etc.).

The metadata system, however, does not offer a field for *branch of knowledge*. Such information is approximated to a certain extent by the *topic and keywords* field; however, firstly, this field is not present in all the Polish digital libraries, and secondly, the descriptions contained in this field are not taken from a predetermined catalogue of branches. As a result, these descriptions can be regarded as chaotic, or under another interpretation – unreliable. To address this problem, it was decided to perform an experiment: the research task involved assigning appropriate branches of knowledge to documents retrieved from *dLibra*, the assignment being made on the basis of the procedure described in Section 2. This procedure was based only on metadata; by contrast, in Section 3.1 we present the preliminary results of experiments with automatic clustering performed on the *content* rather than on metadata.

To begin with, a modified version of the Universal Decimal Classification was adopted. This classification, introduced in the early 20th century, serves the purpose of ordering library collections. It has nine main classes, as listed in Table 2.

For the conduct of the test, an extension of each of the branches was defined, assigning each of them a two-digit code as shown in Table 4.

2 Manual assignment of documents

The XML schema used in OAI-PMH describes the complex type element *oai_dcType* as containing one or more elements such as *title*, *creator*, *subject*, *description*, *publisher*, *contributor*, *date*, *type*, *format*, *identifier*, *source*, *language*, *relation*, *coverage*, *rights*.

Each of the above elements may occur multiple times within the structure, and in fact all are used in such a manner (with a frequency ranging from 3% for "identifier" to 85% for "subject"). The practice of use of each element varies between and within libraries, but by an analysis of random samples, "title", "subject" and "description" were identified as useful for the purposes of determining the thematic fields of publications. The example data structure obtained by parsing selected elements of the XML file may appear as in Table 1 (the underlined words are essential for thematic classification of the text).

Lack of standardization is a particularly onerous problem when it comes to "date", which had to be normalized before further work, because only texts from a specific period were intended to be analysed. The algorithm developed for this purpose was able to parse 93% of all textual date-times into a time stamp, based on the content of multiple elements. Date normalization is a complex problem in data sets like the one described, and is not the topic of this paper, so it is not discussed further.

Selected elements from publications were then morphosyntactically tagged using *LanguageTool*, which is based on a set of disambiguation rules (defined in the `resource/pl/disambiguation.xml` file) and *Morfologik 2.0 PoliMorf* (an open-source morphological dictionary of Polish). In cases of ambiguity the first proposed interpretations were chosen. Next, all the unique lemmas were inserted into the database in one table along with the tags provided by the LanguageTool, while another table holds a one-to-many relation between lemmas and correspond-

identifier	oai:ebuw.uw.edu.pl:148209
date	1997-
	1837-1938
title	Pamiętnik Towarzystwa Lekarskiego Warszawskiego
	Pamiętnik Towarzystwa Lekarskiego Warszawskiego oraz Rocznik Zarządu Towarzystwa Lekarskiego Warszawskiego za Rok ... T. 133 = nr 1 (1997)-
	Pamiętnik Towarzystwa Lekarskiego Warszawskiego oraz Rocznik Zarządu Tow. Lek. Warsz. za R. ... T. 127, 1931-[?]
	Pamiętnik TLW T. 142 = nr 10 (2006) [?]
	Pamiętnik Towarzystwa Lekarskiego Warszawskiego. T. 126, 1932
subject	Medycyna – Polska – 1990- – badania – czasopisma.
description	Warszawa
	Częstotliwość: rocznik.
	Rocznik Zarządu Warszawskiego Towarzystwa Lekarskiego początkowo drukowany wewnątrz zeszytów Pamiętnika, następnie dodawany osobno do niektórych tomów, najprawdopodobniej od t.105 (1909).
	Od 1931, T. 127, do [?] tytuł: Pamiętnik Towarzystwa Lekarskiego Warszawskiego oraz Rocznik Zarządu Tow. Lek. Warsz. za R. ...

Table 1: Example data structure obtained by parsing selected elements of the XML file

0	Science and knowledge. Organization. Computer science. Information. Documentation. Librarianship. Institution. Publications
1	Philosophy. Psychology
2	Religion. Theology
3	Social Sciences
5	Mathematics. Natural sciences
6	Applied sciences. Medicine. Technology
7	The arts. Recreation. Entertainment. Sport
8	Language. Linguistics. Literature
9	Geography. Biography. History

Table 2: The main UDC classes

ing publications.

The next stage began with semi-automatic filtering of the lemma table against lists of Polish stop words. Some human-supervised exclusions were made based on part of speech (for example function words not included on the stop word lists were additionally marked as useless for the purpose of determining the thematic field) and lemma length (lemmas shorter than 5 characters were excluded). Moreover, a review of the remaining proper nouns led to all of them being omitted. Around 20 000 lemmas remained after the above procedures with the most frequent having approximately 450 000 occurrences within the "title", "subject" and "description" fields.

Next, roughly 5 000 most frequent lemmas were assigned by humans to the appropriate Universal Decimal Classification (UDC) codes. Because some of them match many categories, and the proper assignment of others is not trivial, the same task was performed by three people independently. A simple Web application was made for this purpose, allowing users to enter the UDC code into the input next to a lemma or to check a record as unspecific, which means that it was impossible to assign it to any of the categories. About 15% of lemmas were marked as unspecific. Table 3 presents the similarities and differences between independent assignments.

For the purposes of further work, only those lemmas were used for which there was agreement between all of the UDC codes assigned, or for which two codes were exactly the same while the third belonged to the same class. All of the possibly unspecific lemmas were omitted.

Thereafter, the text layer of each corresponding DjVu document was decoded using the `djvutxt` command, and saved in the appropriate directory.

3 Document clustering and category analysis

This section describes an experiment performed on the data described in Section 2. It was conducted in order to determine, which documents are similar in terms of

	$P1_n = P2_n$	$P1_n = P3_n$	$P2_n = P3_n$
all classes	40%	33%	34%
main UDC classes	58%	55%	53%

	$P1_n = P2_n = P3_n$	$P1_n \neq P2_n \neq P3_n$
all classes	23%	47%
main UDC classes	40%	35%

Table 3: The similarities and differences between independent assignments

vocabulary and language and how these similarities are reflected in the UDC categories assigned to these documents. The experiment was inspired by research in the field of automatic content-based document clustering (see [2]) and topic-modelling (e.g. [7]). The list of selected UDC categories used in this experiment is presented in Table 4. The experiment consisted in performing document clustering on a set of 67 296 documents divided into 57 UDC categories.

3.1 Detailed experimental procedure

3.1.1 Pre-processing

In the first step, the whole corpus of documents underwent pre-processing routines. These included: (1) selecting at most 100 documents per category at random, in order to balance the data (over 90% of documents in the raw corpus fell into only three categories: 32 – politics, 65 – communication and 91 – geography); (2) removing function and stop words from the texts; (3) changing words to lower case; (4) truncating words to at most 7 characters, in order to avoid distinction between inflected/derived forms of the same lexeme.

3.1.2 Document clustering

The pre-processed documents were then converted to numeric vectors by means of the HashingVectorizer algorithm described in [3]. This was done to enable document clustering. The document vectors were clustered (into 100 groups) using two different algorithms: mini-batch k-means (see [4]) and spectral clustering (described in [5]). These steps resulted in the assignment of all the documents to 100 automatically generated clusters. These clusters, however, did not provide synthetic information about the similarity of categories. For that reason, a separate analysis was performed on the UDC categories.

Firstly, the categories were converted to numeric vectors using the following scheme. Let n_{docs} be the total number of documents assigned manually to a given category. Let f_i be the number of documents in the category falling into the i-th

UDC	Code	Description	UDC	Code	Description
01	BIB	Bibliography	54	CHM	Chemistry
06	ORG	Organizations	57	BIO	Biology
14	PHI	Philosophy	61	MED	Medical
15	PSY	Psychology	62	ENG	Engineering
16	LOG	Logic	63	AGR	Agriculture
23	RIN	Indian	65	COM	Communication
31	STA	Statistics	69	BUI	Building.
32	POL	Politics	72	ARC	Architecture
34	LAW	Law	76	GRA	Graphics
35	GOV	Government	79	SPO	Sport
37	EDU	Education	81	LAN	Linguistics
39	CUL	Culture	82	LIT	Literature
50	ENV	Environment	91	GEG	Geography
51	MAT	Mathematics	93	HIS	History
53	PHY	Physics	94	GNR	General

Table 4: UDC categories

Size	Categories
11	COM, ECO, EDU, GEG, GEN, GLN, GOV, MAT, POL, REF, RIN
8	AST, GNR, LIT, MUS, PSY, RFE, SPI, ZOO
6	AGR, CUL, LAW, SAF, SPO, STA
4	CHT, ENG, LAN, MED
3	BIO, BOT, GEO

Table 5: Largest groups of categories

automatically computed document cluster. The category vector has the form: $[\frac{f_i}{n_{docs}}]$ for i in the range $[1,100]$, as there are 100 document clusters.

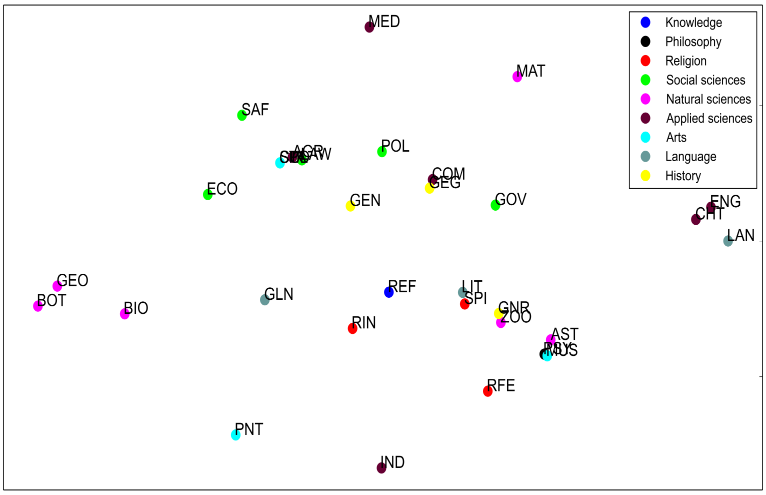
For each pair of category vectors, the cosine similarity (described in [6]) is computed. The distances between categories are then visualized on a 2D plane (see Section 3.2).

3.2 Experimental results

Figure 1 presents the distances between UDC categories computed after mini-batch K-means document clustering. UDC main categories are marked with different colours. It is expected that categories of the same colour will form groups. The largest groups of categories shown in Figure 1 are described in Table 5.

The results show that documents of certain categories tend to be similar. For example, it is virtually impossible to distinguish between the categories Biology,

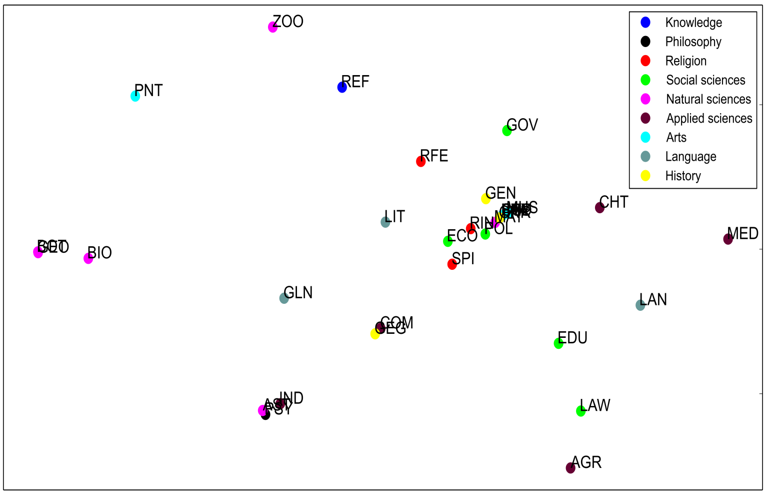
Figure 1: Categories after mini-batch K-means document clustering.



Botany and Geology, i.e. documents assigned to these three categories are very similar. The same can be concluded for the category pair Engineering and Chemical Technology and for Communication/Transport Industries and Geography.

Figure 2 presents the categories after spectral document clustering. The results obtained by this method of clustering also indicate the closeness of the categories Biology, Botany and Geology. Additionally, spectral clustering revealed similarities between pairs of categories such as Economy and Politics and Education and Law.

Figure 2: Categories after spectral document clustering.



4 Conclusions

A multi-phase experiment has been described, conducted with the assistance of human annotators, as well as with automatic clustering. The first part of the experiment consisted in assigning UDC category codes to documents available in Polish digital libraries. This was achieved by first preparing a list of 5000 keywords, which were manually assigned as belonging to a specific UDC category. Then the documents were automatically assigned a two-digit category code based on these keywords. The second part of the experiment served to verify the consistency of this approach. By using clustering algorithms on the full texts of documents, clusters of similar documents were identified. Then, by analysis of UDC categories of similar documents, clusters of similar categories were induced. As these clusters tend to reflect the human intuition of category similarity, it can be concluded that the experiment was a success. The approach described can therefore be used to perform thematic classification of documents in large text collections.

References

[1] Filip Graliński. Polish digital libraries as a text corpus. In *Proceedings of 6th Language & Technology Conference*, pages 509–513, Poznań, 2013.

[2] A. Huang. Similarity measures for text document clustering. In *Proceedings of the Sixth New Zealand Computer Science Research Student Conference (NZC-SRSC2008), Christchurch, New Zealand*, pages 49–56, 2008.

[3] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[4] David Sculley. Web-scale k-means clustering. In *Proceedings of the 19th international conference on World wide web*, pages 1177–1178. ACM, 2010.

[5] J. Shi and J. Malik. Normalized cuts and image segmentation. In *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*, pages 731–737, Jun 1997.

[6] Amit Singhal. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 24(4):35–43, 2001.

[7] Chong Wang and David M. Blei. Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 448–456, New York, NY, USA, 2011. ACM.

[8] Piotr Wierzchoń. Fotodokumentacja 3.0. *Language, Communication, Information*, 4:63–80, 2009.

Measuring Tradition, Imitation, and Simplicity: The case of Attic Oratory[*]

Eleni Bozia

Lehrstuhl für Digital Humanities
Institut für Informatik
Universität Leipzig

Department of Classics
University of Florida
E-mail: bozia@ufl.edu

Abstract

This paper utilizes syntactically annotated texts of Attic orators to quantify their compositional styles. A unified node-based metric formulation was established and used in implementing various syntactical construction metrics, indicative of the compositional characteristics of the sentences. The developed metrics were applied to the texts of several authors, which were then comparatively examined using Principal Component Analysis, attempting to determine structural simplicity and complexity in Attic oratory.

1 Introduction

Classical studies as other humanities disciplines focus on the study of the language. Whether it has to do with grammatical, morphological, syntactical constituents, or word order and sentence construction that can potentially contribute to genre classification and fragment or author attribution, manual study has been predominant. The traditional process, however, in our case is problematic, as it proves unfeasible. The number of texts that a scholar needs to study comparatively is so prodigious that any sort of profound study proves nonviable. Furthermore, when a researcher delves into the text, there is a certain degree of subjectivity especially when we are examining stylistic attributes. The way to approach this type of examination is twofold—employ digital tools that will increase significantly the pace of the progress and

[*] This project was in part funded by the Humboldt Chair for Digital Humanities, Institut für Informatik, Universität Leipzig. The author would like to thank the sponsor and Dr. Giuseppe Celano for his insightful comments regarding Treebank annotation.

computational techniques that can contribute to the quantification and consequently the objectification of the results.

There are different computational approaches for the comparative analyses of texts. Passarotti et al. (2013) perform a lexical-based comparative examination between Seneca, Cicero and Aquinas, employing clustering techniques and Principal Component Analysis (PCA). Eder et al (2013) and Burrows (2002) and (2006) used stylometry to achieve a multilevel analysis of texts. Network analysis has also proven efficient in determining dependency relations between lemmas. Ferrer I Cancho et al. (2004), Ferrer I Cancho (2010), and Passarotti (2014) examined linguistic constituents against the backdrop of network theory. Ancient Greek and Latin present multifarious issues as they are less-resourced languages and the high non-projectivity and complexity of Ancient Greek that renders parses hard to train, hence interfering and impeding any automatic analysis. Dik (1995) provides a comprehensive account of Greek word order. Nivre et al. (2006) work on dependency parsing. Böhmová et al. (2001) and Hajičová (2004) discuss projectivity in the Prague Dependency Treebanks that have been employed for the annotation of Greek. Bamman et al. (2008) and (2009), Passarotti (2010), and Mambrini et al. (2012) and (2013) examine the issues in Latin and Ancient Greek respectively.

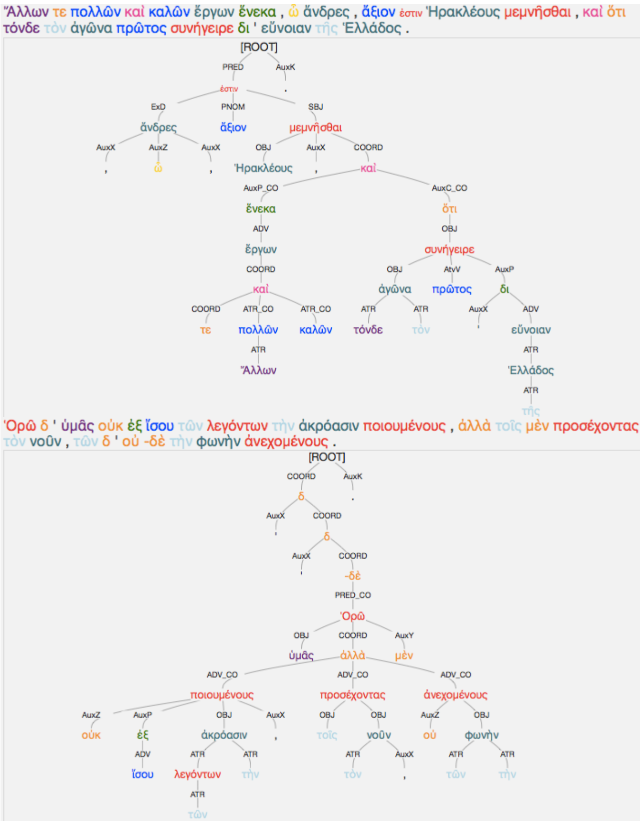
This paper attempts to quantify compositional simplicity and mannerism in Attic Greek oratory via the syntactical annotation of Classical Attic orators and orators of the Imperial Era as well as Dionysius' rewriting of some of the original passages. Dionysius of Halicarnassus wrote extensively on composition in theoretical treatises, such as *On Literary Composition* and his *Letters*, while he also acknowledged the palmary role of practice, hence discussing his theoretical considerations against the backdrop of actual rhetorical speeches. His *oeuvres* on the orators provide the readers with invaluable insight into *ars oratoria*, as Dionysius pursues comparative cogitations over the orators with simultaneous cross-references within the *On Literary Composition*. What is worth considering is that Dionysius not only explores writing styles and techniques, but he also critiques, offers his viewpoint, compares and contrasts, and occasionally suggests revisions of certain passages. Examining, therefore, his works, we may be in a position to demarcate the components of rhetorical writings and achieve an understanding of Greek oratory that will consequently shed light onto the different styles as well as the evolution of oratory and its influence on the Roman counterparts.

2 Philological Primer

Lysias is the par excellence rhetorical example to which Dionysius resorts. He praises the orator for clarity, simplicity, and straightforward expression.

"*As to his composition, it is absolutely simple and straightforward. He sees that characterization is achieved not by periodic structure and the use of rhythms, but by loosely constructed sentences*".[1] (Dion.Hal. *Lysias* 8)

Using Lysias as his framework, Dionysius proceeds to cognitively interpret and describe the other Attic orators. Isocrates' diction is no less pure than that of Lysias.[2] However, *"it is not compact, closely-knit style like the other…it sprawls and overflows with its own exuberance"*.[3] Lysias excels in succinctness, while Isocrates in amplification (see comparison in Fig. 1).[4]



Ἄλλων τε πολλῶν καὶ καλῶν ἔργων ἕνεκα , ὦ ἄνδρες , ἄξιον ἐστιν Ἡρακλέους μεμνῆσθαι , καὶ ὅτι τόνδε τὸν ἀγῶνα πρῶτος συνήγειρε δι ' εὔνοιαν τῆς Ἑλλάδος .

Ὁρῶ δ ' ὑμᾶς οὐκ ἐξ ἴσου τῶν λεγόντων τὴν ἀκρόασιν ποιουμένους , ἀλλὰ τοῖς μὲν προσέχοντας τὸν νοῦν , τῶν δ ' οὐ -δὲ τὴν φωνὴν ἀνεχομένους .

Fig 1. Syntactical comparison of Lysias' (top) and Isocrates' (bottom) sentences.

In addition to the internal comparison between the orators, Dionysius offers his own suggestions when he believes that an orator has transgressed the boundaries of eloquent artistry, thus vilifying unnecessary opulent mannerisms. Therefore, in his attempt to further explicate his points of condemnation, rewrites Isocratean passages, such as the following. The first

[1] Translation by Usher

[2] καθαρὰ μέν ἐστιν οὐχ ἧττον τῆς Λυσίου, Dion.Hal. *Isocrates* 2

[3] στρογγύλη δὲ οὐκ ἔστιν, ὥσπερ ἐκείνη, καὶ συγκεκροτημένη…ὑπτία δέ ἐστι μᾶλλον καὶ κεχυμένη πλουσίως, Dion.Hal. *Isocrates* 2

[4] ἐν δὲ τῷ συντόμως ἐκφέρειν τὰ νοήματα Λυσίαν μᾶλλον ἡγούμην ἐπιτυγχάνειν. περὶ τὰς αὐξήσεις Ἰσοκράτη κατορθοῦν ἄμεινον ἐδόκουν. Dion.Hal. *Isocrates* 11

picture in Fig. 2 shows Isocrates' sentence and the second Dionysius' *scriptum facilior*. Observing the two sentences, it seems that Dionysius substituted the secondary clause with a participle, making Isocrates' elongated style more succinct.

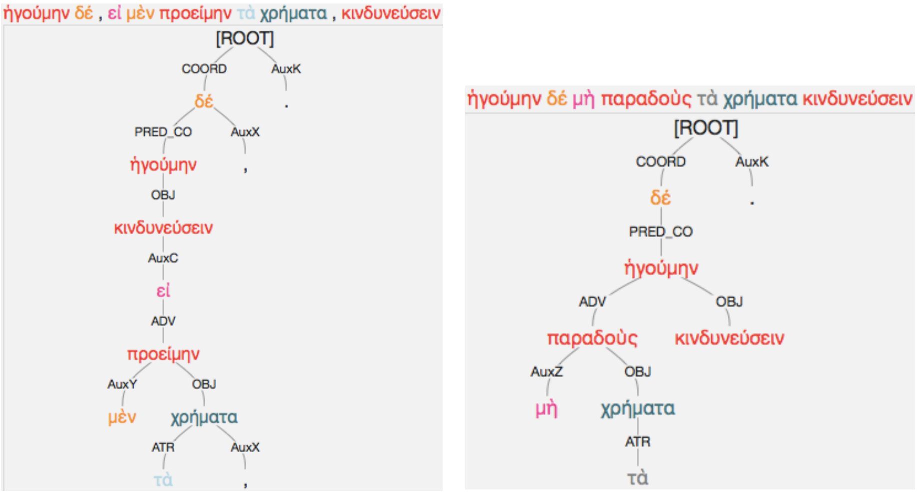


Fig 2. Syntactical comparison of Isocrates' (left) and Dionysius' (right) version.

3 Quantifying Rhetorical Styles

The way I approach this particular study of Atticism is to perform a quantitative analysis which in turn will formulate parameters of Atticism and also provide us with some distinguishable measures pertaining to its stylistic modulations. For the stylistic analysis of the rhetorical structure certain rhetorical speeches were annotated syntactically in the form of treebank annotation, using the Arethusa annotation framework through Perseids.[5] Such an organization of the words can also lead to a further in-depth study of statistical distributions of syntactical phenomena. The purpose of such a parsing of the text is to perform comparative analysis of the syntactical structures used by each author—the morphology of the tree, such as the width and depth of the branches, and the syntactical tags, describing each lemma—which may translate into a more or less convoluted form of expression.

The structure of a syntactically annotated sentence is defined as a linearly ordered set of elements $S=\{n_0, n_1, n_2, \dots n_k\}$, where each element n_i is a tree node ($n_i \in T$) and T denotes the space of tree nodes. In this space the operator *children of* is defined, which maps each node to a set of children nodes that are also elements of the same sentence (*children of*: $T\rightarrow \{T^x\}$). A node without children is mapped to the empty set through this operator (i.e. when $x=0$). The word order in a sentence defines the linear order of the tree

[5] www.perseids.org

nodes in the set S. Additional operators can also be defined to implement other characteristics of the nodes, such as syntactical tags, for example the mapping *isATR*: T→{0,1} could indicate if a given node is an attributive.

In order to quantify the use of Attic in this case, one needs to extract numerical descriptors for each annotated sentence in a given corpus. Therefore, a set of metrics could be defined within the space *S* of sentences that will then allow us to perform further comparative analyses. The syntactical morphology of the sentence is depicted in the connectivity of the nodes. This paper explores the possibility to establish a node-based metric so as to quantify the local morphology of each individual node and ultimately assess the complexity of the sentence. A generalized sentence metric can be expressed as the weighted sum of node-based metrics:

$$w_0\mu(n_0) + w_1\mu(n_1) + \cdots + w_k\mu(n_k) = \sum_{i=0}^{k} w_i\mu(n_i) \qquad (1)$$

where $\mu(n_i)$ is a metric that operates on node n_i and computes a numerical value (μ: T → ℝ). The weights w_i determine each node's degree of syntactical and positional contribution in the sentence. An example of a simple node-based sentence metric is the *number_of_words* in which $w_i=1$ and $\mu(n_i)=1$ for all nodes in the sentence. In this case Eq. 1 will calculate the number of words in a given sentence. Another simple example is the *percentage_of_leaves* in which $w_i=1/k$ and $\mu(n_i)=1$ if n_i is a leaf, or 0 if otherwise. Similarly more complex metrics can be defined by setting the weights w_i and the node metric μ accordingly. It should be noted that, despite the linear form of Eq. 1, non-linear metrics can also be established, using $w_{root}=1$, $w_{i>0}=0$, and setting $\mu(n_{root})$ to be a non-linear function that can operate on the entire sentence tree by traversing it from the root.

This analytical framework was implemented in JavaScript as a generic web-based programming interface. The developed interface includes API for defining custom node-based sentence metrics and is compatible with the Ancient Language Dependency Treebank (ALDT) format version 1.5 currently used by Perseids. In this interface a custom metric is defined as:

```
var example=new NodeMetric(“Num of attrib.”);
example.weight=function(node){return 1;};
example.metric=function(node){
return node.getRelation()==“ATR”; };
var value=example.apply(sentence);
```

The above example shows a case in which $w_i=1$ and $\mu(n_i)=1$ if n_i is ATR (attributive), or 0 if otherwise. The last line in the example shows how a metric can be applied to a Treebank sentence, which is given as a `TreebankSentence` object variable (here named `sentence`).

The purpose of this framework is to set certain metrics and apply them on a collection of treebanks. This process will produce several numerical descriptors of each sentence that can quantify syntactical construction. These numbers can then be used as an input in any classification or pattern analysis algorithm, such as Principal Component Analysis, to examine similarities between works, authors, and writing styles.

For this initial study, a pilot sample of sentences was selected from the Attic orators Lysias, Isocrates, Demosthenes, as well as Lucian, Dio of Prusa, and Aelius Aristides—Imperial orators who revived Attic style—and Dionysius of Halicarnassus, the literary theorist who discusses their writing styles. Approximately one hundred sentences were annotated, using the Perseids Treebank Annotator. Five node-based metrics were implemented, using the presented framework, which are the ratios of the: a) number of leaf nodes, b) number of ATR nodes, c) tree height, d) tree width, and e) max number of branches in a single node, to the total number of nodes in the sentence. Considering that this pilot study was conducted on a relatively small corpus, in an attempt to produce trustworthy results, the metrics that were used were selected as robust descriptors of the morphology of the sentence and are not prone to noise deriving from local syntactic variations. The proposed framework is generic enough to allow for more complex metrics, such as the average number of nodes per noun phrase. Using these metrics, each sentence becomes a point in a 5D space. This point cloud was processed using PCA to map the data onto the plane of the largest spread. Furthermore, the sentences of each author were approximated by a Gaussian distribution that is shown as an ellipse in Fig. 3.

The plots indicate that Lysias and Lucian display proximity. Demosthenes appears between the two, which can also be explained chronologically and through Dionysius' analysis of Lysias' and Demosthenes' style. Finally, Dio displays *imitatio* of Demosthenes, which can be explained when one considers their socio-political awareness. Lucian who openly embraces and imitates Attic style appears to converge significantly with Lysias. Finally, Aelius, who, albeit Attic, admittedly invented his own authorial identity, relates to but does not register with the majority of the other orators.

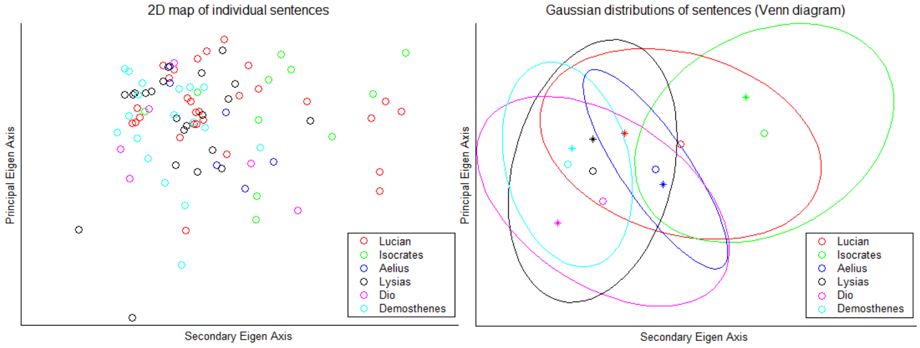


Fig 3. Plots of the sentence dataset on the plane of the two dominant eigenvectors.

To conclude, in this paper a framework was presented for the analysis of stylistic similarities in certain Attic orators. A limited pilot study was performed as a proof of concept. In the future, I intend to annotate a larger amount of those authors' *corpora*, create more complex metrics that will better encapsulate stylistic characteristics, and ultimately produce results, regarding the use of Atticism in Greek and Roman oratory.

References

[1] Bamman, D. et al. (2008) The Annotation guidelines of the Latin Dependency Treebank and Index Thomisticus Treebank. In *Proceedings of the 5th SaLTMiL Workshop*, 71-76. Morocco.

[2] Bamman D. at al. (2009) An Ownership Model of Annotation: The Ancient Greek Dependency Treebank. In *Proceedings of the Eighth International Workshop on Treebanks and Linguistic Theories*, 5-15, Milan.

[3] Böhmová, A. et al. (2001) 'The Prague Dependency Treebank: A three-level Annotation Scenario.' In Abeillé, A. (ed.) *Treebanks: Building and Using Syntactically Annotated Corpora*, 103-127. Boston.

[4] Burrows, J.F. (2002) 'DELTA': a measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing* 17(3), 267-87.

[5] Burrows, J.F. (2006) 'All the Way Through: Testing for Authorship in Different Frequency Strata.' *Literary and Linguistic Computing* 22(1), 27-48.

[6] Dik, H. (1995) *Word Order in Ancient Greek: A Pragmatic Account of Word Order Variation in Herodotus*. Amsterdam.

[7] Eder, M. et al. (2013) Stylometry with R: a suite of tools, In *Digital Humanities 2013: Conference Abstracts*, 487-489, Lincoln.

[7] Ferrer I Cancho, R. et al. (2004) Patterns in Syntactic Dependency Networks. In *Physical Review* E69, 051915(8).

[8] Ferrer I Cancho, R. (2010) Network Theory. P.C. Hogan (ed.). In *The Cambridge Encyclopedia of Language Sciences*, 555-557, Cambridge.

[9] Hajičová, E. et al. (2004) Issues of Projectivity in the Prague Dependency Treebank. In *Prague Bulletin of Mathematical Linguistics* 81, 5-22.

[10] Mambrini, F. et al. (2012) Will a Parser Overtake Achilles? First Experiments on Parsing the Ancient Greek Dependency Treebank. In *Proceedings of the 11th Workshop on TLT*, 133-144. Lisbon.

[11] Mambrini, F. et al. (2013) Non-projectivity in the Ancient Greek Dependency Treebank. In *Proceedings of the 2nd International Conference on Dependency Linguistics*, 177-186.

[12] Nivre, J. et al. (2006) Maltparser: A data-driven parser-generator for dependency parsing. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, 2216-2219. Genoa.

[13] Passarotti, M. (2010) Leaving Behind the Less-resourced Status. In *Proceedings of the 7th SaLTMiL Workshop*, Malta.

[14] Passarotti, M. et al. (2013) A Statistical Investigation Into the Corpus of Seneca, In *Proceedings of The 17th International Colloquium on Latin Linguistics,* 1-16. Rome, Italy.

[15] Passarotti, M. (2014) The Importance of Being sum. Network Analysis of a Latin Dependency Treebank. In *Proceedings of La Prima Conferenza Italiana di Linguistica Computazionale*, 291-295. Pisa.

[15] Usher, S. 1974. *Dionysius of Halicarnassus*. Harvard University Press: Cambridge, MA.

Ranking corpus texts by spelling error rate

Philip Diderichsen, Sune Just Christensen & Jørgen Schack
Danish Language Council
Email: phildi@dsn.dk, sunejc@dsn.dk, schack@dsn.dk

Abstract

We are developing a method to rank corpus texts according to text quality. If the number of orthographic errors in a text correlates with the quality of the text more generally, error counts could be used as a quality metric. We are currently investigating whether this is in fact the case in a newspaper corpus. In a first stage of the project, a pilot study was conducted involving counts of non-word errors. The results did not conflict with such a correlation. We have therefore now moved on to the next stage of the project, where we investigate whether this also holds for real-word errors.

1 Introduction

As one of its main tasks, the Danish Language Council codifies Danish orthography by editing and publishing the official dictionary of Danish standard orthography (Retskrivningsordbogen)[1]. The two main principles stated in the legal framework of the council are the principle of tradition and the principle of language use. According to the principle of tradition, words and word forms are written in accordance with the practice laid down in an executive order on orthography from 1892. The principle of language use acts as the main exception to this. It states that adjustments of the orthography may be made in accordance with the written language use of "competent language users".

As a concrete example, the Language Council might be considering the codification of a new spelling variant for the word *autenticitet* (Eng. *authenticity*). The use of the variant *autencitet* (displaying a kind of quasi-haplography) has been observed, and it might be a candidate for inclusion in a future edition of Retskrivningsordbogen. In order to depart from the principle of tradition and authorize the new variant, the council must verify that the variant is in accordance with competent language use. As it happens, *autenticitet* is used about 75% of the time, *autencitet* about 25% of the time in Danish newspaper text. But is this enough to authorize the new form? And are the texts in which the new variant occurs in fact written by "competent language users" in the sense of the above-mentioned legal framework?

The latter question is the more important one of the two. We have no reason to suspect that journalists are not generally competent language

[1] Cf. http://www.dsn.dk/om-os/about-the-danish-language-council (retrieved September 2015).

users. However, the Language Council is very interested in being able to actually verify whether new variants occur in the highest quality texts or not.

To this end, the aim of the current project is to develop an orthography-based metric of text quality — and preferably a general method of deriving such a metric for various text types other than newspaper text.

2 Orthography-based ranking of corpus texts

The idea of the project is to rank the texts in a newspaper corpus according to the number of deviations from the official standard orthography. For this to be a meaningful proxy for "competent written language use", it must correlate with text quality more generally. This has been shown to hold for elementary school children as well as adults (see Abbott et al. (2010) and references therein), and in the field of automated essay scoring (AES), it has been established that this kind of correlation can be harnessed in software systems automatically grading text submitted as part of academic aptitude and placement tests and English as a foreign language tests (Deane 2013, Enright & Quinlan 2010, Ramineni 2013, Östling et al. 2013). An important part of our project is to verify whether such a correlation also exists in real texts which are not produced in an experimental setting and which, importantly, have already undergone varying degrees of revision before being published. We have completed an initial, small pilot study based on non-word counts, to be summarized below, which did not conflict with this assumption. Encouraged by these results, we have continued to the next stage of the project involving counts of real-word error counts. An outline of this work in progress will be given below.

2.1 Pilot study on non-word-based ranking

The first stage of the project consisted of a pilot project with the following steps:

1. Selection of orthographic deviation patterns and semi-automatic generation of non-word search expressions (regular expressions) on this basis.
2. Corpus analysis to find error texts (i.e. with many non-words).
3. Text judgment experiment with 'disguised' error texts found in step 2 vs. matched control texts not identified in step 2.

Step 1 proceeded as follows. Nearly 90 different known types of non-word spelling deviation were chosen. We define non-words as spelling variants that never appear as correct words in any contemporary Danish language context. The types of spelling deviation chosen could comprise single-word patterns such as *hieraki* (correct form: *hierar**k**i*, Eng. *hierarchy*), patterns which apply to several different lexemes such as *-tion* for the correct *-sion* (as in *reflektion*, correct: *refleksion*, Eng. *reflection*), or patterns with a large number of affected lexemes such as dropping the joining interfix *-s-* before *-s* in compounds (where it is often mandatory e.g. *elskovsyg*, correct: *elskovssyg*, Eng. *love sick*). For each pattern, we made sure to generate all possible (non-word) occurrences of the deviation in the vocabulary covered in Retskrivningsordbogen. Also, where appropriate, we added regular expression extensions to the patterns in order to be able to capture compounds, derivations and inflected forms including the deviating pattern. E.g., the deviant form *hieraki* was extended to *.*hieraki.** using standard regular expression wildcards. In this way, a non-word error list containing about 20 000 error patterns was created.

Using a positive error list was chosen over identifying word forms not appearing on a list of all officially codified word forms because we would otherwise be overwhelmed by false positives in the lexically rich and quite error-sparse text of the newspaper corpus. In other words, we prioritized precision over recall.

Step 2 was to analyze our corpus for occurrences of the patterns on the non-word list. The corpus used consists of about 370 000 newspaper texts primarily from 2010-2014 from 7 contemporary national Danish newspapers (Berlingske, Ekstra Bladet, Information, Jyllands-Posten, Kristeligt Dagblad, Politiken, and Weekendavisen). The texts are copies of the XML files used as source files for the printed newspaper articles; thus, OCR-processing of the texts is not necessary. The corpus contains a total of nearly 200 M word tokens. The total number of non-word occurrences + their correct counterparts was nearly 10 M tokens. Of this, the deviations constituted approximately 1‰, i.e. nearly 10 000 tokens. About 8 000 texts contained one or more occurrences of the non-word patterns on our list.

In step 3, we identified the texts with the highest numbers of our non-word types in order to investigate how they would be judged by a panel of language professionals — after removal of all orthographic errors. The three texts with the highest numbers of types were (1) a literature review (7 non-word errors of 6 different types), (2) an article about housing policy (5 errors of 5 types), and (3) a Formula 1 racing report (5 errors of 5 types). For each of these texts we found a matching text using the metadata of the articles. Thus, we found match texts from the same newspaper and year, of approximately the same length, with the same topic, and with a similar author (same gender and approximate age (judged on the author's name), and with about the same number of texts in the corpus). Each text pair was slightly edited in order to remove any orthographic errors, including the ones on our non-word list (we left morphological and syntactic errors as well as comma errors and random typos in the texts). An overview over the number of errors in the texts is given in table 1, 2, and 3.

Error type	Error text 1	Match text 1
Orthographic errors from list	7	-
Additional orthographic errors	4	0
Morphological errors	1	0
Syntactic errors	1	0
Typos	1	0
Comma errors	11	1

Table 1. Number of errors in text pair 1. The shaded area covers the errors we removed from the texts.

We asked a panel of 8 proofreading and text revision experts at the Danish Language Council to act as judges of the texts. They were asked to judge the texts on general text quality considering all the usual levels of linguistic description: orthography, morphology, syntax, text coherence and stylistic aspects. For each pair, they were asked to pick a 'winner' and freely formulate comments about why one or the other text was chosen. A sketch of the preparation and judgment of the first text pair can be seen in figure 1.

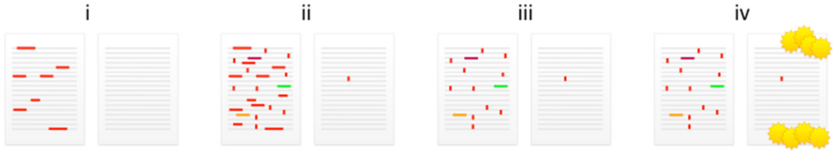


Figure 1. The preparation and judgment of the first text pair. (i) The error text contained 7 non-word errors of 6 different types from our list. (ii) The error text contained several additional errors; the match text only contained one in this instance (cf. table 1). (iii) All orthographical errors, with certain exceptions, were removed before the texts were judged. (iv) All 8 judges preferred the match text.

The first error text contained a lot of comma errors as well as a few other errors in addition to the ones we removed before the comparison. The match text only contained a single comma error. Therefore, it is hardly surprising that all 8 judges preferred the match text.

In the second text pair, the error count was as can be seen in table 2. After removing orthographical errors, the texts had a comparable amount of remaining errors.

The result of this comparison was that 7 out of 8 judges preferred the match text despite the lower amount of error cues. This suggests that the more textual and stylistic aspects of the texts were congruent with the degree of orthographic deviance of the texts. The judges' comments support this interpretation in that several of the judges pointed out stylistic problems and problems with coherence.

Error type	Error text 2	Match text 2
Orthographic errors from list	5	-
Additional orthographic errors	5	2
Morphological errors	0	0
Syntactic errors	0	0
Typos	0	1
Comma errors	3	1

Table 2. Number of errors in text pair 2. The shaded area covers the errors we removed from the texts.

In the third comparison, there was again a 7 against 1 majority — however, this time favouring the error text. This is rather interesting in view of the numbers in table 3, because as it turned out, the match text actually contained more errors than the error text. When the orthographic errors were removed, the texts were again rather equal with respect to remaining errors, but like in comparison 2, the judges preferred the text that originally contained the fewest errors, which this time, incidentally, was the error text[2].

Error type	Error text 3	Match text 3
Orthographic errors from list	5	-
Additional orthographic errors	1	9
Morphological errors	0	0
Syntactic errors	0	1
Typos	1	0
Comma errors	2	2

Table 3. Number of errors in text pair 3. The shaded area covers the errors we removed from the texts.

Thus, the results of this first pilot study do not contradict the assumption that must hold for our larger endeavour to make sense: that the higher textual and stylistic aspects of quality correlate with the number of orthographic errors, even in published newspaper texts.

2.2 Real-word errors join the ranks

We found the above results encouraging enough to launch the next stage of the project: ranking on the basis of real-word as well as non-word errors. This stage of the project can be outlined as follows:

1. Automatic annotation of the corpus with POS-tags and error tags using the constraint grammar-based tool DanProof.

[2] The rather large number of errors in the match text that were not covered by our error list were mainly split compounds (e.g. *Le Mans bil* for the correct *Le Mans-bil,* Eng. *Le Mans car*), which are not trivially handled by a non-word list.

2. Development of supplementary constraint grammar rules with the purpose of refining the error annotation (mainly to get rid of false positives).
3. Corpus analysis to find error texts.
4. Another round of human evaluations of error texts vs. match texts.

We had the corpus annotated with the DanProof-tool developed by GrammarSoft (Bick 2015). DanProof is a constraint grammar-based tool employing constraint grammar rules specially designed for texts which may contain spelling errors (i.e., it is essentially a spell checker engine). It will add POS tags and error tags, among other things, to an input text, but in a biased manner so that real-word errors are caught and annotated as such rather than being taken at face value (resulting in an erroneous POS annotation). Our exploration of the tool indicates that it has quite good recall on real-word errors in newspaper text of the kind we are currently working with. This does come at the price of a lack of precision, however. Therefore, the task at hand is to try to find reliable grammatical patterns which will identify the true positives among the error-tagged tokens.

The approach is quite parallel to the one taken in the first stage of the project: We need to find reliable search patterns which will give us high precision, probably at the cost of some recall — only this time patterns in the more powerful constraint grammar formalism rather than mere regular expressions.

We are currently working on developing the rules for identifying positive occurrences of a few notorious Danish real-word error types:

- *r*-drop in word forms ending in *-rer* which sound the same as words ending in *-re*, and vice versa (*r*-appending).
- Confusion of the suffixes *-ende* and *-ene* (conventionally a present participle verb suffix and a definite plural noun suffix, respectively).
- Split compounds (in Danish, the individual elements of compounds are joined together).

When a number of these error types have been reliably captured, a new set of texts will be extracted and examined to see whether the assumption of a correlation between orthographic and more general text quality still holds. If so, more error types will be added.

3 Discussion

At this stage of the project, we do not yet have a fully developed metric of text quality. We are still exploring which orthographic deviations we are able to identify automatically, and to what extent they are associated with reduced text quality. Therefore, it is yet too early to determine how forms like *aut-*

encitet and *autenticitet* are distributed over higher and lower quality texts in our corpus.

Since spelling deviations are the very basis for the quality metric we envision, it may seem unclear how these new spellings can ever be part of the quality stratum covered by the formulations about "competent language use" mentioned above. In practice, this may turn out to be a relatively minor problem. The deviations that will eventually be searched for will not be an indiscriminate collection of all spelling deviations, but rather at set of carefully selected deviations that reflect quite severe deviations from the standard orthography, or otherwise deviations that the Language Council has no intentions of authorizing in the foreseeable future. Potential new spellings are not likely to be in this set.

However, spelling errors may turn out to be too one-dimensional a basis for the kind of quality metric we want. We may have to include other more or less readily quantifiable measures inspired by the AES literature.

4 Conclusion

Our pursuit of finding a valid method of ranking corpus texts by quality is work in progress. Our first, modest pilot study results based on orthographic non-words have shown that the number of orthographic errors in published newspaper texts may indeed correlate with their quality more generally. This has encouraged us to go on to real-word errors.

Acknowledgment

We would like to thank our three anonymous reviewers for their comments and suggestions.

References

Abbott, R. D., Berninger, V. W. & Fayol, M. (2010). Longitudinal relationships of levels of language in writing and between writing and reading in grades 1 to 7. *Journal of educational psychology*, *102*(2), pp. 281-298.

Bick, Eckhard (2015). *DanProof: Pedagogical Spell and Grammar Checking for Danish*. In: Galia Angelova, Kalina Bontcheva & Ruslan Mitkov: Proceedings of RANLP 2015 (Hissar, Bulgaria, 7-9 Sept. 2015), pp. 55-62.

Deane, Paul (2013). *On the relation between automated essay scoring and modern views of the writing construct*. Assessing Writing 18, pp. 7–24.

Enright, Mary K. and Thomas Quinlan (2010). *Complementing human judgment of essays written by English language learners with e-rater® scoring*. Language Testing 27(3), pp. 317–334.

Ramineni, Chaitanya (2013). *Validating automated essay scoring for online writing placement*. Assessing Writing 18, pp. 40–61.

Östling, Robert, Andre Smolentzov, Hinnerich, B. Tyrefors, Erik Höglin (2013). *Automated Essay Scoring for Swedish*. In: Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 42-47. Association for Computational Linguistics.

Toiling with the Pāli Canon

Frederik Elwert[1], Sven Sellmer[2],
Sven Wortmann[3], Manuel Pachurka[1], Jürgen Knauth[4], David Alfter
(1) Ruhr-Universität Bochum, (2) Adam Mickiewicz University in Poznań,
(3) Universität zu Köln, (4) Georg-August-Universität Göttingen
E-mail: frederik.elwert@rub.de, sven.sellmer@amu.edu.pl,
swortmann@uni-koeln.de, manuel.pachurka@rub.de,
alfter.david@gmx.net

Abstract

The paper describes the preparation of a Buddhist corpus in the Middle Indo-Aryan language Pāli, which is available only in a flat TEI format, for content-based analysis. This task includes transforming the file into a hierarchical TEI P5 representation, followed by tokenisation (including sandhi resolution), lemmatisation, and POS tagging.

1 Introduction

The Pāli Canon is the oldest corpus of Buddhist texts and is considered authoritative by the Theravāda sect (for an overview see von Hinüber [4]). Its name is derived from the Middle Indo-Aryan language in which it was composed (Oberlies [7]). It originated as oral literature in Northern India in the 4th c. BCE, spread and grew in the course of the expansion of Buddhism and was finally committed to writing ca. in the 1st c. BCE on the island of Sri Lanka. In the subsequent period, the text still underwent some changes, but only to a much smaller degree. It consists of three text collections:

- *Vinayapiṭaka* (disciplinary rules for monks and nuns)
- *Suttapiṭaka* (religious and philosophical teachings)
- *Abhidhammapiṭaka* (detailed systematic accounts of philosophical and psychological doctrines).

The general aim of our project was conducting IT-based research on the *Suttaṭipaka* (ca. 1.75 million lexical units) from the point of view of religious studies using methods of network and topic analysis. Unfortunately, the text was not available in such a form that we could start our analyses right away; rather, we had to prepare the data first. In our paper both these (rather time-consuming) preparatory steps and some outlines of the main results will be presented. (The other two *piṭaka*s have not been dealt with in any detail by us so far, but a first assessment shows that generally the same steps as described in this paper can be applied to them, only minor modifications being necessary.)

2 Preparation

2.1 Text structure

The text of the *Suttaṭipaka* is available in digital form ([13]) in a kind of flat TEI format in the now outdated version P4. The encoding focuses mainly on the presentational aspects and was apparently designed to reflect the typographical features of the printed edition and so to allow for an adequate HTML publication.

The text is contained in 541 XML files with names bearing encoded information about their position in a four-tiered hierarchy of:

- 1st level: *piṭaka*
- 2nd level: *nikāya*
- 3rd level: *pāḷi*
- 4th level: variously named divisions, like: *vagga* ("group"), *sutta* ("teaching"), *saṃyutta* ("collection") etc.

E.g., the file name "s0202m.mul0.xml" is to be interpreted as follows:

• s	*Suttapiṭaka*
• 02	2nd *nikāya*, i.e., *Majjhimanikāya*
• 02	2nd *pāḷi* in the *Majjhimanikāya*, i.e., *Majjhimapaṇṇāsapāḷi*
• m	no function
• mul0	1st *vagga*,[1] i.e., *Gahapativagga* ("mul" indicating that we are dealing with the source text, not with a commentary, which is included in a separate file).

The hierarchical structure of these four levels is also represented in an XML file that uses the tree format for the WebFX framework in order to present a browsable hierarchy on the tipitaka.org website. This XML format is non-standard, but can still be used to extract the hierarchy of the individual TEI P4 files. However, inside the TEI P4 files, a flat structure is used that

[1] This index is zero-based.

gives no information about further nesting of sections in the text. The text is merely divided into a sequence of paragraphs on the same level enclosed between <p> (paragraph) tags. The attribute @rend (rendition) is used in order to record representational features of the text, but no semantic information about the role of the text parts as headings, text body, or verse is encoded, let alone text divisions. The format used thus looks like this:

```
<text>
  <body>
    <p rend="centre"> Namo tassa bhagavato arahato sammāsam-
buddhassa</p>
    <p rend="chapter">1. Gahapativaggo</p>
    <p rend="subhead">1. Kandarakasuttaṃ</p>
    <p rend="bodytext" n="1"><hi rend="paranum">1</hi><hi
rend="dot">.</hi> Evaṃ <pb ed="T" n="2.0001" /><pb ed="M"
n="2.0001" /><pb ed="P" n="1.0339" /><pb ed="V" n="2.0001" />
me sutaṃ – ekaṃ samayaṃ bhagavā campāyaṃ viharati gaggarāya
pokkharaṇiyā tīre mahatā bhikkhusaṅghena saddhiṃ. Atha kho
pesso <note>peyo (ka.)</note> ca hatthārohaputto kandarako ca
paribbājako yena bhagavā tenupasaṅkamiṃsu; […]</p>
  </body>
</text>
```

In addition to the rendition of paragraphs, some additional editorial information is given, though in a very basic form. Page breaks of different editions are indicated, and text variants in the editions are inserted as <note> elements. The paragraphs are usually numbered, though the paragraph number is redundantly encoded twice, using the @n attribute and as part of the text.

This structure was not fine-grained enough for our purposes; e.g., the *Gahapativagga* just mentioned contains accounts of ten separate conversations of different householders with the Buddha, which should be accessible individually. We therefore chose to prepare first of all a hierarchically structured TEI P5 representation with one or more additional levels (as required) by processing the whole *Suttapiṭaka* with a dedicated XSLT script. The script mainly works by recognising beginnings and endings of sections and enclosing them with <div> tags. This is possible by combining different pieces of information, partly contained in XML attributes, partly based on expert knowledge. To the first type belong attributes added to certain kinds of headings by the VRI editors (e.g., <p rend="subhead">); as an example for the second type of information formulaic expressions at the end of sections (e.g., XY-*suttaṃ niṭṭhitaṃ* “teaching XY is finished”) may be adduced. Technically this task proved to be rather tricky, mainly due to inconsistencies both in the source file and in the printed edition, so that a considerable amount of manual control, adjustments and corrections proved necessary. Eventually we managed to arrive at a file with a regular structure of up to six textual levels, the paragraph level being the lowest:

```
<text>
  <body>
    <div type="vagga">
      <head n="1">Gahapativaggo</head>
      <opener>
        <salute> Namo tassa bhagavato arahato sammāsambud-
dhassa</salute>
      </opener>
      <div type="sutta">
        <head n="1">Kandarakasuttaṃ</head>
        <p n="1"> evaṃ <pb ed="T" n="2.0001"/><pb ed="M"
n="2.0001"/><pb ed="P" n="1.0339"/><pb ed="V" n="2.0001"/> me
sutaṃ – ekaṃ samayaṃ bhagavā campāyaṃ viharati gaggarāya
pokkharaṇiyā tīre mahatā bhikkhusaṅghena saddhiṃ. atha kho
pesso <note place="app">peyo (ka.)</note> ca hatthārohaputto
kandarako ca paribbājako yena bhagavā tenupasaṅkamiṃsu; […]
       </p>
      </div>
    </div>
  </body>
</text>
```

Editorial information, like page breaks, has been carried over. Instead of the representational categories like “chapter” or “subhead,” which do not relate to the semantic structure of the corpus, object-language types like “vagga” and “sutta” have been introduced.

Additionally, poems in the text have been transformed into <lg>/<l> structures. In the original files, again also rendition information hints to the logical structure of the text:

```
<p rend="bodytext" n="30"><hi rend="paranum">30</hi><hi
rend="dot">.</hi> ‘‘Brahmunāpesā, mahānāma, sanaṅkumārena
gāthā bhāsitā –</p>
<p rend="gatha1">‘Khattiyo seṭṭho janetasmiṃ, ye
gottapaṭisārino;</p>
```

```
<p rend="gathalast">Vijjācaraṇasampanno, so seṭṭho de-
vamānuse’ti.</p>
```

The resulting files represent the structure in a more transparent way:

```
<p n="30"> “brahmunāpesā, mahānāma, sanaṅkumārena gāthā
bhāsitā –</p>
<lg type="verse">
  <l>‘khattiyo seṭṭho janetasmiṃ, ye gottapaṭisārino;</l>
  <l>vijjācaraṇasampanno, so seṭṭho devamānuse’ti.</l>
</lg>
```

2.2 Linguistic information

In addition to coming to terms with the text structure, our second goal consisted in applying lemmatisation and POS tagging because for our project it was essential to be able to access information about actors and basic concepts. Here, a considerable amount of groundwork had to be done, as practically no tools from computational linguistics were available for Pāli.

2.2.1 Tokenisation and sandhi resolution

A major problem for computational approaches to texts in Sanskrit and Middle Indo-Aryan languages is based on the fact that traditionally all kinds of euphonic sound changes, known by the term sandhi (of Sanskrit origin), are reflected in written texts, including such that result in the merging of two or more words to a single character string (e.g., *cāhaṃ* ← *ca ahaṃ* “and I”). Therefore for virtually all tasks of computational analysis the sandhi changes first have to be reversed, which mostly cannot be done in a mechanical manner because often two or more solutions are theoretically possible. Whereas for Sanskrit, sophisticated programs are available that perform quite well though not entirely without human supervision (Hellwig [2]; [3]), there are no such resources for Pāli.

The chief difference between Pāli and Classical Sanskrit consists in the fact that in Pāli, sandhi is not obligatory in all cases but mainly occurs where certain high-frequency words (especially particles like *ca* “and”, *pi* “also”, or *eva* “even”) are involved. This feature of sandhi in Pāli on the one hand excludes a truly systematic solution as it is possible and required in Sanskrit, but on the other hand enables a pragmatic approach consisting in the manual analysis of frequent “sandhi triggers”. As a result of such an analysis we formulated a set of ca. 175 rules, based on regular expressions, with the help of which the majority of sandhi changes could be undone.

The performance of this rule-based approach was evaluated by checking the number of false positives (i.e., incorrect changes) and false negatives (i.e., unreversed sandhis) in appropriate samples. In 1,555,172 input strings[2] the application of the mentioned ruled caused 93,294 changes. Of these, a random sample of 900 instances was inspected, but no errors were found, so the number of false positives seems to be very low. In order to estimate the percentage of unresolved sandhi changes we undertook a check of a random sample of 1,600 strings. In these, 13 strings featured sandhi changes, which indicates that in the whole text about 13,500 sandhi changes were left unreversed, i.e., probably ca. 87% of all changes were correctly resolved. Further optimisation is quite possible, but would be time-consuming because the most frequent types of changes have already been dealt with by the present set of rules, so every new rule will cover only a rather small number of cases.[3] As the results are good enough to serve as a basis for explorative analyses of the kind intended, we decided to leave the enhancement of the sandhi module for a later stage.

2.2.2 Lemmatisation and POS tagging

Because Pāli is a highly flective language, lemmatisation is anything but a trivial task. Our strategy was a twofold one. In a first step we prepared a reasonably comprehensive list of Pāli word forms that might occur in our text; in a second step we used these forms, together with manually prepared gold data, to train a readily available lemmatiser. For the first step, it was of great help that we could use the data of the digitised version of the *Pāli English Dictionary* (= *PED* [10]) though — because the markup of the dictionary data was very rudimentary — we had to create an improved version by extracting grammatical information (on parts of speech and other aspects) from the plain text of the dictionary articles.[4] In addition we prepared a list of proper names (which are not part of the *PED*) on the basis of the digitised version of the *Dictionary of Pali Proper Names* [7], which involved a manual selection of relevant items because this dictionary contains many items that were not only useless but actually harmful for our analyses, e.g., be-

[2] It is important to speak of "strings" here because, to refer to the example given above, *cāhaṃ* is one string representing two words: *ca* and *ahaṃ*.

[3] This claim can be made quite confidently because as a result of the lemmatisation described in the following section strings resulting from sandhi are almost always tagged as "<unknown>", the only exceptions being sporadic cases where these strings happen to be identical with possible word forms. Analysing the set of "unknown" strings with the help of suitable regular expressions, it is therefore possible to obtain a good overview of the different types of unresolved sandhis and of their frequencies.

[4] Some aspects of this step are discussed in Knauth and Alfter [6].

cause of homonymy. Pāli also features a large number of compounds, of which we managed to identify and split about 7,500 (in terms of types). The list of word forms was built of four kinds of data:

- list of indeclinables, taken directly from the improved dictionary
- list of irregular forms of nouns, pronouns, and numerals (obtained by manual input)
- list of verb forms (by courtesy of Yukio Yamanaka, a scholar specialising in Pāli verbs); this list is still being supplemented by Dr Yamanaka and presently comprises about 1/3 of all verb forms occurring in the Pāli canon
- list of regular declension forms of nouns and adjectives; this list was prepared by a simple generator algorithm that combined noun stems with the appropriate endings (here, a certain amount of overgeneration and some false forms were inevitable, but on the whole the generator yielded quite acceptable results).

The main step of the lemmatisation and the POS tagging were both done by the well-known NLP tool "TreeTagger" (Schmid [11]; [12]); the gold data being obtained by a manual annotation of 1,090 sentences with in total 14,017 words, of which a portion of 92 sentences with 1,496 words were not used in training the tagger, but kept for evaluation purposes. As for our main task fine-grained results were not required we chose a simple tag set with the following items:

- noun
- proper noun
- verb
- particle
- punctuation
- sentence delimiter
- adverb
- adjective
- pronoun
- numeral

The performance was checked with the help of the just mentioned evaluation data set of 1,496 words. In 18 instances a sandhi combination was unresolved, so these items were not taken into account. Of the remaining tags, 121 attributions turned out to be wrong, which yields a performance rate of 91.75% correctly tagged items. Taking a closer look at the mistakes, the following types can be distinguished:

Type of mistake		Frequency
Attributed tag	Correct tag	
NOUN	ADJ	58
NOUN	VERB	37
ADJ	NOUN	15
ADJ	VERB	7
NOUN	PRON	1
PROP	ADJ	1
PROP	NOUN	1
PROP	VERB	1

From these figures it clearly emerges that two types of mistakes have by far the greatest impact:

- verbs not recognized: 45 (37.2%)
- noun for adjective or vice versa: 73 (60.3%).

Here, it can be remarked that the first type is bound to disappear almost entirely as soon as the list of verb forms will be complete, which would improve the overall performance to a range of about 95%. As far as the second error type is concerned, it will be much more difficult to minimise, because it is often objectively difficult to distinguish between adjectives and nouns in Pāli: Both categories are formally identical, nominalisation is a very common process, and the word order is quite free (though it may provide some clues), so that even a human reader regularly finds it difficult to take a decision.

The lemmatising performs at somewhat lower precision rates though a detailed assessment would require a careful reading of many passages because Pāli features a particularly high percentage of homonyms and these are treated as separate items by the TreeTagger only if they belong to a different POS class.

It is planned to make the file we prepared according to the above description available on a web repository for Old Indian texts as soon as we will have solved some minor technical issues; all tools used will be published in open source form in the near future.

2.3 Workflow

Since the TEI format is not well suited for computational linguistics tasks and analysis, the TEI representation was first transformed on the fly into the simpler weblicht TCF format (Hinrichs [5]). These data were then passed to the TreeTagger with the help of a small adapter module. Further analysis was carried out on the POS tagged and lemmatised data using the TCFnetworks package (Elwert [1]) and MALLET (McCallum [8]).

3 Corpus-based textual research

The described preparation enabled us to doing content-related research by performing topic analysis with the help of MALLET. Concretely, we selected 102 thematically more or less homogeneous text passages with a median length of 1,607 lemmata (after removal of stopwords) and generated several sets with different numbers of topics, which offer promising starting points for further, in-depth analyses by specialists. At this point, it may only be remarked that a three-topic set yielded quite an interesting thematic division, which can very roughly be described as meditation—self-cultivation—philosophy; here it is represented by the three highest-rated lemmata in each topic:

- *samādhi* ("concentrated meditation"), *kāya* ("body"), *samaṇabhrāhmaṇa* ("ascetics and brahmins")
- *vedanā* ("feeling"), *magga* ("path"), *pajānāti* ("understand")
- *dukkha* ("leading to suffering, unsatisfactory"), *citta* ("mind"), *anicca* ("impermanent")

References

[1] Elwert, Frederik (2014). TCFnetworks. URL: https://github.com/SeNeReKo/TCFnetworks.

[2] Hellwig, Oliver (2009). SanskritTagger, a stochastic lexical and POS tagger for Sanskrit. In: *Proceedings of the First International Sanskrit Computational Linguistics Symposium*, pp. 37-46.

[3] — (2010). Performance of a lexical and POS tagger for Sanskrit. In: *Proceedings of the Fourth International Sanskrit Computational Linguistics Symposium*, pp. 162-172.

[4] von Hinüber, Oskar (1996). *A handbook of Pāli literature*. Berlin: de Gruyter.

[5] Hinrichs, Erhard W., Marie Hinrichs, and Thomas Zastrow (2010). WebLicht: Web-Based LRT Services for German. In *Proceedings of the ACL 2010 System Demonstrations*, pp. 25–29. http://www.aclweb.org/anthology/P10-4005.

[6] Knauth, Jürgen and David Alfter (2014). A Dictionary Data Processing Environment and Its Application in Algorithmic Processing of Pali Dictionary Data for Future NLP Tasks. In *Proceedings of the 5th Workshop on South and Southeast Asian NLP*, pp. 65–73.

[7] Malalasekara, Gunapala P. (1937–1938). *Dictionary of Pāli proper names*. London: Murray. Digitised version URL: http://www.palikanon.com/english/pali_names/dic_idx.html

[8] McCallum, Andrew Kachites (2002). *MALLET: A Machine Learning for Language Toolkit*. URL: http://mallet.cs.umass.edu.

[9] Oberlies, Thomas (2001). *Pāli: a grammar of the language of the Theravāda Tipiṭaka*. Berlin: de Gruyter.

[10] Rhys Davids, T. W. and William Stede (1921–1925). *The Pali Text Society's Pali-English dictionary*. Digitised version URL: http://dsal.uchicago.edu/dictionaries/pali/

[11] Schmid, Helmut (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In: *Proceedings of the International Conference on New Methods in Language Processing*, 12, pp. 44–49.

[12] — (1995). Improvements in Part-of-Speech Tagging with an Application to German. In *Proceedings of the ACL SIGDAT-Workshop,* 1–9. Dublin. URL: ftp://ftp.ims.uni-stuttgart.de/pub/corpora/tree-tagger2.pdf.

[13] Vipassana Research Institute. *Chaṭṭha Saṅgāyana Tipiṭaka* version 4.0. URL: http://www.tipitaka.org.

Flexible annotation of digital literary text corpus with RDF

Andrew U. Frank and Andreas Dittrich

Geoinformation, TU Wien
Comparative Literature, University Vienna
E-mail: frank@geoinfo.tuwien.ac.at, dittricha@gmail.com

Abstract

A corpus of text to be used for literary analysis, but generaly for other corpora in many other document oriented fields of digital humanities, must provide methods
- to add and remove texts from the corpus,
- to inquire about the texts in the corpus and
- to query the content of the corpus.

RDF to structure document storage and SPARQL as a flexible query language are suitable to build, maintain and use corpora.

We report on a system to prepare text for inclusion in a corpus for literary analysis, where text structure is annotated partially automatic and the result of linguistic analysis included in the same corpus. Analysis of the text can use SPARQL to extract text parts or produce statistics.Three experiments have shown us that requirements for (i) literary analysis of a single authors work, (ii) the analysis of specific aspects of a limited set of texts – the ontology of fairy tales – or (iii) a wide ranged, mostly statistical, analysis of a large number of texts pose the same fundamental requirements.

1 Introduction

The standardized Resource Descriptor Framework (RDF) can be used for the annotation of text corpora in the humanities. Annotations in text corpora are currently mostly done with standardized tags (e.g. Penn Treebank[11], Stuttgart Tuebingen POS codes[15]), in either sequential files with simple formats per line (e.g. the Portuguese fairy tale corpus [9]), or via XML encoding. Other important collections of texts have a nearly uniform format (e.g. the texts in the Gutenberg project) and can therefore be included in systematic studies with little manual effort required.

We started experiments for building corpora for computational comparative literature research with annotations encoded in RDF, in order to understand the practical requirements and limitations of possible technical solutions. RDF overcomes one of the most vexing obstacles of annotating complex texts with network structures. The annotation of a textual structure (sections, paragraphs, sentences...) and layout (pages) requires the combination of two hierarchies, resulting in a network. This is required in many projects where the layout of the text on the pages is relevant.

The focus of our experiments is on comparative literary studies, i.e. comparing texts for literary analysis. The requirements are somewhat different for projects where critical editions of a text are used primarily; TEI serves well for such projects, and discussions of future developments are ongoing in the TEI community[1]. We suggest RDF as a widely recognized standard which can be used to annotate text corpora in a flexible and connectable way. It has the advantage of utilizing the widely used XML format (and can be encoded in XML, if this is desired), as it is was designed to connect the extremely large number of documents on the web in a searchable network.

In the next section, the shortcomings of XML for literary annotations and the characteristics of RDF are introduced. The third section describes briefly three experimental projects, and the concluding section argues for RDF as the most promising method for flexible annotations of literary texts which can be linked to other data.

2 Annotation languages

Annotations add additional data to the text – similar to handwritten glosses in the margin of a printed volume. To make annotations useful, their contents must be structured. Two standardized methods are commonly available:

XML Markup languages go back to the early years of computer technology (e.g. RUNOFF and TEX), and were standardized and extended with the advent of the web. The eXtended Markup Language (XML) is a language that can structure text in a hierarchical tree[2]. Data in XML encoding is flexible, easy to parse, and in principle human-readable. There are many efficient tools for processing it, and tag sets can be adapted to special requirements of application domains, including, but by far not limited to, NLP and documenting different version of a text with the TEI tag set[13].

The major strength of XML, namely the hierarchcial structure which allows very efficient processing, is also its major limitation. Representing non-hierarchical situations is difficult, e.g. the structure of pages in a newspaper, each containing multiple articles or articles spanning multiple pages[3].

[1] see contributions for TEI conference 2015 by
Pierazzo, Elena: TEI: XML and Beyond; and
Ciotti / Tomasi / Vitali: An ontology for the TEI (Simple): one step beyond

[2] http://www.w3.org/TR/REC-xml/

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

RDF The extension of the web to a semantically linked set of documents[2] has resulted in a standardized method of representing arbitrary networks of linked data[10]. An RDF-encoded collection can be thought of as a labeled graph (i.e. a network), where the nodes are the objects and the labels describe the relations between the objects. The fundamental representation is the triple: object – property – value, describing a named link between two nodes. The nodes stand for uniquely identified concepts which are further described with values, which are either simple values (e.g. names encoded as character strings) or links to other objects (encoded as the identifiers of these objects)[3]. Representations which are easier for human comprehension do exist (e.g., the Turtle format[4]). SPARQL is the query language for collections of RDF triples, searching the graph. It is standardized[5], very general, and syntactically related to SQL. A number of very efficient implementations exist, which permit searches for objects fulfilling complex conditions, connecting values, and values in other linked objects. Benchmarks are reported for trillions of triples with modest hardware requirements[6].

Many resources available on the web utilize the RDF format, e.g. a large part of wikipedia (with more than 3 million concepts), a collection describing 7 million geographic features, and many others. These data can be connected – automatically or manually – with units in the literary text. The advantages of RDF for corpus annotations are:

1. RDF builds graphs and captures a network of links of non-hierarchical situations without difficulties (e.g. textual and layout relations can be represented on equal footing).

2. RDF can represent both NLP granularity (a word in a text as a linked element) and "humanities granularity" (a document, a chapter, or a text block like a paragraph).

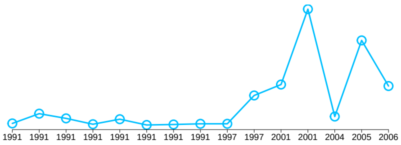
3. RDF is optimized for graph search in networks of billions of elements (many NLP operations are graph searches, e.g. finding "Hearst" (1992) patterns).

Several authors have advocated the use of RDF for NLP encoding (e.g [7]). Others have shown how the combination of NLP and RDF can be used to build resources for other disciplines (e.g. [14, 16]). A text with treebank tags, dependencies, and coreferences produces approximately 10 triples for each word, which means that current large corpora (BNC, ANC) yield less than a trillion triples when fully annotated and triplified.

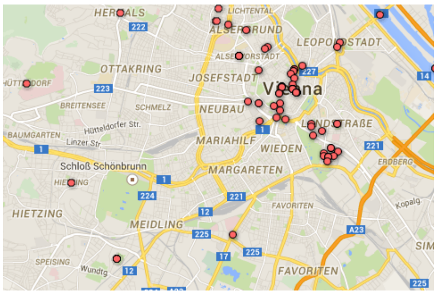
[4] http://www.w3.org/TR/turtle/

[5] http://www.w3.org/TR/sparql11-query/

[6] http://www.w3.org/wiki/RdfStoreBenchmarking
https://www.w3.org/wiki/LargeTripleStores#
Oracle_Spatial_and_Graph_with_Oracle_Database_-_1.08_Trillion_triples_.28edges.29
http://wifo5-03.informatik.uni-mannheim.de/bizer/berlinsparqlbenchmark/results/V7/



(a) A chronological ordered frequency of occurrences of real place name references



(b) Named places in Vienna mentioned in the text

3 Experiments

A text corpus in general is - in the language of software engineering - an abstract data type with methods to

- add and remove texts to the corpus,
- obtain information about the texts included, and
- query the corpus with a flexible query language.

When building an annotated corpus, one must decide (i) on the units to annotate and (ii) the information that should be annotated. The experiments we report are representative for three typical forms of inquiry of comparative literature; each motivates a specific requirement, but we find that all requirements are present in most literature studies to varying degree:

study of the *œuvre of a single author*: annotation on a high semantic level to study the use of space, time, and persons in the author's œuvre;

study of a *literary genre*: annotation of the linguistic structure of text (parts of speech, dependencies, correferences etc.) to analyze style;

large *collections of texts* with bibliographic data and annotations of different levels of granularity, to compare stylistic and vocabulary characteristics.

3.1 A corpus-based examination of the œuvre of a single writer

Goal Subject of the examination is the œuvre of author Ilse Aichinger, which is the object of the project `:aichinger:`. The goal is to briefly show (a) its conceptual framework and basis in the literary studies, and (b) its technical methods[7].

The focus of the study is on the spatiality of Aichinger's texts. In her work, "places carry the plot" [1], and places are important to her in all her writings, spanning more than half a century: they vocalize memories of her experiences during the Second World War in Vienna. "The places, which we looked at, look at us," as she writes in her prose-poetic short text "City Center" ("Stadtmitte"'[1]). It is

[7] more detail can be found at http://gams.uni-graz.at/o:dhd2015.v.032

assumed that the places trigger a process of remembrance and evoke collocated events from different periods. Nevertheless, it is only in Aichinger's later texts that references to recognizable place names become frequent, as was shown during exploratory analysis. The frequency of the occurrence of real place names (identified as words which end with "{*}gasse", "{*}straße" or "{*}platz") varies enormously throughout her work, as is shown in figure 1a.

Approach To construct the corpus systematically, we had to (a) include the text from an authoritative edition, (b) anotate each paragraph, and (c) identify places, persons and times. It seemed to us that a paragraph (or a similarly-sized unit of text) is the optimal granularity for the intended literary analysis. The text therefore had to be broken up into paragraph units, and annotations linked to paragraphs. This is in contrast to many corpora that are built to produce annotated editions of literary text and are annotated with text variants etc. linked to words or short sequences of words and tagged with TEI tags.

The steps are:

1. Obtain a text with simple character encoding in UTF8 from scanning and optical character recognition (OCR) processing.
2. Manually mark the different parts of the text as titles, subtitles, tables of content, and break text up into paragraphs.
3. Manually annotate the paragraphs for places, persons, and times mentioned.
4. Convert the text to RDF format and insert it into a SPARQL endpoint.

We found it advantagous to just annotate the persons, places, and times within the paragraphs with triples (paragraph-id, person-relation, person-id), where person-id was produced on the spot (e.g. theFishmonger). A similar procedure was used for the places and times. Adding additional details and links to other data, e.g. dbpedia, or addresses to link with Google Maps is better left for later.

Analysis is possible using SPARQL; for example, one can retrieve all paragraphs where a location is mentioned with the following query, which finds all google names gname and the title of the text tit for all locations place mentioned in a paragraph para in the text :

```
SELECT ?tit   ?gname WHERE {
?txt lit:titel ?tit .
?para lit:in ?txt .
?para lit:ort ?place .
?place lit:google ?gname .
}
```

From this data, a map can be produced using Google Maps and Fusion Tables[8]. The output for part of a volume of the work and places mentioned in Vienna is shown in Figure 1b.

[8] https://support.google.com/fusiontables/answer/2571232

3.2 Ontology in literary text

Goal We wanted to test the hypothesis that different literary texts use a different "conceptualization of part of reality" [5]. We constructed a corpus of literary texts, for which we assumed that the ontologies differ. The texts are mostly from the Gutenberg project and include fairy tales (from Grimm, Hauff, the Arabian Nights, and similar) but also other texts with non-standard ontology (e.g. science fiction). Some fairy tales differ in their ontology clearly from the real world as we experience it daily. As a first - easy - target, we want to identify the stories in which animals act as rational agents and communicate verbally. Fairy tales should prove a relatively easy genre for such an analysis, and we expect to identify the class of fairy tales where animals act as rational agents automatically.

Approach Texts are manually broken into paragraphs marked up with a schema similar to the previous project. A version of the text is marked in a way which passes transparently through the CoreNLP suite and can be identified in the output. The text paragraphs are linked with the sentence granules of CoreNLP. The translation of the XML-tagged output to RDF and loading it in a SPARQL endpoint is straightforward. Compared to the processing time required for NLP, the conversion and loading into the RDF storage is negligible (a few minutes). Most time-consuming is the linguistic parser and tagger, which for a text of 100,000 words on an (slow) PC takes approximatively one hour; the other steps require less than a minute and load 1.5 billion triples.

The ontology is hidden in the text [4], and we extract from the treebank annotations for example the capabilities of animals for verbal speech, as indication of ontological commitments in a fairy tale. The plan is to identify first (human) persons and the verbs of rational actions typically associated with them and then find texts where these verbs are used with animals as agent. We expect that linking our text with RDF-encoded POS with the RDF-encoded wordnet[9] and RDF Framenet [8] data will be helpful. The tales in which animal agents act rationally, i.e. are subject to active forms of verbs of rational action, are the desired class.

To identify the "rational animals as agents" class of fairy tales automatically and compare it with human judgment is a first step in an effort to understand the ontological differences between texts. To achieve this goal, the analysis must combine observations at every structural level of a text - from the word to groups of texts - and multiple methods of analysis (lexical, grammatical, or narrative structure).

3.3 "All methods for all texts"

Goal This experiment plans to show how to use Big Data methods for computational comparative literature. Comparative literature shuld be corpus-based in order to document the texts included in a study, and it must include a fixed set of

[9] http://wordnet-rdf.princeton.edu/

analytical methods which are applicable to the texts in the corpus. The application of a fixed collection of methods to a fixed corpus of text results in comparable and repeatable results.

A corpus from as many literary texts as we could get hold of is assembled, and a collection of digital analysis methods are applied to each text. The methods vary from simple counts of words, to methods of comparing vocabulary or use of syntactic constructions, etc. Any method published with sufficient detail to be implementable, e.g. from vocabulary or syntactic similarity, to use of space and time in the text or narratology[12].

Each method produces a characteristic of the text analyses, and for each text a vector of characterizations is obtained. Cluster analysis of these vectors results in various groupings which are then used for further studies by comparative literature researchers.

Technical solution For this project, the possibility of RDF to include bibliographic details of the texts included in the corpus is crucial for the handling of a very large number of texts, each with its own vector of characterizations. The results of NLP processing at the word and sentence level are used for style analysis; vocabulary analysis, syntactic structure statistics and similar apply to larger units.

The methods will interface the corpus with a SPARQL query; tools to access a SPARQL endpoint and to process the data are available in most current computer languages, giving maximum freedom in the implementation of observation methods. The key is the fully automated application of the methods to texts in the corpus, which can be repeated if new methods are added or existing methods are changed. This gives reliable and repeatable evaluations of the texts.

4 Conclusion

A common set of requirements emerges from the three different experiments:

- multiple granularity of the analysis and appropriate subdivisions of the text,
- network links between text units,
- automated processing with minimal human intervention only when a text is integrated in the corpus.

A corpus-based analysis has the advantage of generating reproducible results. A corpus useful for the humanities, e.g. for literature analysis, must merge the Natural Language Processing results, which are mostly expressed at the word and sentence granularity, with the structuring tools for literary texts in paragraphs, chapters etc., but also for the layout on pages. Each text must have the relevant bibliography data, and many texts must be available for concurrent analysis. A corpus for literary analysis, especially for comparative literature, must include many texts, possibly from multiple languages; a corpus for literary analysis is different from a corpus constructed for editorial work and publishing critical editions, where a single œuvre is included.

In the language of software engineering, a corpus can be seen as a abstract data type with methods to

- insert and remove texts,
- obtain bibliographic information about the texts included, and
- query the texts with a flexible query language.

The construction of corpus software can begin with readily available open-source software, mostly an RDF triple store (e.g. 4store[10]), extended with smaller, discipline specific programs to convert input documents into an RDF format with the necessary annotations. The analysis programs access the store using SPARQL in the same way administrative programs access databases using SQL and may store the results again as RDF triples in the same store. This layered structure [17] is fundamental in software engineering, and is probably the major contribution to the success of the www; it could be used to build the core software to manage corpora for diffferent application domains in digital humanities. The particulars of different domains and applications are isolated in relatively thin and inexpensive layers which are stacked on top of the foundations. For example, a program to convert the result of the Stanford NLP processor to RDF is a dozen pages long and can be reused with small adaptations for treebanks of other languages.

The generalities of many applications of corpus techniques in digital humanities have certain requirements:

- The analysis of the corpus must be possible at different levels of granularity of the text (e.g. paragraph, page, chapter, volume or multi-volume work, but also pages with their layout), and with different subsets of the corpus.
- It is necessary that treebank tags and other results from Natural Language Processing are available for each subset of a text, such that arbitrary larger subsets from the corpus can be analyzed.

It is likely that many additional widely usable tools can be identified. For example, the annotation of the Aichinger corpus could have been advanced by using a Named Entity Recognizer (e.g. the output form the CoreNLP) to annotate each paragraph with the entities automatically recognized, and to link them with a gazetteer for cartographic output or a list of places of historic times (e.g. for the medival period[11]), reducing further the amount of human-produced annotations. The same process is likely useful for preprocessing data for historic research.

[10] http://4store.org/

[11] http://www.leeds.ac.uk/arts/info/125136/international_medieval_bibliography

The potential for connection of annotations with other sources is important: in the Aichinger project, a connection to Google Maps (for the production of interactive maps), and to the movie database[12] has proven valuable in assisting the literary analysis.

Some manual preparation of texts to be included in a corpus cannot be avoided. Not only are there non-relevant text parts to delete, but there is also the implicit structure of the text with titles and subtitles, which must be marked up to be preserved in the corpus; the use of directories and subdirectories can achieve the same end, but is far less flexible. With a brief markup, taking a few minutes for a text, this structure can be identified.

We found that RDF is flexible and efficient; the openness of RDF and the ease of connecting one RDF collection with other resources is an additional benefit. We recommend standardizing a simple markup language to prepare texts for their conversion into RDF resources.

References

[1] Ilse Aichinger. *Zu keiner Stunde*. Fischer, 1991.

[2] Tim Berners-Lee, James Hendler, Ora Lassila, et al. The semantic web. *Scientific american*, 284(5):28–37, 2001.

[3] Peter Pin-Shan Chen. The entity-relationship model - toward a unified view of data. *ACM Transactions on Database Systems*, 1(1):9–36, 1976.

[4] Philipp Cimiano, Christina Unger, and John McCrae. *Ontology-based interpretation of natural language*, volume 7. Morgan & Claypool Publishers, 2014.

[5] Thomas R Gruber et al. Toward principles for the design of ontologies used for knowledge sharing. *International journal of human computer studies*, 43(5):907–928, 1995.

[6] Marti A Hearst. Automatic acquisition of hyponyms from large text corpora. pages 539–545, 1992.

[7] Sebastian Hellmann, Jens Lehmann, Auer, Sören, and Martin Brümmer. Integrating NLP using linked data. In *The Semantic Web–ISWC 2013*, pages 98–113. Springer, 2013.

[8] Nancy Ide. Framenet and linked data. In *Proceedings of Frame Semantics in NLP: A Workshop in Honor of Chuck Fillmore*, volume 1929, pages 18–21, 2014.

[12] http://datahub.io/de/dataset/linkedmdb

[9] Paula Vaz Lobo and David Martins De Matos. Fairy tale corpus organization using latent semantic mapping and an item-to-item top-n recommendation algorithm. In *LREC*, 2010.

[10] Frank Manola, Eric Miller, Brian McBride, et al. RDF primer. *W3C recommendation*, 10(1-107):6, 2004.

[11] Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330, 1993.

[12] Vladimir Jakovlevič Propp, EM Meletinskij, and Christel Wendt. *Morphologie des Märchens*. Carl Hanser Verlag, 1972.

[13] Adam Przepiórkowski. TEI P5 as an XML standard for treebank encoding. In *Proceedings of the Eighth International Workshop on Treebanks and Linguistic Theories (TLT8)*, pages 149–160, 2009.

[14] José Saias and Paulo Quaresma. Using NLP techniques to create legal ontologies in a logic programming based web information retrieval system. In *Workshop on Legal Ontologies and Web based legal information management of the 9th International Conference on Artificial Intelligence and Law, Edinburgh, Scotland*, 2003.

[15] Anne Schiller, Simone Teufel, and Christine Thielen. Guidelines für das tagging deutscher Textcorpora mit STTS. *Manuscript, Universities of Stuttgart and Tübingen*, 66, 1995.

[16] Erich Schweighofer. *Semantic indexing of legal documents*. Springer, 2010.

[17] Hubert Zimmermann. OSI reference model–the ISO model of architecture for open systems interconnection. *Communications, IEEE Transactions on*, 28(4):425–432, 1980.

Towards a Hittite Treebank. Basic Challenges and Methodological Remarks

Guglielmo Inglese
IUSS Pavia\Università di Pavia\Università di Bergamo
guglielmo.inglese01@ateneopv.it

Abstract

The creation of a Hittite treebank constitutes quite a challenging task for computational linguists, as texts require a certain amount of preliminary work on philological issues before linguistic annotation can be effectively implemented. The aim of this paper is to survey a number of problems in laying the foundation of a resource which complies both with current digital annotation standards, as provided by UD, and with the philological practices established in the field of Hittitology.

1 Introduction

In this paper, I outline the first steps towards the creation of a treebank of the Hittite language, built within the framework of Universal Dependencies (UD). Hittite is the most anciently attested Indo-European language, and as such it is of primary importance for Indo-Europeanists, as well as for scholars interested in the puzzling linguistic scenario of the Ancient Near East.

Despite the compilation of grammars of the language (cf. Hoffner & Melchert [3]), and the development of electronic resources dedicated to Hittite texts (cf. Giusfredi [1]), many linguistic issues still remain open, so that a Hittite treebank is nowadays a *desideratum* of research. Still, the peculiarities of the sources pose a number of problems to computational linguists. First, unlike modern languages with digital-born texts, such as English, for which a number of NLP tools has been developed, Hittite texts must be manually annotated. Second, up-to-date linguistic annotation following current trends in NLP should be paired with the encoding of philological notes. The importance of the interaction between these two components cannot be underestimated in building a resource able to reach an audience as wide as possible. On the one hand, the design from scratch of a Hittite treebank constitutes an interesting case study for digital humanists, as it provides important clues as to how to deal with the digital encoding of cuneiform languages.[1] On the other hand, working with UD allows one to build a resource valuable for computational linguists as well. In what follows, I present the basic issues in the annotation of Hittite, taking as a case study the so-called ‘Zalpa’s text’ (120 sentences, 1270 words).

2 Cuneiform script and philological problems

Hittite texts are written in cuneiform script, a syllabic script native of ancient Mesopotamia, which is exemplified in figure 1.

[1] Projects currently working on similar issues are the Ugaritic corpus by Zemánek [5], the *Annotated Cuneiform Luwian Texts* (http://web-corpora.net/LuwianCorpus/search/).

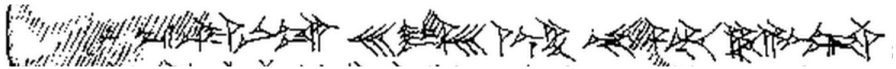


Figure 1: autography of the first line of KBo 22.2, from *HPM* [2]

This script displays a certain degree of complexity, as it contains both syllabic signs and logograms. In Hittite texts, logograms consist of Sumerian and Akkadian words used as graphic shortcuts for underlying Hittite words. To make Hittite texts available to non-specialists, cuneiform signs must be first transliterated, either in 'narrow transliteration' or in 'broad transcription' (cf. Hoffner & Melchert [3]). The former matches closely the script, with each sign transliterated as a syllable, as in *e-eš-zi* 'he is', whereas the latter consists of a phonological interpretation of words, as in *ešzi*. In addition, syllabic signs and logograms are graphically kept distinct. Hittite syllabic signs are written in lowercase italics letters, as in *e-eš-zi* 'he is', Sumerograms are transliterated in uppercase, as in MUNUS 'woman', and Akkadograms in uppercase italics, as in *TUP-PI* 'tablet'. Moreover, Sumerograms occur either in their root-form or bearing phonetic complements, i.e. final syllables marking Hittite case endings, as in DUMU-*an* 'son (acc.)'. Finally, a handful of Sumerograms, the so-called 'determinatives', were graphically preposed to nouns, and indicated the semantic class that nouns referred to. For instance, the sign URU in $^{\mathrm{URU}}$*Ne-e-ša* indicates that the noun *Ne-e-ša* belongs to class of city names.

In the treebank, words are given in broad transcription, thereby allowing users lacking a philological training to easily look into the corpus. Narrow transliteration and determinatives are stored as philological features. Also, both Sumerograms and Akkadograms are temporarily transcribed in uppercase.

Finally, one must consider the conservation status of tablets. As a matter of fact, most tablets are not entirely preserved, but rather broken or otherwise damaged. As a result, scholars usually need to reconstruct missing parts to assemble readable texts, either referring to less damaged copies, or drawing upon their expertise of the language. The conservation status of tablets brings about at least two practical issues: first, the integration status of each word should be properly annotated; second, it is necessary to develop a schema dedicated to the syntactic annotation of incomplete sentences.

3 Tokenization and philological features

Once transliteration and transcription have been performed, the first task to attend to is the tokenization. Luckily, we possess a great clue as to how to segment Hittite texts, as Hittite scribes separated words through blank spaces.

Still, tokenization cannot be restricted to the observation of blank spaces. First, one needs to split off clitics. For instance, the graphic word *nu-wa-aš-ša-an* must be split up as *nu=wa=šan*, that is, as the connective *nu* plus the particles *wa* and *šan*. Second, one needs to normalize texts by resolving elisions and assimilations, in order to improve the searchability of the treebank. Elision and assimilation often take place within clitic chains. For instance, the sequence

[2] URL: http://www.hethport.uni-wuerzburg.de/hetkonk/hetkonk_abfrage.php

an-da-ma-pa should be split as *anda=m=apa*, with *m* being the enclitic pronoun *mu* 'me', which undergoes elision before the particle *apa*. Assimilation takes place in sequences such as *n=at=ši*, in which the pronoun *at* 'it' assimilates to the pronoun *ši* 'him', thus yielding the graphic word *na-aš-ši*. Once normalized, the raw text is converted to CoNLL-U and further annotated.

The CoNLL-U format employed includes ten fields for each word line: ID, FORM, LEMMA, CPOSTAG, POSTAG, FEATS, HEAD, DEPREL, DEPS, MISC. Under FORM, I put words in broad transcription. Within this field, the two operations of clitic splitting and normalization are performed thanks to the token vs. word indexation. Each multiword is tokenized as one word, and it is indexed with integer ranges, whereas words are indexed with simple integers.

Moreover, philological data discussed in the previous section are associated to each word in the form of philological features, which are stored in the MISC field. First, I add three language-specific close-ranged philological features: Integration=0,1,2; Language=Hitt, Sum, Akk; Determinative=1-16. The Integration feature takes three values, and indicates whether a word is actually attested on the tablet (0), or, if restored, whether it is restored after other copies (1), or by the editor himself (2). The Language feature indicates whether cuneiform signs should be read as Hittite (Hitt), Sumerian (Sum), or Akkadian (Akk). Furthermore, determinative signs are handled thanks to the feature Determinative, which takes as value a numeric code corresponding to a specific determinative sign, based upon the list in Hoffner & Melchert [3]. Note that, since determinatives are stored as philological features, they do not visually surface in the treebank as tokens. Second, I add two open-ranged features, that is, Ntrans and HLemma. The former takes narrow transliteration of words and multiwords as value, whereas the latter indicates the corresponding Hittite lemma of logograms, if available. It should be stressed that these features do not aim at a full coverage of all issues in Hittite philology, but merely at providing users with notes essential for a proper reading of Hittite texts.

4 Morphological annotation

UD employs a three-layered model of morphological annotation, whereby each entry is lemmatized, assigned a POS label, and tagged for its morphological features. First, each word is given a LEMMA, based on dictionaries such as Tischler [4]. Note that I lemmatize Sumerograms and Akkadograms with Sumerian and Akkadian words, and store available Hittite equivalents through the Hlemma feature. Also, since forms can instantiate different lemmas, as in the case of *iyanzi*, which is a form of either *iya1-* 'go' or *iya2-* 'make', the LEMMA field can host multiple lemmas, separated by simple comma. POS tagging will be done in accordance with UD guidelines. Of the POS tags featured in UD, only PUNCT is left out, as Hittite texts display no punctuation.

The finer-grained morphological analysis is carried out through the use of morphological Feature=Value pairs, encoding both lexical and inflectional features of words. A general problem one is faced with in morphological annotation is that a number of features cannot be inherently assigned to single tokens. As an example, let us consider the feature Aspect. Hittite displays the well-known IE verbal derivational suffix -*šk*-, which is often associated with imperfective aspect (Hoffner & Melchert [3]). Still, not every *šk*-suffixed verb is imperfective, nor unsuffixed verbs are always perfective. As a result, aspectual interpretation of verbs cannot simply rely on morphological marking. In the treebank, I prefer to exclude such borderline features, as they ultimately depend on the linguist's judgment on specific tokens, and their annotation is thus liable to a high degree of inconsistency.

I adopt the following UD lexical features: NumType=Card, Ord; Poss=Yes; PronType=Prs, Int, Rel, Dem, Tot, Neg, Ind. Note that though lexical features are inherent features of lemmas, this is not always the case, as for the pronoun *kuis* 'who', which is either interrogative, relative, or indefinite, depending on the context. To these, I add the language-specific feature Clitic, which indicates whether a token constitutes an independent word or a clitic item. Finally, I leave out the Reflexive feature, as reflexivity is not morphologically marked.

UD includes both nominal and verbal inflectional features. Nominal features adopted for Hittite are: Gender=Com, Neut, Masc, Fem; Number=Sing, Plur; Case=Nom, Acc, Dat, Gen, Voc, Inst, Abl, Dir, Erg; Definiteness=Red. Remarkably, not all UD features are relevant for Hittite. For instance, neither Animacy nor Degree have been included, as neither of them is morphologically coded. As for Gender, to the bipartite Com vs. Neut Hittite system, I added the Masc and Fem values, in order to account for the distinction in Akkadian pronouns such as =*ŠU* 'his' vs. =*ŠA* 'her'. Among the cases, it is much disputed whether the -*anza* ending on neuter nouns is an ergative case ending (Goedegebuure [2]). I do not take a stand in this scholarly debate, but for our purposes, it is more parsimonious to treat this ending as an ergative case, since treating it as a derivational suffix would artificially increase the amount of lemmas in the treebank. Sumerograms lacking phonetic complements are not tagged for Case. Definiteness of noun phrases is not explicitly marked in Hittite, but the Red value is retained to mark the reduced state of Akkadian nouns.

The verbal features adopted include VerbForm=Fin, Inf, Sup, Part; Mood= Ind, Imp; Tense=Pres, Past; Voice=Act, Mid; Person=1,2,3. Not every verbal features available in UD has been employed, as for Aspect, and some features display a reduced range of values. For instance, only indicative vs. imperative, and present vs. past distinctions are morphologically encoded. As for voice, only a two-fold opposition active vs. middle is marked by verbal endings.

5 Syntactic annotation

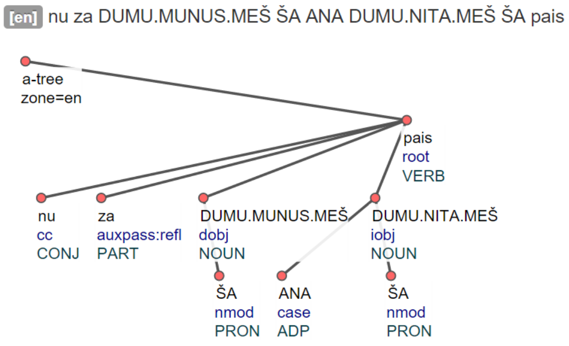
In UD, syntactic relations are represented as dependency relations between words, and are stored in the HEAD and DEPS fields. UD's universal set of dependency relations has been expanded with the following language-specific relations: *acl:relcl*, *advmod:emph*, and *auxpass:refl*. Moreover, I introduce the newly created relation *advmod:loc* to annotate Hittite so-called 'local particles'.

For reasons of space, I focus here on the annotation of complete sentences only, and leave the annotation of broken sentences for further research.

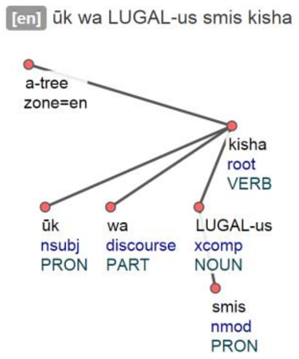
5.1 *Clausal predicates and core arguments*

Predicates constitute the head of the predication, and as such receive the *root* relation. When predicates are omitted, the *root* label is conventionally assigned to the first word in the sentence. In nominal predications, the complement of the copula takes the *root* relation, and the verb ‘be’ is tagged as *cop*.

Non-clausal core arguments of the predicate are tagged as *nsubj*, *dobj*, and *iobj*, as in (1) and (2). In my corpus, clausal arguments are not attested. Secondary predicates of predicative verbs are tagged as *xcomp*, as in (2).



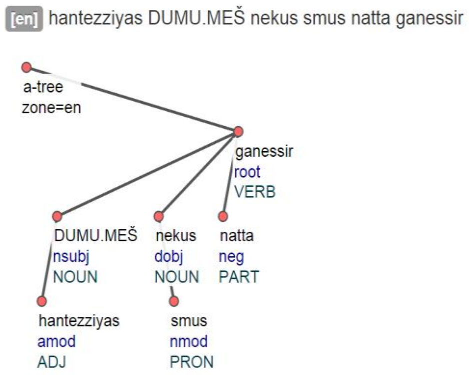
(1) *Core arguments of clausal predicates*
“And she gave her own daughters to her own sons.”



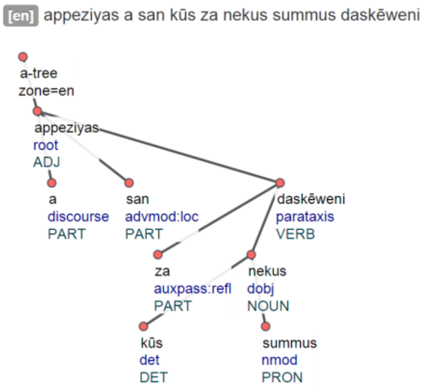
(2) *Secondary predicates*
“I have become your king.”

5.2 *Noun dependents*

In the treebank, noun dependents include adjective (*amod*) and determinatives (*det*), as in (3) and (4). Numeral modifiers (*nummod*) are attested as well. Nouns can be modified by nominal modifiers (*nmod*) as well (cf. sec. 5.3).



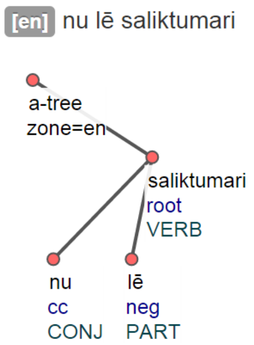
(3) *Adjectives*
“The older sons did not recognize their sisters.”



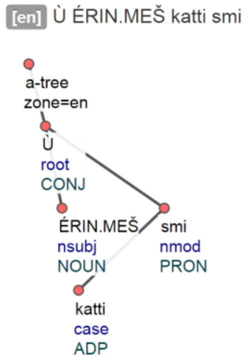
(4) *Determinatives*
“But the first one (said): we are taking these sisters of ours (in marriage).”

5.3 *Non-core predicate dependents*

Predicates can be modified by several non-core dependents, such as negation (*neg*), as in (5), or other adverbs (*advmod*). Note that preverbs are consistently tagged as *advmod*. Bare nominal modifiers are tagged as *nmod*, and adpositions depend on their nominal or pronominal head through the *case* relation, as in (6). The same annotation is adopted for Akkadian prepositions.

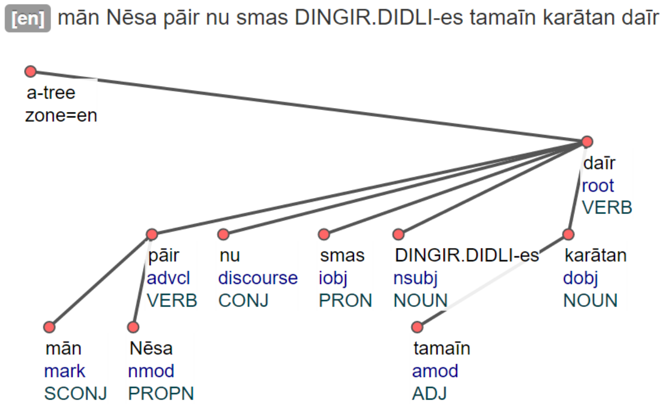


(5) *Negation*
"Now do not commit an outrage!"



(6) *Nmods and adpositions*
"And the troops (are) with me."

Clausal modifiers include adverbial subordinate clauses. In UD, subordinators depend on the predicate of the dependent clause as *mark*, which in turn depends on the predicate of the main clause as *advcl*, as in (7).



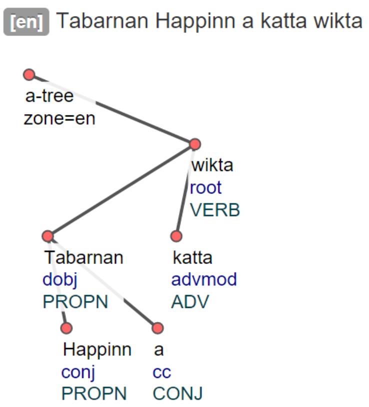
(7) *Adverbial clauses*
"When they went to Nesa, the gods gave them a different appearance."

5.3.1.Clause-linking devices

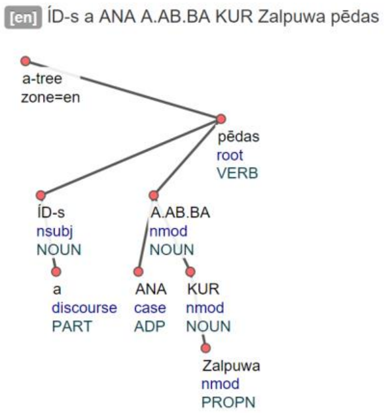
Hittite displays a number of non-subordinating connective devices, that is, sentence initial connectives *nu*, *šu*, and *ta*, and enclitic *=(y)a* and *=(m)a*.

The only proper coordinative conjunction is *=(y)a*, which links both sentences or phrases, as in (8). In UD, the first coordinand is annotated as the head of the coordination, on which the second one depends as *conj*. The conjunction depends on the first coordinand as well, and takes the *cc* relation.

The function of enclitic *=(m)a* is disputed. It arguably serves either as a topic-switching device or as a generic adversative connective. This difference is reflected in the annotation, as *=(m)a* depends via the *discourse* relation either on a topicalized noun, as in (9), or on the main predicate.



(8) *Coordinating* =(y)a
"And he demanded Tabarnas and Happis."

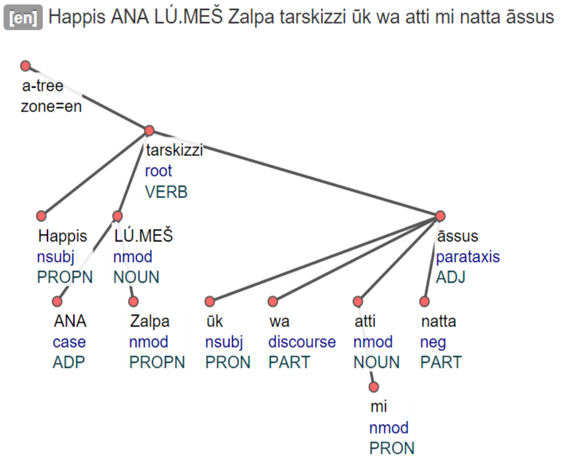


(9) *Enclitic* =(m)a
"But the river took (them) to the sea at Zalpuwa."

Unlike coordinating *=(y)a*, sentence initial connectives *nu*, *šu*, and *ta* establish additive links between sentences. The annotation mirrors these semantic and syntactic peculiarities. Sentences featuring connectives are always treated as independent from each other, and connectives are annotated as paragraph-initial conjunctions, that is, they depend upon the main predicate via *cc*, as in (5). In addition, connectives can be placed at the juncture between a preposed subordinate clause and its main clause. When this is the case, connectives depend on the predicate of the main clause as *discourse*, as in (7).

5.4 Direct speech

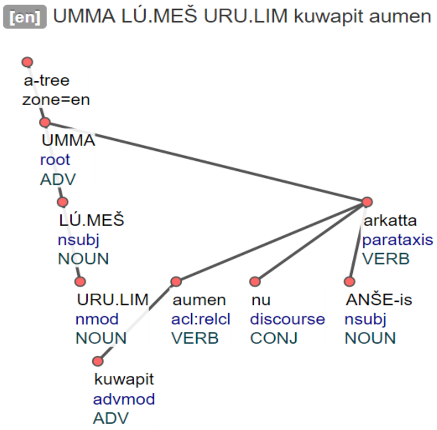
Hittite makes extensive use of reported direct speech, marked by the clitic particle *=wa(r)*. The particle depends as *discourse* on the predicate of the reported speech, which in turn depends via *parataxis* on the predicate of the main clause, as in (10).



(10) *Direct speech*
"Happis says to the men of Zalpa: «I (am) not dear to my father»."

5.5 *Relative clauses*

UD’s *acl* relation is not suitable for the annotation of most Hittite relative clauses, as they usually do not modify a lexical head. Therefore, they are better treated as parallel to adverbial clauses, with their predicate depending on the predicate of the main clause via the *acl:relcl* relation, as in (11). Relative pronouns and adverbs depend on the predicate of the relative clause. In my corpus, only relative clauses introduced by the adverb *kuwapit* ‘wherever’ are attested, but this schema can also be extended to other kinds of relative clauses.

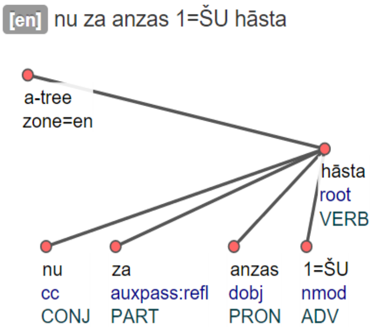


(11) *Relative Clauses*

“The men of the city (speak) as follows: «Wherever we have looked, a donkey will *arkatta*».”

5.6 *Reflexive* =za

The so-called reflexive particle *=za* occurs in a number of contexts. First, it occurs with reflexive middle verbs and with transitive verbs indicating some sort of subject involvement. Moreover, in some cases it slightly modifies the semantics of the main predicate, or it occurs in nominal predications adding no detectable semantic contribution (Hoffner & Melchert [3]). Given the uncertainty in assigning *=za* a semantic and syntactic role, I conventionally annotate it as depending on the main predicate via the language-specific *auxpass:refl* relation, as in (12).

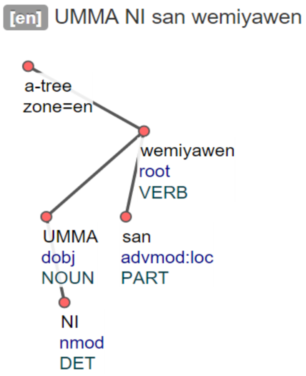


(12) *Reflexive* =za

“And (she) bore us at one time.”

5.7 Local particles

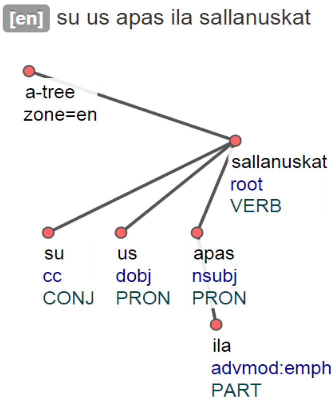
The cover term ‘local particles’ is employed to refer to the clitic particles =*an*, =*apa*, =*ašta*, =*kan*, and =*šan*. These particles arguably modify either the verb or some other local expression by adding spatial information, though this is much disputed. Therefore, I take them as depending either on the predicate or on adverbial modifiers via the newly created *advmod:loc* relation, as in (13).



(13) *Local particles*
“We found our mother.”

5.8 Emphatic particles

Enclitic particles =*pat* and =*ila* give emphasis of some sort to nouns and pronouns which they are attached to. Both particles are annotated as depending on their phonological hosts as *advmod:emph*.



(14) *Emphatic particles*
“And she brought them herself.”

6 Metadata

CoNLL-U format licenses the insertion of metadata in the form of comments to sentences. This proves extremely useful to store a number of textual information. So far, I have included sentence ID, reference to the text and the tablet that the sentence belongs to, place of retrieval of the tablet, dating of both the tablet and the text, and possible translations. Note that these data differ in scope from the philological features discussed in section 2, as they concern the text in general, and it is consequently more parsimonious to store them as comments to sentences. I am well aware that these metadata could be more fruitfully stored by editing Hittite raw texts adopting TEI guidelines,[3] but this constitutes a long-term task which goes beyond the initial stage of this project.

7 Conclusion and future work

In this paper, I have addressed some preliminary issues in designing a treebank for the Hittite language built within the framework of Universal Dependencies. Crucially, in order to grasp all relevant linguistic and textual features of Hittite, UD's template needs to be enriched with a number of language-specific dependency relations and morphological features. Also, the need to add a set of new features dedicated to the encoding of philological data has emerged.

This paper constitutes only the first step of a larger project, and much work still needs to be done. First, a schema for the syntactic annotation of broken sentences should be worked out, along the lines of what discussed by Zemánek [5]. A simple solution would be to treat gaps in tablets as instances of ellipsis, but the relation *remnant* employed to annotate ellipsis in UD is seemingly not suitable for the purpose. An alternative option could be the insertion of empty nodes in the expected position of missing words, but the insertion of empty nodes would go against the very tenets of UD, and should be avoided.

Finally, it should be stressed that the annotation outlined throughout this paper is based upon a text written in the oldest variety of the language, that is, Old Hittite. It would be thus intriguing to test whether this annotation holds for more recent phases of the language, that is, Middle and New Hittite, in which a number of significant morphological and syntactic changes have occurred.

References

[1] Giusfredi, F. 2014. Web resources for Hittitology. *Bibliotheca Orientalis* 71: 358-362.

[2] Goedegebuure, P. 2013. Split-ergativity in Hittite. *Zeitschrift für Assyriologie und vorderasiatische Archäologie*, 102 (2): 270-303.

[3] Hoffner, H. A. & Melchert, C. H. 2008. *A Grammar of the Hittite Language. Part I: Reference Grammar.* Winona Lake: Eisenbrauns.

[4] Tischler, J. 2001. *Hethitisches Handwörterbuch. Mit dem Wortschatz der Nachbarsprachen*. Innsbruck: Institut für Sprachwissenschaft.

[5] Zemánek, P. 2007. A Treebank of Ugaritic. Annotating Fragmentary Attested Languages. In *Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories*, De Smedt, K, Hajič, J & Kübler, S. (eds.), 212-218. NEATL: Bergen.

Universal Dependencies, <https://universaldependencies.github.io/docs/>

[3] URL: http://www.tei-c.org/index.xml

Corpus-based Research in Computational Comparative Literature

Christine Ivanovic and Andrew U. Frank

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

Abstract

Comparative literature investigates texts simultaneously in multiple languages. It is by definition conducting corpus-based research (each study is based on a quantitatively discernible number of texts chosen by single sets of qualities), in practice however it is facing diverse limitations which restrict the value of generalized statements. We propose to extend well approved methods of corpus-based research to computational comparative literature in order to overcome those limitations. The core is the construction of a large corpus of annotated literary texts with a related set of computational analysis methods. The aspects of the proposed technical solution make it even suited for the comparison of multi-lingual text corpora, as the same methods can be applied to all texts irrespective of language. Computational comparative literature has a supportive function for established approaches in comparative literature. It allows for systematic evaluation of a large number of texts that goes beyond established questions.

1 The Task of Comparative Literature

Comparative literature designates an academic discipline within the humanities characterized by its specific way of investigating literature (as well as other media) by comparison. Comparison however is widely understood as one of the most basic cognitive practices, and nowadays it can be performed even by machines; yet it has proven indispensable for all scientific approaches in all fields of research. Comparative literature as a discipline has to clarify why it considers comparison its basic task by definition, as well as explain the means and the goals of its specific approach.

As an academic discipline, comparative literature emerged relatively late. It arose as an attempt at comparing the observations and results of at that point already highly sophisticated philogical works in different fields of national literatures. Goethe's concept of "World Literature" is considered one of the initial terms for the development of Comparative literature while H.M.Posnett was the first to publish a book length study under this title in 1886; cf. Damrosch [7]. Comparative literature thus implies not only a comparison of texts, but a reconsidering of long established categories that are held to be determinative for literary texts, such as language, genre, nation, and others.

By investigating texts in different languages, or from different cultural, national, ethnical, religious, historical, or even media contexts, comparative literature is conceived of as a way of either transcultural or crosscultural examination. It asks for qualities of texts which are considered comparable throughout all literatures, without regards to (cultural) space and time, as well as for the manifold ways of transferring (cultural) concepts through space and time. Comparative literature is therefore as much about texts and concepts as it is about contexts and relations (which also includes investigation of the relationship between literature and other media, e.g. arts or film).

To compare in comparative literature means to simultaneously consider texts which must belong to different systems, and which are related to or referring to different contexts. Additionally, the comparison is also determined by the whole spectrum of theoretical approaches typical for the humanities, for example aesthetic and/or literary theory, history, sociology, political, cultural, religious or gender studies, and many others[2].

The analytical goal of comparative literature is not the hermeneutic interpretation of one single text, but (a) the determination of constant characteristics of (literary) textuality, (b) the determination of historical changes of (literary) textuality, (c) the discerning of manifest proof of cultural contact (dependencies), and (d) the evaluation of literary representations and their references to “the (social, economic, political or religious ..) world”. In other words: Comparative literature investigates texts simultaneously in multiple languages in order to better understand cultural progress and cultural systems as they are represented in texts.

Comparing texts is the exclusive means considered appropriate to reach this goal. To that end, several steps must be performed one after the other: (a) determination of text characteristics; (b) investigation of the same categories of text characteristics in multiple texts in different languages; (c) measuring of the specific qualities of text characteristics of the investigated texts; (d) evaluation of similarities and differences of characteristics of texts formulated in different languages and situated in different contexts; (e) evaluation of the different reference frames of the text or the text characteristics, respectively.

2 Corpus-based research in Comparative Literature

Comparative literature is based on comparing selected characteristics in a specified number of texts. As a starting point, each study in Comparative literature has to determine (a) the selection of texts to be compared, and (b) the characteristics to observe. Comparative literature is very much corpus-based, i.e. based on a quantitatively discernible number of texts chosen by single sets of qualities. The analysis depends on the type and number of the selected and analyzed texts as well as on the type of the characteristics observed.

If a study aims at investigating e.g. the importance of Dante Alighieri's *Divine Commedy* in respect to James Joyce's *Ulysses*, the comparison can be restricted to those two texts. But it might also be argued that including Homer's *Odyssee* as well as Vergil's *Aeneid* among other reference texts important to both of them into the corpus would also be advisable.

If another study investigates e.g. the typical subject of the "European Novel", the number of texts to be included as well as the selection of characteristics are less easy to determine [12]). The decision depends on prior concepts outlining which texts are considered to be a "novel" and which texts belong to "European literature". Likewise, the number of texts which by definition should be included may exceed the texts which are de facto accessible for investigation. As in the case of the study on Joyce's novel, it must be explicitly argued which texts are included and by which qualities they are chosen over other possible texts. The same goes for studies in typology, image studies, influence studies and other fields of comparative literature.

Comparative literature has to be corpus-based when it aims at generating (universally) valid results from the comparison of texts. But in order to conduct valuable corpus-based research, it is not just important to construct well-argued text corpora suited for this research; investigating the categories and concepts which constitute a corpus in literary studies is also task to be solved by the discipline.

3 Computational corpus-based research overcoming the current limitations for comparative literature by scale

The tasks of comparative literature exceed the capacities of one single investigator. In theory, all (literary) texts ever written are potentially the object of comparative literature. In practice, however, only a limited selection of texts is evaluated. Traditional comparative literature is limited by the abilities, time, and resource constraints of the researchers. Proficiency in many different languages, cultural systems, and literary traditions is an inevitable precondition for a scholar engaged in comparative literature. But even if the scholar is broadly educated or working in a well-prepared team of investigators, and even if the study is restricted to a relatively small number of languages, as it was the case in our example of the "European Novel", the number of literary texts which should be included may be abundant, and beyond what is manageable even in a lifetime of a researcher. The effect of such restrictions is the approximative nature of all such investigations, based on a more or less personal and biased choice and a relatively small number of texts evaluated; that is the regular case, not an exception. As comparatists, we most often conduct (case) studies on a very limited textual base - with accordingly limited results.

For those reasons research in comparative literature only recently turned to considering a corpus-based computational approach in comparative literature[1]. It makes explicit, in the construction of the corpus and in the identification of observations, which texts and which characteristics will be used for comparison. One difference to current practices is the ways in which the selection of texts and characteristics are being documented. The extension of textual comparability by enlarging the number of characteristics to be compared might be another advantage. Computational methods are intended to overcome current limitations in comparative literature especially in two regards:

Problem of subjective selection and circular logic The subjective selection by the individual scholar is hard to justify objectively. The selection criteria and the result of a study are often directly related. For example, in a study to describe the characteristics of the "European Novel", decisions on text selection are necessary. The choice of which texts to include anticipates the results of the study. This is especially problematic when following the classical credo of comparative literature: researchers are expected to only analyze texts that they can read in the "original language".

In order to perform a study on e.g. railway novels, one would recollect all the texts encountered so far mentioning railways, and then search for more texts which might fit. Then a selection of the most appropriate pieces is subjected to closer examination. This mode of operation is corpus-based, but the corpus is not clearly determined. It consists of the texts already known by the investigator, and of texts searched for more or less systematically. Using this method, it is very difficult to develop or argue the criteria which lead to the choice of the texts used for closer examination.

A corpus-based computational approach would evaluate texts not through the personal choices of the investigator, but determined by the construction of the corpus. If using e.g. the Austrian Academy Corpus [2] in the investigation of railway novels, a selection of over 6 million tokens from representative German language texts between 1848 and 1989 would be analyzed. This enables valid statements concerning this corpus. The findings can be compared with results based on another comparable text corpus in one or more other languages, and reproduced by other investigators. Researchers are not required to read all texts in all languages, but nevertheless will be able to compare them. It is evident, that with another corpus different results could be obtained. How to conduct a valid, i.e. a unbiased corpus is worth further studies. For now, if results for similar studies executed with different corpora coincide one might assume that the corpus is unbiased (or both corpora are similarly biased).

[1] Computational comparative literature as an academic subdiscipline in Digital Humanities does not yet exist. A recent handbook edited by Behdad and Thomas at least mentions "Comparative Data Studies" in its Chapter 13: "Comparative Literature in the Age of Digital Humanities: On Possible futures for a Discipline". [3]

[2] http://www.aac.ac.at/

Problem of capacity Any scholar can only analyze a limited number of texts, which limits comparative studies to very small subsets of the vast number of literary texts ever written. This problem is especially pertinent to the aim of comparing texts in virtually all existing languages. Traditionally comparative literature rarely includes texts in so called “small” languages (e.g. Finnish, Urdu). Comparisons between texts in non-related language systems (e.g. Chinese to Arabic) are also conducted far less than those between “European” languages (e.g. English to French).

Intended Solution Currently, the vast majority of digital literary studies is focused on digital editing. There is a rising number of national initiatives to digitize huge text corpora in the respective languages. There are also Google Books as well as crowdsourcing initiatives like Project Gutenberg whose aim is to digitize more and more texts in many languages. As comparatists we should be prepared to harvest these treasures in order to seriously evaluate “global” or “world literature”. Computational methods help to overcome some of these limitations:

- A digitally processed corpus includes a fixed, known set of texts.
- Digital text processing makes it possible to collect and analyze large text corpora / large amounts of data in any language, far beyond what an individual researcher can digest.
- Conclusions in a study can be corroborated in repeated studies and further checked with larger corpora.
- The selection of the corpus is logically separated from the conclusions.
- Results of corpus-based research are valid “for all texts included” and justify negative statements (“no text in the corpus has property x”) - comparable negative statements are not justified without a fixed corpus.

4 Challenges for corpus-based computational comparative literature

The challenges for a computational approach to comparative literature research are twofold: (i) design an efficient structure for the corpus and its maintenance; and (ii) construct the computational methods to evaluate the texts in the corpus systematically and report results suitable for analysis.

The overall design must make it easy and fast to add a text to the corpus, or add or change one of the methods used for the evaluation of a text. The collection of results must be maintained up to date, which means: (a) if texts are added, all the methods must be applied automatically to them; and (b) if methods are added or changed, the new method must be applied automatically to all texts. Such a system of maintaining a literary corpus can be realized with today's information technology.

Construction of the corpus Adding a text to the corpus means to bring it into a form the automatic methods for evaluation can use it. Initially, for each text in the corpus must be available:

- bibliographic detail, at least a BibTex entry from which author, date of publication, edition etc. can be retrieved but better a description following the Dublin Core standard,
- the text with markup for the text structure, non-literary text parts etc.
- a treebank tagged version, preferably with dependency trees (parse trees), coreferences and named places recognized.

The markup of the text identifies the structure of the text in parts, chapters, paragraphs (or whatever a text uses), and the layout of the text on pages. It is also necessary to identify the title page, author names, and other material which appears in a printed book and identifies the source of the text but is not part of the literary text properly and should not be included in the analysis. The text should be in UTF-8 encoding to allow for non-alphabetic language texts, including Chinese, Arabic etc., and processed with a Natural Language Processor to get parts of speech tags, named entities etc. recognized. Markers are to be placed into the text such that the tagged text can be connected with the original text structure. Eventually, sentiment analysis should be included.

The marked-up text and the output from the speech tagger are integrated into a single framework. We currently experiment with RDF[11], which seems to fulfil the requirements and is able to deal with the amount of data expected. Our observations indicate that for each word in a text with rich linguistic annotations, we obtain 10 RDF triples; for a literary text of 100,000 words we therefore obtain a million triples, or for a corpus of 10,000 books, 10 billion triples, which is well within the possibilities of today's RDF storage systems (benchmark results report that 1 trillion triples load in few hours [3]). Response times for different types of queries are acceptable even for very large data collections (for example dbpedia with 3 billion triples[4]), and hardware requirements are within the reach of a properly funded project (the test configuration for the above mentionmed project was acquired for Euro 70,000 in 2012).

Collection of methods The pertinent literature reports a large number of methods to evaluate texts, many of which are available on the web. Corpus-based examination of literary texts allows us, for example, to identify citations or reuses [10].

[3] http://wifo5-03.informatik.uni-mannheim.de/bizer/berlinsparqlbenchmark/results/V7/

[4] https://en.wikipedia.org/wiki/DBpedia

It serves for authorship attribution, for recognition of author gender/gender difference [13], or for the identification of literary movements [1]. It supports pattern discovery and similarity computation [14]. It has also proven useful in detecting emotions[8], or in knowledge mining[6]. If a well-structured corpus is available, new methods can be realized, for example:

- compare temporal sequences of actions with the sequence of narration in the text,
- spatial reference in the text: mapping the named places, amount of "movement" in the text, use of locations to convey e.g. sentiment, use of locations to link persons, etc.
- number of persons acting in the text and distribution of actions to them (hero-centered text, Bildungsroman, ..),
- vocabulary: literary sense vs. metaphorical, size of the vocabulary.

Each method produces for each text a value - True/False, a count, or a statistical value. For each text a vector of values represents the result of applying all methods to this text; this vector characterizes the text in the corpus.

Cluster analysis Applying all methods to all texts produces a matrix of evaluation values; a vector of values characterizes each text. Cluster analysis applied to these vectors indicates which texts are in some respect similar to each other. The resulting similarity structure can be mapped and analyzed in two directions, for example:

- Reconstruction of traditional genre attributions for the texts. We ask which evaluation methods give similar results for a group of texts which are considered of the same genre, and gain an understanding of the connection between traditional literary research approaches and the computational methods
- Reconsidering literary genres and groups: Analyzing groups of texts which appear similar on some evaluation methods but are not typically thought of comparable (for example, compare the *Arabian Nights* with Boccaccio's *Decamerone*).

From the analysis of the results, new and refined ideas for evaluation methods will result, refining the computational characterizations of texts.

5 Examples for corpus-based computational comparative literature

Moretti's *Atlas of the European Novel* In comparative literature, computational corpus-based text examination is still in its infancy. Franco Moretti has been one of the pioneers of the field. His “Atlas of the European Novel” [12] from 1999 was one of the first studies that outlined some of the possible directions comparative digital literary analysis might take. It was the starting point of corpus-based research methods still to be developed in computational comparative literature. Moretti has developed parameters by which the analysis of a large number of texts could reveal connections between variables that have so far never been related to one another, for example: language and distribution of a novel in space and time (including distribution as a translation), diegetic space of a novel, characteristics specific to the sub-gerne (adventure novel, ghost story, Gothic novel...).

Novel networks In the scope of a current project conducted on the Austrian Academy Corpus (AAC hosted at the ICLTT/Austrian Academy of Science: [4, 5]) as well as using other corpora, we plan to analyze the connection between the development of the railroad network from the second third of the 19th century until today in relation to the structural development of the (European) novel.

Multilingual corpora With new computational methods, comparing texts in multilingual corpora becomes possible. By using the approach outlined above of “all methods for all texts”, structural text characteristics in particular can be compared beyond language borders. Possible research questions include:

- the evaluation of language characteristics, e.g. a comparative analysis of the proportion of multilinguality in texts,
- stylometry irrespective of language, for example, which characteristics of a style are preserved when translations are published [9]
- the structure of character depictions: the comparative analysis of the number of characters in texts in relation to the texts time of writing, text category, and text complexity; comparative analysis of character depictions (how they are being described through indirect speech, through attributes, through concrete action, through movements in space, etc.),
- the grounding of the plot through identifiable place names or historical dates.

6 Conclusion

Corpus-based research in Computational Comparative Literature has a supportive function for established approaches in Comparative literature. It allows for systematic evaluation of a large number of texts that goes beyond established questions. The aspects of the proposed technical solution make it even suited for the comparison of multi-lingual text corpora, as the same methods can be applied to all texts irrespective of language. In the terminology of software engineering, a corpus is an abstract data type (object) with methods to

- add and remove texts,
- inquire about the texts included, and
- query the corpus with a flexible query language.

For each discipline the units of texts which are annotaded and the annotation content must be adapted to the specific tasks; for many disciplines, an automated production of annotations is desirable to achieve reproducible results. The annotations inserted into the text determine the queries which can be asked - or reversed, what must be annotated follows from the expected queries.

The core is the construction of a large corpus of literary texts with sub-corpora and a related set of computational analysis methods. Each text is systematically analyzed with all methods and we obtain for each text a comparable collection (a vector) of characterizing values. Such a combination of corpus and method collection is feasible with today's information technology, and the text resources are available on the web. All methods for all texts!

References

[1] Diego R. Amancio, Osvaldo N. Oliveira Jr., and Luciano da F. Costa. Identification of literary movements using complex networks to represent texts. *CoRR*, abs/1302.4099, 2013.

[2] Emily Apter. The translation zone: A new Comparative literature. Princeton University Press 2006.

[3] Ali Behdad and Dominic Thomas. A companion to comparative literature. John Wiley & Sons 2011.

[4] Hanno Biber and Evelyn Breiteneder. Fivehundredmillionandone tokens. loading the AAC container with text resources for text studies. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012), Istanbul, Turkey, May 23-25, 2012*, pages 1067–1070, 2012.

[5] Hanno Biber, Evelyn Breiteneder, and Karlheinz Mörth. Words in contexts: Digital editions of literary journals in the "aac - austrian academy corpus". In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco*, 2008.

[6] Amedeo Cappelli, Maria Novella Catarsi, Patrizia Michelassi, Lorenzo Moretti, Miriam Baglioni, Franco Turini, and M. Tavoni. Knowledge mining and discovery for searching in literary texts. In *Proceedings of the Third International Conference on Language Resources and Evaluation, LREC 2002, May 29-31, 2002, Las Palmas, Canary Islands, Spain*, 2002.

[7] David Damrosch. Rebirth of a Discipline: The Global Origins of Comparative Studies. In Comparative Critical Studies, 3(1):99–112, 2006

[8] Daniel Dichiu, Ana Lucia Pais, Sunita Andreea Moga, and Catalin Buiu. A cognitive system for detecting emotions in literary texts and transposing them into drawings. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, Istanbul, Turkey, 10-13 October 2010*, pages 1958–1965, 2010.

[9] Umberto Eco. *Dire quasi la stessa cosa*. Bombiani, 2003.

[10] Jean-Gabriel Ganascia, Peirre Glaudes, and Andrea Del Lungo. Automatic detection of reuses and citations in literary texts. *Literary and Linguistic Computing*, 29(3):412–421, 2014.

[11] Frank Manola, Eric Miller, Brian McBride, et al. RDF primer. *W3C recommendation*, 10(1-107):6, 2004.

[12] Franco Moretti. *Atlas of the European novel, 1800-1900*. Verso, 1999.

[13] Urszula Stanczyk. Recognition of author gender for literary texts. In *Man-Machine Interactions 2, Proceedings of the 2nd International Conference on Man-Machine Interactions, ICMMI 2011, The Beskids, Poland, October 6-9, 2011*, pages 229–238, 2011.

[14] Masayuki Takeda, Tomoko Fukuda, and Ichiro Nanri. Mining from literary texts: Pattern discovery and similarity computation. In *Progress in Discovery Science, Final Report of the Japanese Discovery Science Project*, pages 518–531, 2002.

Detecting Reuse of Biblical Quotes in Swedish 19th Century Fiction using Sequence Alignment

Dimitrios Kokkinakis
Department of Swedish
University of Gothenburg
Sweden

Mats Malm
Department of Literature, History of Ideas, and Religion
University of Gothenburg
Sweden

Email: first.last@gu.se

Abstract

Text reuse, a form of text repetition, recycling or borrowing, is a theoretically and practically interesting problem that has attracted considerable attention during the last years e.g. in the cultural heritage context (historical and comparative linguistics); in the context of social network propagation of ideas and in the measuring of journalistic reuse. In this paper we briefly outline and experiment with a method used for biological sequence alignment that has been also used in humanities research for e.g. the detection of similar passages in the complete works of Voltaire and 18th century French encyclopedias or for tracing how and which ideas spread in 19th century US-newspaper collections. We use available software (text-PAIR: Pairwise Alignment for Intertextual Relations) and experiment with the Charles XII Bible translation into Swedish, completed in 1703, against the content of the Swedish prose fiction 1800-1900, in order to automatically detect passages taken from this particular Bible translation in the prose fiction corpus.

1 Introduction

The notion of intertextuality, coined by Julia Kristeva in the mid-60s [14], i.e. the "shaping of a text's meaning by another text", has recently become an exploration playground for scholars in digital humanities worldwide due to the growing availability of large electronic (historical) corpora of various kinds in the field that has opened new possibilities for intertextual-based language data explorations. In parallel, there is a growing demand and an implied necessity to develop and integrate natural language processing (NLP) methodologies and techniques to novel applications in the area, such as intertextuality detection, since methods from NLP can shed light on the question of (historical) text reuse and contribute to the discovery of, e.g., the origin of quotes or allusions. This paper describes an experiment with a method used for biological sequence alignment that has also been applied in humanities research for e.g. the detection of similar passages in the complete works of Voltaire and 18th century French encyclopedias [5] and for tracing how and which ideas spread in 19th century US-newspaper collections [10,11]. For our experimental setup we use available software, text-PAIR [13] (Pairwise Alignment for Intertextual Relations) which we apply to automatically detect similar passages between the content of the Charles XII Bible translation into Swedish, completed in 1703, against the content of the Swedish prose fiction 1800-1900 database. After some parameter tuning and also some thorough, but manual, evaluation of the obtained results by experimenting with different configurations, the “best” results and parameters are discussed and examples are provided. The paper ends by a discussion of the limitations of the approach and particularly the current implementation of sequence alignment and gives some thoughts regarding new promising approaches [1] to overcome some of these limitations. Note, that while literary fiction often alludes to Biblical narratives and, depending on the text, may quote much from the Bible, sequence alignment only catches verbal correspondences but may neglect structural correspondences. The choice of using the genre of literary fiction and of this particular time period has simply been a pragmatic one, since we wanted to use available data at hand to test the technology.

2 Background

Multifaceted relations between texts can be (linguistically) complex, abstract, diverse or subtle. Researchers in e.g. digital humanities are particularly interested in identifying pairs of text passages likely to contain substantial overlap and empirically supporting (hopefully) new interpretations of historical texts. For instance, Cordell [3] discusses how digital interpretive tools can help make better sense of enlarged bibliographies, and the continuous digging into digital archives promises to effect exciting revisions to our literary history. Intertextual similarities between *historical* texts embrace a larger set of morphological, linguistic, syntactic, semantic and copying variations, thus adding a complication to text-reuse detection. Recycled text chunks are frequently only small portions of a document and may be significantly modified [5]. Intertextuality mining is inspired by plagiarism detection, and even more it has to take into consideration numerous alternations that may have transformed a text segment to a completely new counterpart [4]. Older language variants and dialects are less standardized; their evolution spanning centuries [1]; unlike today, e.g. verbatim quotes in older texts were not visually enclosed in quotation marks, making it hard for us to discern reuse from ‘original’ text; some authors quote other authors we know nothing about or whose works do not survive. Moreover, spelling and orthographic variations as well as OCR-errors can be problematic for the identification of (historical) text reuse [1,2]. Therefore, the task of detecting text re-use is challenging and requires a significant amount of experimentation and specialized expertise from several disciplines, with NLP having a major role to play in this process but also other areas of research could provide useful insights in the problem, e.g. computational philology and comparative phylogenetics [12,14]. NLP techniques to discover intertextual similarities between historical texts – i.e. how (ancient) authors copied, reused, borrowed, paraphrased and (even) translated earlier or contemporary authors as they spread their knowledge in writing, is a major topic of considerable interest among scholars from both a theoretical and practical point of view. Biological sequence alignment is one of available methods used for detection of similar passages in various literature, newspaper or other collections [2,5,8,10,11].

3 Material and Methods

3.1 The Charles XII Bible translation into Swedish and the Spf Corpus

For the experiment described in this paper we use as the sources of textual material, the Charles XII Bible translation into Swedish and the content of the Swedish prose fiction (Spf). We use the Charles XII Bible which is a translation completed in 1703 and remained the official Swedish Bible translation until 1917. An excerpt from the book is given below: *…ty bokstafven dödar, men Anden gör lefvande* i.e. "…for the letter kills, but the Spirit gives life". Spf is all prose fiction, written in the original and published separately for the first time, that appeared in Swedish during the years 1800, 1820, 1840, 1860, 1880 and 1900. Spf contains 300 publications, approximates 55 000 pages, written by 93 different authors. More information about Spf can be found here: <http://spf1800-1900.se/#!/om/inenglish>.

3.2 Sequence alignment for humanities research

A sequence alignment in bioinformatics is a way of arranging sequences of e.g. DNA in order to identify regions of similarity. Comparing known sequences with new sequences is one way of understanding the biology of an organism from which the new sequence comes. As implied, sequence alignments have been also used for comparing non-biological sequences, such as those present in natural language. In the context of humanistic research (string) sequence alignment is preferred to other techniques such as vector space similarity, since it has properties that are considered more appropriate for close readings of texts, not available on the latter (e.g., vector space approaches do not consider the notion of word order while the proposed similarity of such techniques is unrelated to text reuse).

The *Pairwise Alignment for Intertextual Relations* (PAIR or actually text-PAIR [13]) is a very simple implementation of a sequence alignment algorithm for humanities text analysis which supports one-against-many comparisons. That is, it is designed to identify "similar passages" shared by two strings or sequences (*aka*: longest common substring problem, that is to find the longest string that is a substring of two strings) in large collections of texts, such as direct quotations or other forms of borrowings including commonplace expressions. A corpus is indexed and an incoming text (the Bible in our case) is compared against the entire corpus for text reuse. text-Pair uses a small set of parameters (around 30) that can be easily tuned; for instance *stop words*; *start tags* (a reference to an array of tags that indicate the start of document for processing purposes; if no start tags are submitted, the entire document will be processed); *omit tags* (a reference to an array of hashes containing opening and closing tags that describe sections of text to be skipped during indexing); *normalize case* (default is 1, i.e. to flatten upper case characters to lower case when building shingles); *max doc percent* *matches* (a decimal 0 to 1, the fraction of shingles that can be shared between two documents before they are considered to be duplicates) etc.

4 Experiments and Parameter Tuning

text-PAIR can efficiently find similar text between a query document and large, pre-indexed corpus. The search can be configured to span minor differences between matched areas to increase the flexibility of matching. The technique used to index and query the corpus draws inspiration from work in genetic sequencing as well as other text similarity approaches. A pairwise comparison in this context implies that for each document in a collection a number of *n-grams* or *shingles* is generated and a number of parameters (such as *maximum gap* and *minimum span match*) are tuned in order to e.g. filter out common shingles that are separated by a large number of words in text order. The approach uses (overlapping) shingle sequences as the basic text unit and the evaluation of document pairs is performed by the intersection of n-gram features in these shingles which can be substrings or blocks of *n* words [9,10,11]. An example of a preprocessed shingle tri-gram generation for the fragment: *Och han grep drakan, den gamla ormen, som är djefvulen och Satan...* (i.e. 'And he seized the dragon, that ancient serpent, who is the devil and Satan…') will be: *grep_drakan_gamla*; *drakan_gamla_ormen*; *gamla_ormen_djefvulen*; *ormen_djefvulen_satan*; here function words *Och*, *han*, *den*, *som*, *är* and *och* (i.e. conjunctions, pronouns and the copula) have been removed.

The preprocessing included tokenization while the parameter tuning involved removing of numerals, the filtering of function words, the specification of the size of the content word n-grams, the minimum word length to consider, the size of shingle overlap between two sets, the span and the maximum number of consecutive gaps between shingles. Some of these parameters can increase flexibility; however by allowing more flexibility may also increase the ratio of uninteresting matches and redundancy. Nevertheless, if the criteria are met, the match is expanded, examining wider contexts in each document, until the criteria are violated, at which point the match is terminated and recorded [8]; as a matter of fact one of the longest passages text-pair could identify in our data is almost 80 tokens long (a segment provided here between angle brackets without translation: …*englars <tungor , och hade icke kärleken , så vore jag en ljudande malm , eller en klingande bjellra. Och om jag kunde profetera, och visste all hemlighet, och allt förstånd, och hade alla tro, så att jag försatte berg, och hade icke kärleken , så vore jag intet. Och om jag gåfve alla mina egodelar them fattigom, och låte min lekamen brinna, och hade icke kärleken, så vore> thet mig intet*). Therefore tuning and experimentation with various parameters is a necessary step, and dependent on whether we focus on high precision or high recall or a balance of the two.

The experiment compared the content of Spf with the Charles XII Bible by considering different tuning parameter settings. The most important of these parameters were the *shingle size* (i.e. the number of words to include in each shingle), the *minimum pair of shingles* (i.e. an integer indicating the minimum number of shingles that must be shared by both matches) and the *maximum gap* (i.e. an integer indicating the maximum number of continuous shingles in a match range that do not appear in the other match range; e.g. if set to 5, then any match range in a pair may contain at most 5 running shingles that do not appear in the other match range). The *minimum word length* of tokens to be indexed (we used 2), the *flattening of upper case characters* to lower case; the *removal of numerals* when building shingles and the use of a list of the 50 most frequent tokens in the corpus (i.e., the stop words) were some of the parameters that were kept intact in all experiments.

5 Results and Evaluation

Table 1 shows some figures based on the results obtained from three slightly different configurations A-C. A (shingle size 3; max gap 5; min pair of shingles 3); B (shingle size 3; max gap 5; min pair of shingles 2) and C (shingle size 3; max gap 12; min pair of shingles 3; and no stop words). Configuration A[1] showed the most conservative results with the highest precision in a qualitative style evaluation we performed (i.e., a close reading of the results). Out of the 300 volumes in Spf, configuration A could detect Bible reuses in 53 volumes, configuration B 139 and configuration C 183.

A [53]	B [139]	C [183]
25 nr of pairs: 1	56 nr of pairs: 1	42 nr of pairs: 1
13 nr of pairs: 2	26 nr of pairs: 2	36 nr of pairs: 2
5 nr of pairs: 3	18 nr of pairs: 3	22 nr of pairs: 4
3 nr of pairs: 4	10 nr of pairs: 4	17 nr of pairs: 3
3 nr of pairs: 5	8 nr of pairs: 5	16 nr of pairs: 5
1 nr of pairs: 6	4 nr of pairs: 7	10 nr of pairs: 6
1 nr of pairs: 7	3 nr of pairs: 6,8,13	7 nr of pairs: 7
1 nr of pairs: 11	2 nr of pairs: 10	6 nr of pairs: 12
1 nr of pairs: 12	1 nr of pairs: 9,11, 14,16,17,25	5 nr of pairs: 8
		4 nr of pairs: 11
		3 nr of pairs: 9,14
		2 nr of pairs: 10, 15, 16,18
		1 nr of pairs: 19, 20,27,31

Table 1. Results from three different configurations (A,B,C).

Configurations B[2] and C could caption more re-used segments but with the cost of also capturing well formulated but trivial associations such as the ones marked with angle brackets in the following examples: (i) *Hon <reste sig upp, och gick in> till grefvinnan* ‘She stood up, and went in to the countess’ (Spf-file lb3040528) and …*<reste han sig upp, och gick in> i staden* ‘…he stood up, and went into the city’ (in the Bible); and (ii) *<Tre dagar och tre nätter>*

[1] <http://demo.spraakdata.gu.se/svedk/seqAlignResults/spf-conf-A.html>.

[2] <http://demo.spraakdata.gu.se/svedk/seqAlignResults/spf-conf-B.html>.

satt Lilia… 'Three days and three nights sat Lilia…' (Spf-file lb99904107) and *…vara i <tre dagar och i tre nätter> i jordene* '…be three days and three nights in the earth' (in the Bible). The results from text-PAIR suggest that thorough (manual) evaluation is a very important step since no scoring mechanism is built into the text-PAIR that could rank the results according to some strength association measure, either statistical (e.g. log likelihood or perplexity) or distance based (e.g. containment or weighted resemblance)[3].

A	B	C
Ringen med svarta stenen; 1860; Wilhelm Aurell (12 pairs)	*Marfa d'Orino*; 1900; Carl Oscar Berg (25 pairs)	*En pilgrims resa genom denna jordens öken*; 1900; Nils Johnsson (31 pairs)
Marfa d'Orino; 1900; Carl Oscar Berg (11 pairs)	*En pilgrims resa genom denna jordens öken*; 1900; Nils Johnsson (17 pairs)	*En idyll på berget Karmel*; 1880; Janne Bruzelius (27 pairs)
En pilgrims resa genom denna jordens öken; 1900; Nils Johnsson (7 pairs)	*Ringen med svarta stenen*; 1860; Wilhelm Aurell (16 pairs)	*Den starkare;* 1900; Mikael Lybeck (20 pairs)
Mannen i röfvarehänder; 1880; C.O. Berg (6 pairs)	*Mannen i röfvarehänder*; 1880; C.O. Berg (14 pairs)	*Mannen i röfvarehänder*; 1880; C.O. Berg (19 pairs)

Table 2. The top-4 novels in the Spf-collection with most common pairs with the Charles XII Bible according to the three configurations previously outlined.

6 Conclusions and Future Work

In this paper we have briefly described a text reuse experiment by comparing the Charles XII Bible with the content of Swedish prose fiction collection. The experiment uses available software which implements sequence alignment and incorporates a number of parameters that need to be tuned and thoroughly tested in order to achieve results that are suitable for close reading of text re-use, by fulfilling a reasonable tradeoff between accuracy and efficiency. Since the results were only evaluated by hand, and by adjusting some of the implemented parameters, until the results were "pleasing" this runs the risk of optimizing precision at the cost of recall. Also, some sort of (shallow) syntactic parsing of the data (or, at least, chunking) might improve the results. However, more advanced implementations and methodologies might be necessary to cope with language phenomena beyond local alignment techniques [10], taking into consideration more linguistically informed

[3] Text re-use is a broad and potentially vague area of research, and gold standards are scarce or not at all available (at least not for Swedish) in this area. Nevertheless, there are thinkable ways to perform a comprehensive evaluation. For instance, in a comparable problem, in Information Retrieval, researchers have developed an evaluation method in terms of 11-points interpolated average precision [7:159] that can provide insight. We intend to apply it in the near future in order to get a "big picture" view of the results' performance.

preprocessing such as lemmatization, spelling variation, synonym identification, better forms of document representations and also methods more robust to noisy texts. For these reasons we have just started experimenting with more advanced software, TRACER [1], which implements several such extensions, by using a feature-based linking and scoring approach for the shingles acquired from texts. TRACER is developed within the project eTRACES[4], while its visualization component TRAViz (Text Re-use Alignment Visualization) [6] can be applied as a distant reading method in order to obtain an overview of the distribution of text reuses between each pair of texts and all texts in a corpus. Moreover, for future work we would also like to compare the content of Spf and also other Swedish digital collections (such as the Swedish literature bank[5]) with each other and also with classical, influential Swedish authors such as August Strindberg and Selma Lagerlöf.

Acknowledgements

This work is partially supported by the Swedish Research Council's framework grant (2012-2016; dnr 2012-5738): "Towards a knowledge-based culturomics" (<http://spraakbanken.gu.se/eng/culturomics>).

References

[1] Büchler M., Burns P. R., Müller M., Franzini E. and Franzini G. (2014). Towards a Historical Text Re-use Detection. In *Text Mining, Theory and Applications of NLP*. Biemann, C. and Mehler, A. (eds). Pp. 221-238. Springer.

[2] Colavizza G., Infelise M. and Kaplan F. (2015). Mapping the Early Modern News Flow: An Enquiry by Robust Text Reuse Detection. In Social Informatics. *Lecture Notes in Computer Science.* Vol. 8852. pp 244-253. Springer.

[3] Cordell R. (2013). "Taken Possession of": The Reprinting and Reauthorship of Hawthorne's "Celestial Railroad" in the Antebellum Religious Press. *Digital Humanities Quarterly 7:1*. Alliance of Digital Humanities Organizations (ADHO).

[4] Ganascia J-G., Glaudes P. and del Lungo A. (2014). Automatic Detection of Reuses and Citations in Literary Texts. *Lit Linguist Computing* 29 (3): 412-421. doi: 10.1093/llc/fqu020.

[5] Horton R., Olsen M and Roe G. (2010). Something Borrowed: Sequence Alignment and the Identification of Similar Passages in Large Text Collections. *Digital Studies / Le champ numérique*. Vol 2:1.

[6] Jänicke S., Geßner A., Franzini G., Terras M., Mahony S. and Scheuermann G. (2015). TRAViz: A Visualization for Variant Graphs. *Digital Scholarship in the Humanities – Digital Humanities 2014 Special Issue*.

[7] Manning C.D., Raghavan P. and Schütze H. 2009. *An Introduction to Information Retrieval*. CUP Cambridge, UK <http://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf>

[8] Roe G.H. (2012). Intertextuality and Influence in the Age of Enlightenment: Sequence Alignment Applications for Humanities Research. *Digital Hum.* 2012. Hamburg, Germany.

[4] <http://etraces.e-humanities.net/>

[5] <http://litteraturbanken.se/#!/om/inenglish>.

[9] Seo J and Croft W.B. (2008). Local Text Reuse Detection. *ACM SIGIR*, Singapore.

[10] Smith D.A., Cordell R. and Maddock Dillon E. (2013). Infectious Texts: Modelling Text Reuse in Nineteenth-Century Newspapers. *IEEE Conf on Big Data*. Santa Clara, CA, USA.

[11] Smith D.A., Cordell R., Maddock Dillon E., Stramp N. and Wilkerson J. (2014). Detecting and Modelling Local text Reuse. *14th ACM/IEEE-CS Joint Conference on Digital Libraries Pages (JCDL)*. Pp. 183-192. ACM.

[12] Swofford D.L. 2001. *PAUP*: Phylogenetic Analysis Using Parsimony (and other methods) 4.0.b5*

[13] text-pair. (2015). *PAIR: Pairwise Alignment for Intertextual Relations*. Online; accessed 31-August 2015: <https://code.google.com/p/text-pair/>

[14] Wikipedia. (2015). *Intertextuality*. Online; accessed 25-August-2015: <https://en.wikipedia.org/ wiki/Intertextuality>[14] Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer Applications in BioSciences* 13:555-556.

A corpus-based analysis of near-synonyms of fear in Russian

Erica Pinelli
University of Pavia
e-mail: erica.pinelli@unipv.it

Abstract

In this paper I investigate the distribution of four Russian near-synonyms of fear, *strach*, *ispug*, *opasenie* and *bojazn'* in prepositional phrases. The data are taken from the Russian National Corpus and they belong to the sub-corpus of 20th century. The quantitative analysis points out a clear opposition between two groups of nouns, *strach* and *ispug* on the one hand, and, on the other hand, *opasenie* and *bojazn'*. The qualitative analysis shows that the discrepancies in distribution are explained by the tendency of nouns and prepositions to express specific semantic roles.

1 Introduction

The aim of this paper is to analyze the semantics of four near-synonyms expressing fear in Russian *strach* 'fear', *ispug* 'fright', *opasenie* 'apprehension, fear' and *bojazn*' 'fear, dread'. In particular, I analyze the distribution of these four nouns of fear in prepositional phrases.

Prepositions, as any other linguistic elements, have their semantics, which must be compatible with the semantics of the nouns with which they co-occur (Mostovaja [7], Zelinsky-Wibbelt [14]). Langacker [6] observes that prepositions are extremely important because they convey relational meaning. Prepositions are able to express the relation between a Trajector, i.e. the entity located and described by the predication, and a Landmark, the entity which functions as the reference point for the predication. Because they express a relation, prepositions can tell us something about the entities involved. In the present analysis the role of prepositions is particularly important for determining the semantics of near-synonyms of fear and their conceptualization.

Cognitive researchers have studied the semantics of spatial prepositions drawing the attention to the relevant role of space in the conceptualization of more abstract concepts (Tyler & Evans [11], Vandeloise [13], Herskovits [4], Zelinsky-Wibbelt [14], Borozdina [1]). Croft [2] elaborates the Causal Order Hypothesis, and identifies the metaphors which are responsible for the conceptualization of causation through spatial metaphors. According to Croft [2], those metaphors are able to map the direction of movement onto the direction of causation. In the process of identification of metaphors prepositions play a relevant role.

Our experience of the world tells us that emotions and causality are strictly linked because emotions can be triggered by external or internal factors or can be themselves the origin of internal or external reactions. Radden [10] observes that emotional causality, as causality in general, is not mainly expressed by causal prepositions such as *because of* or *due to* but it is better expressed through spatial metaphors involving prepositions such as *in*, *with*, *for* and *out of*. Radden [10] also claims that the choice of a preposition is not accidental: the causal meaning of a preposition is motivated by its spatial meaning.

The present analysis focuses on the most relevant prepositional phrases and their semantic compatibility with fear words (see section 3). For this paper I limit the qualitative analysis to causal prepositional phrases and what they can tell us about the semantics of fear nouns. The qualitative analysis allows identifying the semantic roles played by fear nouns and prepositions which let us identify their most central semantic features.

2 The constructional profile and the analysis of near-synonyms

The quantitative analysis has been carried out following the approach of the *constructional profile* suggested by Janda and Solovyev [5]. This method has been tested for the investigation of near synonymy. Considering the construction as a “conventionalized form-meaning pair”, Janda and Solovyev [5] define the *constructional profile* of a word as “the relative frequency distribution of constructions that a certain word appears in”. The main implication is that the difference in the constructional profile reflects the difference in meaning: the constructional profile of two near synonyms will be more similar than the constructional profile of two distant synonyms. In order to test the validity of the constructional profile method, Janda and Solovyev consider the construction [(preposition)[NOUN]case] (the declined noun with or without preposition) and analyze the distribution of non related words in eight constructions. The results show that similar words, for example abstract nouns, behave similarly, while they behave completely differently from other kinds of words, such as concrete object nouns.

Janda and Solovyev [5] also analyze the constructional profile of near synonyms of sadness and happiness in Russian and they observe that these strictly related words occur in the same constructions but with different frequency. The authors demonstrate that the analysis of the constructional profile is a valid method to define the semantic similarity or dissimilarity of near-synonyms.

2.1 The semantics of Russian near-synonyms of fear

The four near-synonyms under analysis in this paper are *strach* 'fear', *ispug* 'fright', *opasenie* 'fear, apprehension', and *bojazn’* 'dread, fear'.

Strach is the most frequent noun of fear in Russian and it has high distribution. *Strach* refers to a general kind of fear and it has very few semantic constraints which allow it to occur in several different constructions. Uryson [12] underlines that *strach*, in some cases, can be controlled by the experiencer while, in other cases, it can lead the experiencer to lose control over his/her actions. In the present analysis, these semantic components of *strach* are confirmed. However, *strach* tends to express prototypically the semantic role of cause (see section 3).

Uryson[12] defines *ispug* as a sudden and short-lasting emotion that cannot strongly affect the experiencer's actions but can only affect facial expressions. However, the present analysis reveals that, in most of the occurrences, *ispug* can heavily affects the experiencer's reactions, although it has more semantic constraints in comparison to *strach*.

Opasenie can be defined as a feeling of anxiety due to a premonition of danger. This suggests that *opasenie* is a future oriented emotion; the analysis shows that *opasenie* can lead the experiencer to a conscious and intentional reaction that has rational motivation.

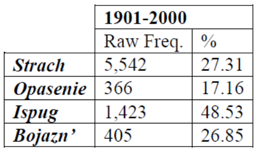
Uryson [12] shows that, on the one hand, *bojazn'* denotes an emotion that arises when the experiencer wants to avoid a perceived dangerous situation. Uryson underlines that in this case *bojazn'* is close to *opasenie* and this is confirmed by the present analysis. On the other hand, *bojazn'* can refer to an irrational fear in particular when the danger is not real, as in *bojazn' temnoty* 'fear of the dark'. The present analysis shows that *bojazn'* can condition the behavior of the experiencer who, after a reasoning process, acts wittingly.

2.2 The distribution of Russian words of fear in prepositional phrases

The present analysis has been carried out on a sub-corpus of data belonging to the 20th century taken from the Russian National Corpus (RNC). The data have been selected by using the RNC interface which allows to select those documents created in the lapse of time between 1901 and 2000.

For the present research I have taken into consideration the occurrences of the four Russian near-synonyms of fear (section 2) in prepositional phrases. In particular, I consider the occurrences in which the preposition occurs in position -1 with respect to the noun.

Table 1: Fear nouns in prepositional phrases. The percentage is calculated on the total number of occurrences of each fear noun in the corpus of the 20th century.



	1901-2000	
	Raw Freq.	%
Strach	5,542	27.31
Opasenie	366	17.16
Ispug	1,423	48.53
Bojazn'	405	26.85

Starting from these data (Table 1), two different discriminating factors have been taken into account to determine whether a prepositional phrase was relevant or not: the first is a quantitative factor, i.e. the frequency of co-occurrence with fear words, while the second factor is a qualitative judgment on the relevance of a prepositional phrase in the conceptualization of fear. The prepositional phrases I have taken into account are the following: *ot*+genitive 'from', *s*+genitive 'of(f)', *iz*+genitive 'out of', *s*+instrumental 'with', *v*+locative 'in', *na*+locative 'on', *v*+accusative 'into', *na*+accusative 'onto' and *pod*+instrumental 'under'. These prepositional phrases have been used to define the constructional profile of fear nouns (Figure 1).

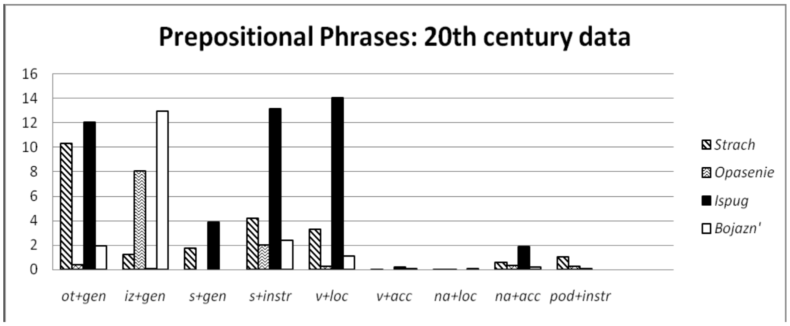


Figure 1: Relative Frequency of fear nouns in the most relevant prepositional phrases. Data of the 20th century from the Russian National Corpus.

In Figure 1, near-synonyms of fear show some preferences for specific prepositional phrases. It can be noticed that *strach* and *ispug* show high frequency in *ot*+genitive 'from' and in *v*+locative 'in'. Moreover, in *s*+genitive 'of(f)' only *strach* and *ispug* can occur. On the contrary, *opasenie* and *bojazn*' occur frequently in *iz*+genitive 'out of'. In the prepositional phrase *s*+instrumental 'with' all four nouns of fear show high frequency.

The hierarchical cluster analysis (Figure 2) obtained by the same set of data clearly points out the opposition of the two groups of nouns, *strach*/*ispug* and *opasenie*/*bojazn'*.

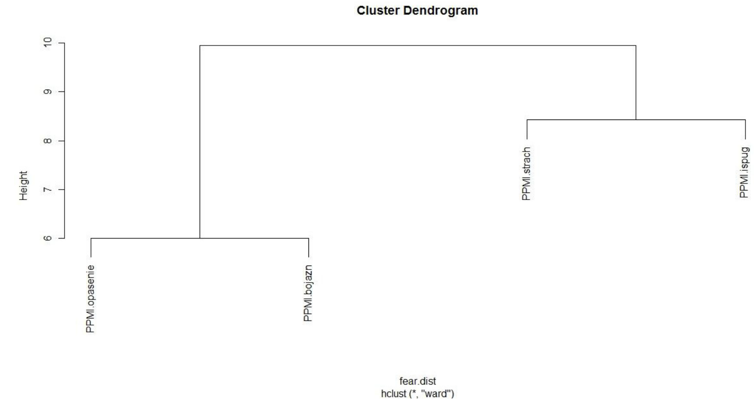


Figure 2: Cluster analysis

Figure 2 presents two clusters which divide *strach* and *ispug* in opposition to *bojazn'* and *opasenie*. It can be also noticed that the similarity between *opasenie* and *bojazn'* is prominently higher than the similarity between *strach* and *ispug*: *opasenie* and *bojazn'* clusterize at distance 6.00, while *ispug* and *strach* clusterize later at 8.42.

The correspondence analysis in Figure 3 shows the attraction between nouns of fear and prepositional phrases.

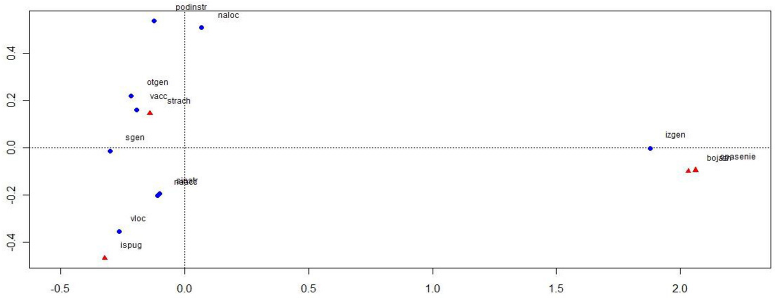


Figure 3: Correspondence analysis.

In Figure 3, *opasenie* and *bojazn'* are located close to each other at the very right of the two dimensional space and they attract the causal prepositional phrase *iz*+genitive. *Strach* and *ispug* are located at the very left but they are clearly separated. *Strach* attracts several prepositional phrases, in particular *ot*+genitive, *v*+accusative and *s*+genitive. *Ispug*, which shares several characteristics with *strach*, attracts *v*+locative.

3 The qualitative analysis of causal prepositional phrases

The quantitative analysis has highlighted the central role played by the three causal prepositional phrases *ot*+genititve, *iz*+genitive and *s*+genitive: they are particularly frequent and determine the opposition between *strach*/*ispug* and *opasenie*/*bojazn'*. In order to understand the reasons of this clear-cut opposition, I have analyzed qualitatively the occurrences of nouns of fear in causal prepositional phrases. The qualitative analysis presented in sections 3.1, 3.2 and 3.3 shows that both prepositions and nouns are more likely to express specific semantic roles, i.e. the role they play in the event.

In particular, there are two semantic roles relevant for the present discussion: cause and reason. Cause is defined by Croft [2] as “an event that causally immediately precedes the event expressed by the main verb”. In particular, the semantic role of cause suggests that the relation between the two events is direct and not mediated by any other means. Considering the specific case of fear, the emotion plays the semantic role of cause when it directly causes some reactions which cannot be controlled by the experiencer. On the contrary, the semantic role of reason implies the presence of an agent who is pushed to act in a certain way. In this case, the experiencer reacts intentionally to the emotion.

As explained in section 1, spatial prepositions can be used as causal prepositions through metaphorical extension. Moreover, the spatial meaning of the preposition can also determine its causal meaning.

3.1 Ot+genitive

The prepositional phrase *ot*+genitive 'from' suggests that the origin of the movement is a general point (or a human) without giving further physical details. Borozdina [1] argues that in the prepositional phrase 'X *ot* Y' there are two main relevant features: proximity and influence. The preposition *ot* implies that the Trajector (X) is removed from its original location, i.e. from the area surrounding the Landmark (Y); at the same time, Y exerts influence over X. These physical properties and the spatial relations between the Landmark and the Trajector are also active in the causal meaning of *ot*+genitive. The influence that the Landmark exerts over the Trajector has its metaphorical and causal counterpart in the reactions that the emotion triggers in the experiencer.

In the prepositional phrase *ot*+genitive *strach* and *ispug* are the most frequent nouns (12.07% and 10.33%), while *opasenie* and *bojazn'* show low frequencies (1.92% and 0.42%).

The analysis points out that *ot*+genitive very frequently co-occurs with verbs of physiological reaction such as *drozat'* 'tremble', *blednet'* 'turn pale', *cholodet'* 'turn cold'. This implies that the experiencer has no control over his/her reactions: we cannot control the tremble or the pallor.

(1) *Ruka eë* ***zadrožala ot stracha*** *i vino prolilos' na stol. [A. Chovanskaja. Avantjuristka (1928)]*
Her hand trembled from fear and the wine spilled on the table.

In (1) the verb *zadrožat'* 'start trembling' denotes a physiological reaction which cannot be controlled by the experiencer. In this case *ot stracha* expresses the semantic role of cause.

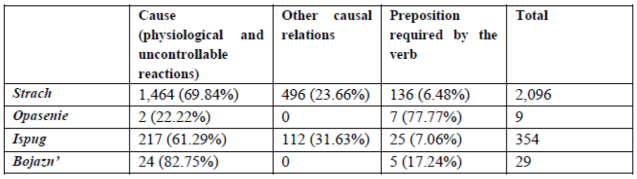
As I have already noticed, *opasenie* and *bojazn'* occur rarely in *ot*+genitive and they are not very likely to express the semantic role of cause. However, the analysis of the few occurrences of *ot opasenija* and *ot bojazni* shows that the preposition forces the noun to express an uncontrollable kind of emotion.

(2) *On snižalsja pri ètom nastol'ko, čto sotni naprjaženno sledivšik za nim glaz nevol'no rasširjalis'* ***ot opasenija*** *za sud'bu lëtčika. [S. Višenkov. Ispytateli (1947)]*
He descended so much that the hundreds of eyes which were intently following him involuntarily widened with [from] fear for the pilot's fate.

In example (2) *ot opasenija* expresses the semantic role of cause: the experiencer cannot control his physiological reaction, the dilation of pupils.

In Table 2, I present the results of the analysis of the occurrences of *ot*+genitive and I show that cause is the most frequent semantic sole.

Table 2:Analysis of the occurrences of ot+genitive.



	Cause (physiological and uncontrollable reactions)	**Other causal relations**	**Preposition required by the verb**	**Total**
Strach	1,464 (69.84%)	496 (23.66%)	136 (6.48%)	2,096
Opasenie	2 (22.22%)	0	7 (77.77%)	9
Ispug	217 (61.29%)	112 (31.63%)	25 (7.06%)	354
Bojazn'	24 (82.75%)	0	5 (17.24%)	29

The analysis shows that *ot*+genitive has a specific semantics which implies the loss of control of the experiencer. The two nouns of fear that mainly occur in this prepositional phrase, *strach* and *ispug*, are strictly linked to the expression of the semantic role of cause. The other two nouns, *opasenie* and *bojazn'*, are not likely to appear in this prepositional phrase which forces them to express the semantic role of cause.

3.2 Iz+genitive

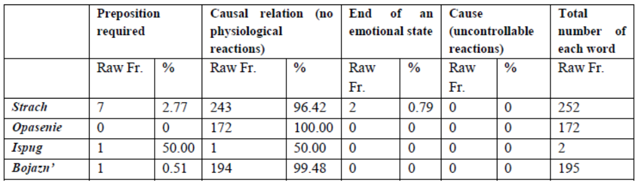
The prepositional phrase 'X *iz* Y' requires that the Landmark (Y) is a closed place or conceptualized as such. This suggests that the Trajector X, in order to be removed from the initial position, has to cross the boundaries of the Landmark Y. Besides the geometrical features of the Landmark, the preposition *iz* also focuses on the function of the Landmark as a container.

The analysis of the occurrences shows that *iz*+genitive never occurs with verbs of physiological and uncontrollable reactions. In this prepositional phrase *opasenie* and *bojazn'* are the most frequent nouns.

(3) *Ognja ne razvodili* ***iz opasenija****, čto kto-libo obnaružit ich mestoprebyvanie.* *[M. A. Šolochov. Tichij Don. Kniga četvërtaja (1928-1940)]*
They did not make a fire for [out of] fear that someone could discover their dwelling place.

In (3) the experiencer has full control over his/her body and s/he becomes an agent who intentionally reacts to the emotion: *iz opasenija* plays the semantic role of reason.

Table 3: Analysis of the occurrences of iz+genitive.



	Preposition required		**Causal relation (no physiological reactions)**		**End of an emotional state**		**Cause (uncontrollable reactions)**		**Total number of each word**
	Raw Fr.	%	Raw Fr.	%	Raw Fr.	%	Raw Fr.	%	Raw Fr.
Strach	7	2.77	243	96.42	2	0.79	0	0	252
Opasenie	0	0	172	100.00	0	0	0	0	172
Ispug	1	50.00	1	50.00	0	0	0	0	2
Bojazn'	1	0.51	194	99.48	0	0	0	0	195

Table 3 shows that reason is the only semantic role which can be expressed by *iz*+genitive. In this prepositional phrase also *strach* and *ispug*, which show low frequency, are forced to express the semantic role of reason.

(4) ***Iz stracha** on **ne berët** daže vzjatok. [V. M. Doroševič. Sachalin (Katorga) (1903)]*
Out of fear he doesn't even take bribes.

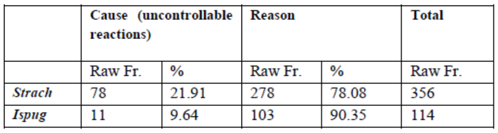
In (4) the experiencer does not lose control over his actions but he reacts intentionally. In *iz*+genitive, the noun *strach*, which more frequently occurs in *ot*+genitive, plays the semantic role of reason.

The spatial meaning of *iz*+genitive is metaphorically mapped onto the causal domain with some specific consequences: the Trajector's effort to cross the boundaries of the container (Landmark) metaphorically corresponds to the reasoning process of the experiencer who, released from the constraints of the emotion, reacts intentionally. This shows that *opasenie* and *bojazn'* denote a kind of fear which can play the semantic role of reason.

3.3 S+genitive

The prepositional phrase *s*+genitive denotes the origin from a surface or an open space. The spatial meaning of *s*+genitive shares some features with *ot*+genitive and some others with *iz*+genitive. The Landmark in *s*+genitive is not a container, but a surface which has limits that must be crossed by the Trajector in order to be removed from the initial position. The preposition *s* also shares with *iz* the support function: in *s*+genitive the surface functions as support for the Trajector. This intermediate position in the spatial meaning of *s*+genitive is reflected metaphorically onto the causal domain and confirmed by the analysis. Let us look at Table 4:

Table 4: Analysis of the occurrences of s+genitive.



	Cause (uncontrollable reactions)		**Reason**		**Total**
	Raw Fr.	%	Raw Fr.	%	Raw Fr.
Strach	78	21.91	278	78.08	356
Ispug	11	9.64	103	90.35	114

Table 4 shows that only *strach* and *ispug* can occur in this prepositional phrase but, contrary to *ot*+genitive, they mainly express the semantic role of reason.

(5) *Skaži mame, čto teper' mnogie moi znakomye sobirajutsja **so strachu** v Finlandiju, no ja tvërdo nameren vyvezti Vas v Petrograd. [K. I. Čukovskij. Pis'ma L. K. Čukovskoj (1912-1917) // "Družba narodov", 2001]*
Tell mum that many acquaintances of mine are now planning to go to Finland out of [off] fear, but I firmly intend to bring you out of here to Petrograd.

In example (5), the experiencer is also an agent who acts intentionally. Even if *strach* very often denotes a non-controllable fear, in this prepositional phrase it is more likely to play the semantic role of reason. Contrary to *iz*+genitive, *s*+genitive is compatible also with non-controllable reactions and, in few occurrences, *so stracha* and *s ispuga* play the semantic role of cause as in example (6).

(6) *U Trusivogo* ***so strachu zuby zastučali****.* *[P. V. Zasodimskij. Razryv-trava (1914)]*
Trusivyj's teeth were chattering from fear.

In (6) the experiencer has not complete control over his body reaction. However, several verbs of physiological reactions such as *belet'* 'turn white', *cholodet'* 'turn cold', *blednet'* 'turn pale' are not found in the data of *s*+genitive.

4 Conclusion

The analysis of four near-synonyms of fear in Russian, *strach*, *ispug*, *opasenie* and *bojazn'* in the most frequent and relevant prepositional phrases has pointed out that prepositions and nouns must be semantically compatible in order to co-occur. The constructional profile of the four near-synonyms has shown the high frequency of fear nouns in specific prepositional phrases; the cluster analysis has shown the clear-cut opposition between the two groups of nouns, *strach* and *ispug* on the one hand, and, on the other hand, *opasenie* and *bojazn'*. The correspondence analysis has reinforced the opposition of the two groups of nouns and has highlighted the relevance of causal prepositional phrases.

Starting from the quantitative results, I focused the attention on the occurrences of fear nouns in three causal prepositional phrases. The qualitative analysis showed that fear nouns tend to express a specific semantic role. In particular, *strach* and *ispug* are more likely to occur in *ot*+genitive and to express the semantic role of cause, highlighting the loss of control of the experiencer over his/her actions. On the contrary, *opasenie* and *bojazn'* mainly occur in *iz*+genitive and express the semantic role of reason. The analysis of *s*+genitive shows that this prepositional phrase can be considered to be in an intermediate position: it co-occurs only with *strach* and *ispug* but it mainly expresses the semantic role of reason.

The analysis of Russian fear nouns in prepositional phrases and their tendency to express specific semantic roles allowed to identify relevant semantic properties of both nouns and prepositions.

References

[1] Borozdina, I. S. (2003), *Semantika prostranstvennych predlogov (na materiale anglijskogo i russkogo jazykov),*

dissertacija na soiskanie učënoj stepeni kandidata filologičeskich nauk, Kursk.

[2] Croft, W. (1991), *Syntactic Categories and Grammatical Relations.* Chicago: University of Chicago Press.

[3] Dirven, R. (1997), *Emotions as cause and the cause of emotions,* in Niemeier, S. & Dirven, R., *The Language of Emotions*, pp. 55-83. Amsterdam/Philadelphia: John Benjamins.

[4] Herskovits, A. (1986), *Language and Spatial Cognition: an interdisciplinary study of the prepositions in English.* Cambridge: Cambridge University Press.

[5] Janda, L. & Solovyev V. (2009), *What constructional profiles reveal about synonymy: a case study of Russian words of sadness and happiness*, in 'Cognitive Linguistics', Vol. 20(2): 367-393.

[6] Langacker, R. W. (*2008*), *Cognitive Grammar: A Basic Introduction*. New York: Oxford University Press.

[7] Mostovaja, A. D. (1998), *On emotion that one can "immerse into", "fall into" and "come to": the semantics of a few Russian prepositional constructions*, in Athanasiadou A. & Tabakowska E., *Speaking of Emotions. Conceptualisation and Expression,* pp. 295-329. Berlin/New York: Mouton de Gruyter.

[8] Pajar, D. & Selivërstova O. N. (2000), *Issledovanija po semantike predlogov, sbornik statej*, Moskva: «Russkie slovari».

[9] Radden, G. (1985), *Spatial metaphors underlying prepositions of causality*, in Paprotté, W. & Dirven, R. (eds.), *The Ubiquity of Metaphor*, pp. 177-207. Amsterdan/Philadelphia: John Benjamins.

[10] Radden, G. (1998), *The conceptualisation of emotional causality by means of prepositional phrases*, in Athanasiadou, A. & Tabakowska, E., *Speaking of Emotions: Conceptualisation and expression*, pp. 273-294. Berlin: Mouton de Gruyter.

[11] Tyler, A. & Evans V. (2003), *The Semantics of English Prepositions.* Cambridge University Press.

[12] Uryson E. V. (2003), *Strach1, Bojazn', Ispug, Užas 1, Panika,* in Apresjan, Ju. D. (ed.), *Novyj Ob"jaznitel'nyj slovar' sinonimov russkogo jazyka*, pp. 1109-1115. Moskva: «Škola "jazyki i slavjanskoj kul'tury».

[13] Vandeloise, C. (1986), *L'espace en français. Sémantique des prépositions spatiales.* Paris: Édition du Seuil.

[14] Zelinsky-Wibbelt, C. (ed.) (1993), *The semantics of prepositions. From Mental Processing to Natural Language Processing*. Berlin/New York: Mouton de Gruyter.

Modal forms expressing probability and their combination with concessive sequences in French and Italian[1]

Corinne Rossari, Claudia Ricci and Margot Salsmann

Department of French linguistics
University of Neuchâtel
E-mail: corinne.rossari@unine.ch

Abstract

This paper deals with the combination between modal forms and concessive constructions in French and Italian, which we investigated in the framework of a tool-based approach to the linguistic analysis of corpora. To this end, we analyzed the frequency of forms expressing epistemic value and compared it to the frequency of their co-occurrence with the conjunction *mais*/*ma* (‘but’), seeking regularities in the link between this co-occurrence and the ability of an epistemic form to convey new meanings.

1 Introduction

This contribution examines the combination between some modal forms and concessive constructions in two Romance languages, French and Italian[2]. Our purpose is to identify regularities in the correlation between how frequently an epistemic form occurs in the environment of *mais/ma* (‘but’), and the ability of such a form to convey new meanings (e.g. a concessive value). To this end, we have analyzed the frequency of French forms expressing epistemic value in correlation with their co-occurrence with the conjunction *mais*. A parallel analysis has been carried out on the corresponding forms in Italian and their co-occurrence with the conjunction *ma*. Our analysis has been conducted in the framework of a tool-based approach to the linguistic analysis of corpora.

The research has been carried out from a synchronic perspective for French and Italian, with the aim of assessing the degree of convergence between two Romance languages, and from a diachronic perspective for French[3], with the aim of assessing a possible evolution concerning the co-occurrence between *mais* and the epistemic forms discussed here.

2 Analysis

[1] This research has been carried out in the framework of a research project entitled *La représentation du sens modal et de ses tendances évolutives dans deux langues romanes: le français et l’italien* (Representation of modal meaning and of the trends in its evolution in two Romance languages: French and Italian, project n. 100012_159458), directed by Corinne Rossari and financed by the Swiss National Science Foundation.

[2] For an overview of modal adverbs in French see Borillo [3].

[3] Our synchronic analysis covers both French and Italian corpora, while the diachronic analysis has been carried out on a French corpus only due to the unavailability of an equivalent corpus in Italian in our analysis tool.

The starting point of our research is the qualitative observation made in Rossari [9] and [11] that a number of French epistemic adverbs may take on a concessive value, which has become more or less conventionalized, while other adverbs of the same category do not acquire this kind of value[4]. We have observed that French adverbs *peut-être* ('maybe') and *sans doute* ('surely', lit. 'without (a) doubt') can either convey an epistemic value or be used by the speaker to cast doubt on the validity and/or relevance of a state of affairs, regardless of all evaluation of its degree of certainty. Let us look at the following case:

1) -Où est la directrice ?
-Elle est *peut-être/sans doute* dans son bureau, son secrétaire vient de lui apporter le café

-'Where is the director?'[5]
-'Maybe/Surely she is in her office, her secretary just brought her coffee'

In such a context, the above adverbs are interchangeable with *probablement* ('probably') in terms of acceptability of the utterance:

2) -Où est la directrice ?
-Elle est *probablement* dans son bureau, son secrétaire vient de lui apporter le café

-'Where is the director?'
-'She is probably in her office, her secretary just brought her coffee'

By contrast, in a context where the speaker resumes a content which is already present in the discursive background – for instance, because it has already been uttered by another speaker – with the intention of casting doubt on the validity and/or relevance of such a content, the adverb *probablement* is less natural, while *peut-être* and *sans doute* trigger a different interpretation of the utterance, less strongly rooted on an epistemic evaluation:

3) -Je trouve que la nouvelle directrice est sympa
-Elle est *peut-être/sans doute/*$^{?}$*probablement* sympa, *mais* elle est super exigeante

-'I think our new director is really cool'
-'Maybe/Surely/Probably she is cool, but she is extremely demanding'

Within this kind of context, the epistemic adverbs *peut-être* and *sans doute* seem to be taking on a concessive value which can go so far as to entirely rule out epistemic evaluation in the interpretation of the utterance. Such is the case for *peut-être*, which can be associated with contents whose degree of certainty cannot be questioned:

4) Je suis *peut-être* italienne, *mais* je n'aime pas la pizza

[4] The framework concerning the conventionalization of a value is related to the grammaticalization theory developed by Hopper and Traugott ([6]). In Traugott [13] the epistemic adverbs are analyzed in relation to their position in the sentence.

[5] Our translations are intended to render the basic meaning of the French and Italian examples and do not necessarily reflect the relevant values discussed here for French or Italian adverbs and tenses.

'I may be Italian but I don't like pizza'

The other epistemic adverbs introduced above are either less natural or even incompatible with this type of context:

5) Je suis $^{?}$*sans doute*/**probablement* italienne, *mais* je n'aime pas la pizza

'I am surely/probably Italian, but I don't like pizza'

Sans doute is less natural than *peut-être* in such a context, whereas *probablement* sounds particularly odd. However, in contexts where the state of affairs expressed is less self-evident, such as in the following example:

6) Le débat est *sans doute* français*, mais* il intéresse tout le monde (Mei Duanmu, Hugues Tertrais (eds.), *Temps croisés I*, Paris, Maison des Sciences de l'Homme, 2010)

'The debate is surely French, but it is of interest to everyone'

sans doute may occur (as could *peut-être*), whereas *probablement* in the same context would exclusively convey an epistemic reading ('it is not sure that the debate is French').

When used in the above way, *peut-être* expresses a concessive value which has become conventionalized. The adverb does not provide epistemic quantification concerning the states of affairs 'being cool' or 'being Italian', but rather takes in its scope the very act of enunciation, as if it were merely potential. In Rossari [10], such a use or value is labelled "enunciative" as opposed to a content/semantic use or value. While not sharing the same contexts of use as *peut-être*, *sans doute* appears to follow a similar trend.

The values these French adverbs convey are comparable to those of their Italian counterparts[6], as shown in the Italian rendering of the above examples:

7) -Dov'è la direttrice?
-*Forse/Probabilmente* è in ufficio, il segretario le ha appena portato il caffè.

-'Where is the director?'
-'Maybe/probably she is in her office, her secretary just brought her coffee'

Similarly to French, when used in a concessive context, the Italian adverbs show different degrees of acceptability. *Probabilmente* is less natural as it only conveys an epistemic meaning:

8) -Penso che la nuova direttrice sia proprio simpatica
-*Forse*/$^{?}$*Probabilmente* è simpatica, *ma* è molto esigente

-'I think our new director is really cool'
-'Maybe/probably she is cool, but she is extremely demanding'

As *peut-être*, *forse* can also be used in a context in which epistemic evaluation is completely ruled out, while we observe that *probabilmente* shows the same behavior as its French counterpart:

[6] Except for *sans doute*, which has no unique equivalent in Italian. To remain as close as possible to the French forms, we have restricted our analysis to the forms that convey very similar values in Italian. Therefore only *forse* (equivalent to *peut-être*) and *probabilmente* (equivalent to *probablement*) are analyzed here.

9) Sono *forse* Italiano per l'anagrafe perché ahimè sono nato lì, *ma* appena ho avuto l'opportunità, me ne sono andato il più lontano possibile. (www.australianboardcommunity.com)

'I may be Italian for the registry office because, alas, I was born there, but as soon as I had the opportunity I went as far away as possible'

10) Sono **probabilmente* Italiano per l'anagrafe perché ahimè sono nato lì, *ma* appena ho avuto l'opportunità, me ne sono andato il più lontano possibile.

'I am probably Italian for the registry office because, alas, I was born there, but as soon as I had the opportunity I went as far away as possible'

The same type of value conveyed by *peut-être* and *forse* can be taken on by the Italian future[7], which can also function at the level of enunciation, when it precedes *ma*:

11) *Sarò* italiana *ma* non mi piace la pizza

'I may be Italian but I don't like pizza'

Such a future could not be used in French:

12) *Je *serai* italienne, *mais* je n'aime pas la pizza

On the basis of these data, evaluated on a qualitative level, we make the assumption that there is an evolution pattern in the conventionalization of the concessive value of epistemic adverbs, or of a tense as the future in Italian. We claim that there is a correlation between this conventionalization process and the frequency of these forms in a context in which they precede *mais*/*ma*. The more frequently an epistemic form or the future occurs in such a context, the more likely such a form will be to lose its literal meaning and acquire a concessive value, thus allowing the speaker to cast doubt on the validity and/or relevance of a content already present in the discursive background.

Our purpose is: (i) from a synchronic perspective, to compare French and Italian in order to verify the convergence between two Romance languages as for the ratio: frequency of the correlation between *mais*/*ma* and an epistemic form (or the future) / conventionalization of a concessive value; (ii) from a diachronic perspective, to verify if such a frequency is stable or has evolved along the centuries. The latter analysis is focused on French data but we intend to extend it to Italian data.

3 French and Italian data

Our quantitative study[8] has been conducted on an annotated French corpus of over twenty million words. We have used the BTLC platform conceived by Sascha Diwersy (see Diwersy [5]), which allows, among other, extracting concordances, calculating frequencies and frequency-related specificities within the corpus, as well as extracting lexical-syntactic co-occurrences. This

[7] For a description of such a use of the Italian future see Berretta [1].

[8] See Legallois [7] for the theoretical background taken into account for our quantitative research.

platform includes synchronic corpora representing different genres of speech as well as a diachronic database gathering corpora from the 16th to the middle of the 20th century, calibrated with respect to the genres of speech (theatre, romance novels, essays) and selected in the framework of a French and German project on the diachronic evolution of French prepositions (PRESTO) (see http://presto.ens-lyon.fr/ for a description of the project). For our synchronic analysis we used a corpus of press speech from the 21th century (*Le Monde* 2008). This corpus includes different subgenres (such as editorials, books reviews), which have been treated separately to verify if the frequencies observed in the whole media press corpus are also relevant in the subgenres, avoiding, as far as possible, corpus biases. For our diachronic analysis, we have used the corpora selected in the frame of the Presto project, restraining them to the period 17th-19th century, as data for the 20th century are only available until 1941.

As regards Italian, we have used the same platform, which contains an annotated Italian press corpus of over thirty million words (*La Stampa* 2002), equivalent to the French corpus in terms of genres, subgenres and synchronic span. Since BTLC does not include a diachronic corpus for Italian, such a corpus will be further constituted in order to conduct a parallel analysis of the diachronic evolution of these forms in Italian.

4 Discussion of results for French

The results including the synchronic and diachronic data for the epistemic adverbs support our initial assumption concerning the correlation between their frequency in the environment of *mais* and their ability to convey or stress a concessive value.

Corpus *Le Monde 2008* (20 410 766 words)

Adverbs	**Total number of occur-rences**	**Relative frequency (occurrences per million words)**	**Number of occur-rences followed by *mais*[9]**	**Relative frequency (occurrences per million words)**	**Rate of co-occurrences with *mais* on all occurrences of the adverb**
Peut-être	2904	142.27	432	21.16	14.87 %
Sans doute	2483	121.65	323	15.82	13.00 %
Probablement	724	35.47	60	2.93	8.28 %

Figure 1: results for the association between epistemic adverbs and *mais* – synchronic perspective

These synchronic data bring out a significant fact: in the case of *probablement,* the rate of the adverb co-occurring with *mais* on its overall number of occurrences is lower by almost half than the same rate for *peut-être* and by 1.5 times than the same rate for *sans doute*.

[9] We have taken into account a text portion in which the adverb co-occurs with *mais* within a span of 0 to 20 words.

Corpus Presto – 17th-19th century
(17th: 7 040 233 words; 18th : 7 756 598 words; 19th : 8 020 248 words)

17th century	**Total number of occurrences**	**Relative frequency (occurrences per million words)**	**Number of occurrences followed by *mais***	**Relative frequency (occurrences per million words)**	**Rate of co-occurrences with *mais* on all occurrences of the adverb**
Peut-être	957	135.93	118	16.76	12.33 %
Sans doute	601	85.34	76	10.79	12.6 %
Probablement	21	2.98	0	–	0 %
18th century	**Total number of occurrences**	**Relative frequency (occurrences per million words)**	**Number of occurrences followed by *mais***	**Relative frequency (occurrences per million words)**	**Rate of co-occurrences with *mais* on all occurrences of the adverb**
Peut-être	2493	321.40	306	39.45	12.27 %
Sans doute	1411	181.79	278	35.81	19.7 %
Probablement	129	16.61	16	2.06	12.4 %
19th century	**Total number of occurrences**	**Relative frequency (occurrences per million words)**	**Number of occurrences followed by *mais***	**Relative frequency (occurrences per million words)**	**Rate of co-occurrences with *mais* on all occurrences of the adverb**
Peut-être	2774	345.87	360	44.88	12.9 %
Sans doute	2053	255.82	523	65.17	25.47 %
Probablement	302	45.09	22	2.74	6 %

Figure 2: results for the association between epistemic adverbs and *mais* – diachronic perspective

The diachronic data[10] show a continuous progression of the overall frequency of *peut-être*, coupled with a parallel progression of its frequency in co-occurrence with *mais*. The rate of occurrences of *peut-être* with *mais* with respect to the overall occurrences of the adverb is remarkably stable throughout three centuries. The same does not hold true for *sans doute*, whose use with *mais* progresses strongly compared to the overall progression of the adverb's frequency during the three centuries. On the contrary, while the frequency of *probablement* increases steadily through the three centuries considered, its co-occurrence with *mais* decreases by half from the 18th to the 19th century.

These data are consistent with the assumption that *peut-être* had the same ability to take on an enunciative value for the three centuries considered, whereas *sans doute* appears to develop this ability throughout the centuries.

[10] The data in the diachronic table cannot be compared to those in the synchronic one because each table results from a different corpus sample.

5 Discussion of results for Italian

The same synchronic quantitative analysis as for French has been conducted for the Italian epistemic adverbs corresponding to *peut-être* and *probablement*, namely *forse* and *probabilmente.* Our aim is to verify if, in Italian also, the value of the adverbs can be linked to their presence in a context containing *ma*.

Corpus *La Stampa 2002* (31 369 484 words)

Adverbs	**Total number of occurrences**	**Relative frequency (occurrences per million words)**	**Number of occurrences followed by *ma***	**Relative frequency (occurrences per million words)**	**Rate of co-occurrences with *ma* on all occurrences of the adverb**
Forse	11773	375.30	1494	47.62	12.69 %
Probabilmente	2771	88.33	273	8.70	9.85 %

Figure 3: results for the association between epistemic adverbs and *ma* – synchronic perspective

Although the gap between the ratios for the different adverbs is smaller, it is a significant fact that *probabilmente* shows a lower rate of occurrences with *ma*, which is comparable to the results obtained for its French counterpart *probablement.*

Since the Italian future can also convey a concessive value when it precedes *ma*, besides conducting a quantitative analysis on the epistemic adverbs, we have also analyzed the frequency of the future with *ma/mais* in Italian and in French. We aimed at verifying if such a use is correlated to a higher frequency of the co-occurrence 'future […] *ma*' in Italian as compared to the frequency of the co-occurrence 'future [...] *mais*' in French.

Corpus *Le Monde 2008* (20 410 766 words) **/ *La Stampa 2002*** (31 369 484 words)

Future	**Total number of occurrences**	**Relative frequency (occurrences per million words)**	**Number of occurrences with *ma/mais***	**Relative frequency (occurrences per million words)**	**Rate of occurrences with *ma/mais* on all occurrences of the future**
Italian future	143 063	4560.58	9902	315.66	6.9 %
French future	53139	2603.48	2754	134.92	5.1 %

Figure 4: results for the association between the future tense and *ma/mais* – synchronic perspective

Although the co-occurrence with *ma/mais* is lower in the case of the future than it is in the case of the adverbs, it is significant that its frequency is still higher in Italian than it is in French[11]. Consistently with the qualitative

[11] The difference between Italian and French does not appear to be so spectacular due to the fact that, within a span of 20 words, there are many cases in which the use of the future is

analysis, which shows that the future in French does not convey a concessive value, we observe that this tense is less likely to appear with *mais*.

6 General discussion

Besides the correlation between the frequency of occurrences of the above forms in the environment of *ma/mais* and their ability to convey a concessive value, the quantitative results show another peculiarity.

We can observe that there is a remarkable disproportion between the overall frequency of the future in French and that of the future in Italian. Our data show half the rate of occurrences of the future in French as compared to Italian. In a previous study (Ricci, Rossari and Siminiciuc [8]) we had observed that the non-temporal uses of the future (which we consider as resulting from an enunciative use of the tense and of which the concessive future is an instance) are much more frequent in Italian than they are in French.

We have verified this on a quantitative level by analyzing the rate of futures expressing probability[12] in relation to the overall frequency of the future. Below are the results of this research on the totality of our corpus:

Corpus *Le Monde 2008* (20 410 766 words) **/ *La Stampa 2002*** (31 369 484 words)

Future	**Total number of occurrences**	**Relative frequency (occurrences per million words)**	**Number of occurrences with an epistemic value**	**Relative frequency (occurrences per million words)**	**Rate of occurrences with an epistemic value on all occurrences of the future**
Italian future	143 063	4560.58	25 in a sample of 1500	76.00	1.66 %
French future	53139	2603.48	4 in a sample of 1500	6.94	0.26 %

Figure 5: results for the epistemic use of the future tense in the global French and Italian corpora

The data in our table confirm that the overall frequency of occurrences of the future is linked to the rate of its occurrences as a future of probability. The relative frequency of the future is 1.75 times higher in Italian than in French. Correspondingly, the rate of occurrences of the future of probability on all occurrences of the future is 6.4 times higher in Italian.

independent of the use of *ma/mais*. Within a span of 10 words, the gap between the rate of occurrences with *ma* in Italian as compared to French increases up to a ratio of 2:1.

[12] Such a future, which we also consider as being an enunciative use in Rossari, Ricci and Siminiciuc [12] has been studied, a.o., by Dendale [4] for French and is referred to in the *Grande grammatica italiana di consultazione* as "epistemic future" (see Bertinetto [2], 118-120). This type of future exists both in French and in Italian. Below is an example of its use:
Non telefonargli! A quest'ora *dormirà* / Ne l'appelle pas! À cette heure-ci, il *dormira*
'Don't call him. He'll be sleeping at this hour'

The same research has been conducted on a subsection of each corpus, respectively *Le Monde des Livres* and *Inserto tuttolibri* (comparable subsections featuring book reviews) in order to make sure that the difference in rates is not due to corpus bias. The results show a very similar disproportion:

Corpus *La Stampa 2002*, subsection *Inserto tuttolibri* (1 421 663 words)
Corpus *Le Monde 2008*, subsection *Le Monde des livres* (1 095 656 words)

Future	**Total number of occurrences**	**Relative frequency (occurrences per million words)**	**Number of occurrences as a future of probability**	**Relative frequency (occurrences per million words)**	**Rate of occurrences as a future of probability on all occurrences of the future**
Italian future	3384	2380.31	80	56.27	2.36 %
French future	2162	1973.24	18	16.42	0.83 %

Figure 6: results for the epistemic use of the future tense in a subsection of each corpus

The overall data make visible not only the correlation between the frequency of the adverb occurring in the environment of *mais* and its ability to convey a concessive value – a frequency which has been observed in synchrony, and diachrony for French – but also the link between the frequency of the future and its ability to have a non-temporal probability use.

7 Conclusion

Quantitative data have allowed: (i) for French, in a synchronic and in a diachronic perspective, to verify the influence of *mais* on the meaning of the adverbs, showing that the more frequently a marker appears in the left environment of *mais*, the more it will be prone to develop enunciative meanings consisting in stressing or taking on concessive values; (ii) to highlight the fact that enunciative meanings such as probability and concession for the future in Italian are correlated to the absolute frequency of the form.

As consequence of the latter observation, we could make another assumption which should be tested in a further analysis: the more frequent an epistemic form is, be it grammatical or lexical, the more it is prone to develop enunciative values.

References

[1] Berretta, Monica (1997) Sul futuro concessivo: riflessioni su un caso (dubbio) di de/grammaticalizzazione. *Linguistica e Filologia* 5, 7–41.

[2] Bertinetto, Pier Marco (1991) Il verbo. In Renzi, Lorenzo, Salvi, Gianpaolo (eds), *Grande grammatica italiana di consultazione*, vol II, pp. 13–161. Bologna: il Mulino.

[3] Borillo, Andrée (1976) Les adverbes et la modélisation de l'assertion. *Langue française* 30, 74–89.

[4] Dendale, Patrick (2001) Le futur conjectural versus *devoir* épistémique : différences de valeur et restrictions d'emploi. *Le Français Moderne* 69, 1, 1–20.

[5] Diwersy, Sascha (2014) La plateforme Primestat.BTLC et l'exploitation lexico-statistique de corpus diachroniques. Lecture given at the Journée d'études à l'Institut des Sciences du Langage et de la Communication, University of Neuchâtel.

[6] Hopper, Paul and Traugott, Elisabeth C. (1993) Grammaticalization. Cambridge, UK: Cambridge University Press.

[7] Legallois, Dominique (2012) La colligation : autre nom de la collocation grammaticale ou autre logique de la relation mutuelle entre syntaxe et sémantique ? *Corpus* [Online] 11, http://corpus.revues.org/2202.

[8] Ricci, Claudia, Rossari, Corinne and Siminiciuc, Elena (in press) La représentation des sens modaux dans trois langues romanes : le français, l'italien et le roumain. Du qualitatif au quantitatif et retour. *Syntaxe et sémantique* 17.

[9] Rossari, Corinne (2014) How does a concessive value emerge? In Ghezzi, Chiara and Molinelli, Piera (eds.) *Pragmatic markers from Latin to Romance languages*. Studies in diachronic and historical linguistics. Oxford: Oxford University Press.

[10] Rossari, Corinne (2015) Une concession implique-t-elle une opposition ? In Ferrari, Angela, Lala, Letizia and Stojmenova, Roska (eds.), *Testualità. Fondamenti, unità, relazioni*, pp. 189–203. Firenze: Franco Cesati.

[11] Rossari, Corinne (in press) La présupposition dans les structures concessives. In Biglari, Amir and Bonhomme, Marc (eds.) *La Présupposition entre théorisation et mise en discours*. Paris: Classiques Garnier.

[12] Rossari, Corinne, Ricci, Claudia and Siminiciuc, Elena (in press) Les valeurs du futur modal en français, italien et roumain. In Baranzini, Laura, Sanchez-Mendez, Juan Pedro and De Saussure, Louis (eds.) *Le futur dans les langues romanes*. Bern: Peter Lang.

[13] Traugott, Elizabeth C. (in press) On the function of the epistemic adverbs surely and no doubt at the left and right periphery of the clause. In Beeching, Kate and Detges, Ulrich (eds.), *The role of left and right periphery in semantic change*. Amsterdam: Benjamins.

Non-accusative null objects in the Homeric Dependency Treebank

Eleonora Sausa and Chiara Zanchi
University of Pavia and University of Pavia/University of Bergamo
eleonorasausa@gmail.com, zanchich2@gmail.com

Abstract

While accusative referential null objects in Ancient Greek are the subject of various descriptions, an investigation on non-accusative null objects is still a *desideratum*. In this paper we analyze two kinds of contexts, semi-automatically retrieved from the Homeric Dependency Treebank, in which one can assume that a genitive or a dative object is omitted. In particular, we focus on contexts with coordinated verbs and with conjunct participles. After a manual revision of the occurrences retrieved, we describe and analyze the data according to a number of parameters, in order to hypothesize the case of the non-overt object.

1 Introduction[1]

While accusative referential null objects in Ancient Greek are the subject of various studies (Haug [6], Keydana and Luraghi [7], Luraghi [8]), an investigation on the omission of non-accusative objects is not currently available. In this paper we analyze the occurrences, semi-automatically retrieved from the Homeric Treebank (HDT),[2] with coordinated verbs and with conjunct participles, in which one can hypothesize the omission of the second argument. We focus on contexts in which the expressed object is in the genitive or in the dative.

This paper is organized as follows. Section 2 is devoted to the theoretical background: after touching upon the issue of non-configurationality in Homeric Greek, we deal with the status of accusative and non-accusative objects in this language and summarize the syntactic and pragmatic conditions which allow or require object omission. Section 3 describes the methodology: after introducing the Homeric Treebank, we illustrate the query that extracts null objects of coordinated verbs and of conjunct participle. In section 4, we describe the contexts with coordinated verbs and with conjunct participles and analyze them according to a number of parameters. On the basis of such parameters, we try to hypothesize the case of the non-overt object. Finally, we draw our conclusions in section 5.

[1] We would like to express our gratitude, on the one hand, to Marco Passarotti and Edoardo Ponti for their support with the Tree query editor *TrEd*, and, on the other hand, to Silvia Luraghi and three anonymous reviewers for their insightful comments. Of course, all mistakes are ours.
[2] URL: http://nlp.perseus.tufts.edu/syntax/treebank/greek.html.

2 Theoretical background

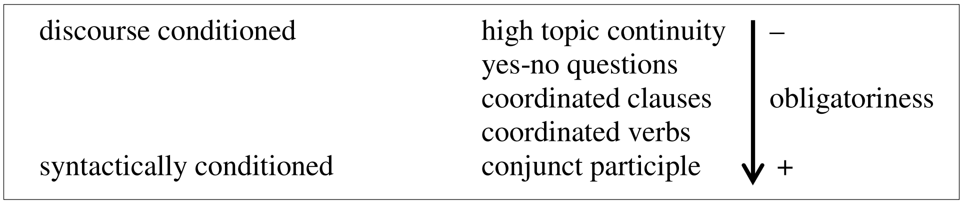
2.1 Non-configurationality in Homeric Greek

It has been argued that Homeric Greek is a non-configurational language, that is, a language in which constituency and government are at their onset. Features connected with non-configurationality are, among others, free word order, discontinuous constituents, and the occurrence of null anaphora (Baker [1], Devine and Stephens [3]).

Such features are shown by Homeric Greek: it displays free word order (in particular, the verb can virtually occur in any position within the sentence), an abundance of discontinuous constituents (that is, Noun Phrases or Prepositional Phrases interrupted by more or less heavy lexical items), and a frequent use of null anaphora for all verbal arguments. Notably, the possibility of omitting even referential direct objects reveals that, at this stage, the relation between the verb and its arguments is not yet ruled by government. In other words, the dependency relation between a verb and its arguments cannot be regarded as obligatory and fixed: the same verb can show case variation, and objects can be frequently omitted.[3]

2.2 Accusative null objects in Ancient Greek

Keydana and Luraghi [7] and Luraghi [8] argue that in Ancient Greek the omission of referential objects occurs under certain syntactic and pragmatic conditions (see also Haug [6]), as shown in figure 1. For reasons of space, we do not describe here those contexts for which the omission of the accusative referential null object is discourse conditioned and less obligatory, that is, cases of high topic continuity and yes-no questions, which have been not taken into account in our investigation.



discourse conditioned	high topic continuity	–
	yes-no questions	
	coordinated clauses	obligatoriness
	coordinated verbs	
syntactically conditioned	conjunct participle	+

Figure 1 Contexts of object omission and their degree of obligatoriness

[3] An anonymous reviewer wonders whether object omission could be simply due to metrical reasons. Remarkably, however, the omission of verbal arguments in Homeric Greek is consistent with omission as found in Classical prose writers. Moreover, omission of dative and genitive null objects is also consistent with that of accusative objects, which has been investigated in detail (Haug [6], Keydana and Luraghi [7], Luraghi [8]).

More syntactically conditioned contexts are those with coordinated clauses and coordinated verbs, with *kai* or other coordinative conjunctions, illustrated in (1):

(1) *allá pou* **autòn** *thumòs epotrúnei kaì anṓgei* ø
but certainly self:ACC heart:NOM move:PRS.3SG and force:PRS.3SG
‘Certainly his heart moves and forces **him.**’ (*Il.*15.43)

Example (1) shows what Harris-Delisle [5] calls ‘conjunction’ or ‘coordination reduction’: when two sentences or clauses are coordinated, a number of features present in the first sentence or clause – including the overt realization of an argument – can be omitted in the second one. In Ancient Greek, in such contexts, the omission of the direct object is a clear trend, but not a strict rule. Certain direct objects, having a special status for pragmatic reasons, can be repeated, as shown in the passage at (2). In example (2), the first person singular pronoun functioning as direct object is repeated:

(2) *hóte* ***m’*** *ṓnato kaí* ***m’***
when 1SG.ACC blame:AOR. MID.3SG and 1SG.ACC
hupémeine
face:AOR.3SG
‘When he blamed and faced me.’ (*Il.*17.25)

The passage in (3) shows the most obligatory (and the most frequent) context of object omission, that is, the presence of a conjunct participle. This can also be regarded as a case of object sharing:

(3) *Pátroklo mèn* ***sîton*** *helṑn*
P.:NOM PTC bread:ACC take:AOR.PTCP.NOM
epéneime ø *trapézēi kaloîs en kanéoisin*
distribute:AOR.3SG table:DAT.F fair:DAT.PL in basket:DAT.PL
‘(lit.) Patroclus, taking **the bread**, distributed (**it**) on the table in fair baskets.’ (*Il.*9.216)

2.3 Non-accusative objects in Homeric Greek

As a matter of fact, in Homeric Greek bivalent verbs can occur with non-accusative second arguments. Adopting a constructionist approach to verb argument structure (see, among others, Barðdal [2], Goldberg [4]), one can state that in Homeric Greek verbs with two arguments can occur in one of the three Nominative-Subject subconstructions: Nom(inative)Acc(usative), Nom(inative)Gen(itive), and Nom(inative)Dat(ive). These three constructions are characterized by their frequencies of occurrence and by the semantics of verbs which instantiate them, as argued in Sausa [9]. For example, verbs typically instantiating the NomDat construction are verbs of ruling, helping, fighting, experiential predicates denoting anger and verbs of motion with animate Goal. Verbs typically instantiating the NomGen construction are verbs of departing from a source, verbs of missing and ceasing, verbs of emotion denoting love, desire and caring, verbs of mental activity and of perception.

Thus, these three Nominative-Subject subconstructions differ in the case assigned to the object. Therefore, one can talk about accusative objects, dative objects, and genitive objects. If the case of the omitted object can be assumed as being different from the accusative, it is possible to talk about a non-accusative null object.

3 Methodology

3.1 The Homeric Dependency Treebank

The HDT, which is included in the Ancient Greek Dependency Treebank (AGDT),[4] collects all sentences of Homeric poems, syntactically annotated and represented as tree structures. The HDT can be queried by means of the tree editor *TrEd.*[5]

3.2 Query for retrieving non-accusative null objects

As already highlighted, the most obligatory contexts of object omission, that is, those of coordination reduction and of conjunct participles, are the easiest to extract, though AGDT syntactic annotation does not provide dedicated nodes for non-overt arguments.

For coordination reduction, we have bypassed this issue by extracting all COORD(ination) nodes, which are parents both of two coordinated verbs (m/tag = “^v”, is_member = 1) and of a non-accusative object (afun = “OBJ”, m/tag != “-------a-”). The label COORD (afun = “COORD”) comprises coordinative devices of different sort, such as correlative particles (e.g. *mén…dé…*) and coordinative conjunctions (e.g. *kai*). Then, we have manually eliminated all extracted occurrences in which one of the verbs is intransitive or does not require a definite referential object.

For conjunct participles, we have extracted all finite verbs (“^v”) that have as children nodes a non-finite verb (“^t”) and a non-accusative object, or a non-

[4] URL: http://nlp.perseus.tufts.edu/syntax/treebank/greek.html. Guidelines at http://nlp.perseus.tufts.edu/syntax/treebank/greekguidelines.pdf.
[5] URL: https://ufal.mff.cuni.cz/tred/documentation/tred.html.

finite verb which in turn has a non-accusative object as child node. We have further added the condition that the (non-)finite verb must not take a non-accusative object. Finally, we have put together and compared all these extracted data, and excluded contexts in which one of the verbs is intransitive or does not require a referential null object.

4 Analysis

4.1 Parameters of analysis

We analyze contexts with coordinated verbs and with conjunct participles with a genitive or dative overt object according to the following parameters:

(a) the type of coordinative conjunction (in contexts of coordinated verbs);
(b) the position of the overt object;
(c) the semantic roles played by the overt and null objects;
(d) the meaning of the verbs;
(e) the constructions which the verbs tend to instantiate.

All these parameters provide us with indications about the case of the non-overt object. As already mentioned, semantic similarity among verbs is crucial for our analysis, as it is regarded as an element in favor of the instantiation of the same construction: similar verbs, indeed, typically tend to occur in similar constructions.

4.2. Coordinated verbs

4.2.1. Dative null objects with coordinated verbs

We now discuss the contexts of null object with coordinated verbs with a dative overt object: we collected 23 Homeric passages of this kind.

In (4) and in similar examples, two verbs are coordinated with correlative particles (here *d'…dè…*) or coordinative conjunctions, the dative overt object occurs in preverbal position, and the two coordinated verbs are near-synonyms. Both the overt and the null objects play the semantic role of Stimulus and the null object can be assumed as a dative object.

(4) ***Hḗrēi*** *d'* *oú* *ti* *tóson*
H.:DAT.F PTC NEG PTC so.much
nemesízomai *ou* *dè* *kholoûmai* ø
feel.indignation:PRS.MID.1SG NEG PTC be.angry:PRS.MID.1SG
'I am not so indignant **at Hera**, I am not angry **(at her).**'
(*Il.*8.407=*Il.*8.421)

The overt object may also occur after both verbs, as in (5):

(5)	*ken*	*égōg'*	*ethéloimi*	*parestámenai*	ø
	PTC	1SG.NOM	want:PRS.OPT.1SG	stand.by:INF.PF	
	kaì	*amúnein*	***Patróklōi***		
	and	defend:INF.PRS	P.:DAT		

‘So I would stand by **Patroclus** and protect **(him).**’ (*Il.*17.563)

In (5) the two infinitives are coordinated by *kai,* and are near-synonyms. Both the overt object and the null one play the semantic role of Theme. The null object can be assumed as a dative object.[6]

4.2.2. Genitive null objects with coordinated verbs

Genitive null objects with coordinated verbs in the Homeric poems are six. In examples below two verbs are coordinated with *kai* or other coordinative conjunctions and the overt object in genitive case precedes both verbs.

In example (6) the null object can be assumed as a genitive object: the coordinated verbs are similar in meaning, and take an argument that play the semantic role of Stimulus:

(6)	***tôn***	*oú ti*	*metatrépēi*	*oud'*	*alegízeis* ø*?*
	this:GEN.PL	NEGPTC	care.for:PRS.M/P.2SG	NEG	mind:PRS.2SG

‘You neither care for **them**, nor mind **(them).**’ (*Il.*1.160=*Il.*12.238)

Interestingly, in (7) the case assumed for the null object is not the genitive but the dative, as the middle voice of *peíthō* ‘obey’ typically occurs in the NomDat construction:

(7)	*ȍs*	*éphth',*	*hoì*	*ára*	***toû***	*mála*
	so	speak:IMPF.M/P.3SG	3PL.NOM	PTC	2SG.GEN	very
	mèn	*klúon*	*ēd'*	*epíthonto*	ø	
	PTC	hear:IMPF.3SG	and	obey:AOR.MID.3PL		

‘So he spoke, and they readily listened **to him**, and obeyed **(him).**’ (*Il.*7.379)

[6] One can wonder whether examples such as that in (5) are simply cases of object sharing. Similar word orders are in fact allowed in configurational languages as well (cf. it. *Marco compra e legge ottimi libri* ‘Marco buys and reads very nice books’).

4.3. Conjunct participle

We have collected 41 occurrences in which a non-accusative null object can be assumed with the conjunct participle. We have also found numerous contexts in which the non-accusative object is the indirect object of one the two verbs (or of both), as shown in example (8):

(8) *tá* ***hoi*** *xeînos… dôke tukhḗsas* ø
REL.ACC.PL 3SG.DAT friend:NOM give:AOR.3SG meet:AOR.PTCP.NOM
'(Gifts) which a friend had given **him** when he met (**him**).' (*Od*.21.13)

In (8) the overt object is the indirect object of the finite verb *dídōmi* 'give', but the null object of the participle of *tugkhánō* 'happen, meet'. One also finds occurrences showing the symmetrical situation, in which the expressed object is the direct object of the finite verb and the null indirect object of the participle. Although interesting, we decided to exclude both these contexts, because indirect objects are out of the scope of our paper. A third type of occurrences is also attested (cf. e.g. *Od.*10.295, 322), in which the direct object of the finite verb is expressed and it is possible to assume a null object for the infinitive that in turn depends on the conjunct participle. We have left out of this paper such occurrences as well.

4.3.1. Dative null objects with conjunct participle

In 29 occurrences out of 41, the overt object is in dative and one can assume a dative null object as both the finite verb and the participle occur in the NomDat construction. In most cases, the finite and non-finite verbs are similar in meaning, as shown in example (9), in which both verbs refer to anger and take the dative Stimulus *toi*:

(9) ***toi*** *déspoina kotessaménē* ø
2SG.DAT mistress:NOM.F be.angry.at:AOR.PTCP.MID.NOM.F
khalepḗnēi
be.severe.with:AOR.SBJ.3SG
'(Being afraid that) the mistress could be angry (**with you**) and severe **with you**.' (*Od*.5.147=*Od*.19.83)

In other cases, the two verbs are different; however, the semantic roles of the expressed and null objects are the same or, at least, they can be both encoded by the dative case, as shown in (10):

(10) *oudè* *mákhesthai* ø *elthṑn* ***dusmenéessin***
NEG fight:PRS.INF.M/P go:AOR.PTCP enemy:DAT.PL
'(I cannot) fight **the enemies**, going (**among them**).' (*Il*.16.520)

In (10) one can say that the finite and non-finite verbs share the dative object *dusmenéessin* 'enemies', which plays the semantic role of human Theme with the finite verb *mákhomai* 'fight' and that of human Goal with the conjunct participle *elthōn* 'going'. Both verbs, as other verbs of fighting and of movement, occur in the NomDat construction in the Homeric poems.

4.3.2. Genitive null objects with conjunct participle

In 12 contexts out of 41, the expressed genitive object of the finite verb in the main sentence allows assuming a genitive null object for the conjunct participle. As observed for dative null objects, we often found that the two verbs are similar in meaning and usually occur in the NomGen construction, as shown in (11):

(11) *kheirì* *labṑn* ø *peirḗsato* ***neurês***
hand:DAT take:AOR.PTCP.NOM try:AOR.3SG.MID string:GEN
'After grasping **the string** with the hand, he tried (**it**).' (*Od*.21.410)

As for dative null objects, in other occurrences the finite and non-finite verbs are different, but still typically occur in the NomGen construction. Consider the example in (12) with the finite verb *éramai* 'love' and the participle of *eîdon* 'see', both experiential verbs typically taking genitive objects encoding the semantic role of Stimulus:

(12) ***tês*** *dè* *kratùs* *Argeiphóntēs* *ērásat',* … ø *idṑn*
3SG.GEN.F PTC strong:NOM A.:NOM love:AOR.3SG.MID see:AOR.PTCP
'Strong Argheiphontes became enamoured **of her**, when he saw (**her**).' (*Il*.16.182)

4.4 *Prepositional phrases as antecedents of null objects*

In examples (13) and (14) the non-accusative null object is recoverable from an antecedent prepositional phrase. In (13) the null object can be assumed as a dative object, as the verb *otrúnō* 'exhort' usually occurs in the NomDat construction; in (14) the null object can be assumed as a genitive object, as the verb *megaírō* 'hate' usually occurs in the NomGen construction.

(13) *autàr* *égōge* *speúsomai* ***eis*** ***Akhilêa,*** *hín'* *otrúnō*
but 1SG.NOM hasten:FUT.1SG to A.:ACC so.that urge:PRS.1SG

ø	*polemízein*
	battle:PRS.INF

‘But I will hasten **to Achilles**, that I may urge **(him)** on to do battle.’ (*Il.*15.401-402)

(14)

táōn	*oú*	*toi*		*egṑ*	***prósth'***
DEM.GEN.PL	NEG	2SG.DAT		1SG.NOM	before
hístamai		*ou*	*dè*	*megaírō*	ø
stand:PRS.M/P.1SG		NEG	PTC	regard.as.too.great:PRS.1SG	

‘I do not stand **before them** in your defense, I do not account **(them)** too greatly.’ (*Il.*4.54)

5 Conclusions

In this paper, based on the existing analysis of accusative referential null objects in Ancient Greek, we have discussed two types of syntactic environments in which the second argument of two verbs is expressed only once, that is, contexts with coordination reduction and with conjunct participles.

For both types of occurrences we have observed what follows. The omission of the genitive or dative objects in the two analyzed contexts seems to be high in the scale of obligatoriness. In addition, the free positioning of the overt object, which can occur either close to the first occurring verb or close to the second one, shows that it is impossible to decide which verb governs it (or whether both verbs do).

From a quantitative point of view, we have noted that the passages in which a dative null object can be assumed are far more frequent than those in which a genitive null object can be assumed. These data can be related to the absolute frequencies of the NomDat and NomGen constructions in Homer. As argued in Sausa [9], there is no great disparity between the frequency of the NomGen and of the NomDat constructions in terms of verb tokens instantiating them, but the latter is slightly higher.

Furthermore, in the majority of contexts, the two verbs are similar in meaning and take a direct object playing the same semantic role. However, exceptions to this tendency also occur, cf. for example (10).

The case assumed for the null object is mostly the same as the case of the overt object. Once again, exceptions are attested, e.g. (7). Interestingly, from our data, which for reasons of space we have only partially discussed here, it did not emerge that differences in the selection of the construction – NomAcc, NomDat, NomGen – can prevent the omission of the object.

The temporary conclusion concerning the behavior of non-accusative null objects of coordinated verbs and of conjunct participles in the Homeric poems is that they seem to behave in the same way as accusative null objects. Of course more data are necessary to draw final conclusions on genitive and dative null objects. In particular, it would be also desirable analyzing contexts of high topic continuity and yes-no questions, though not easily automatically retrievable. Moreover, it could be interesting to extend the analysis to other syntactic environments than those discussed for accusative null objects (cf. examples with finite verb and infinitive depending on a conjunct participle), and to indirect objects. Finally, another intriguing issue might be investigating what constitutes the other side of the same research question: detecting under which syntactic, semantic, and pragmatic conditions objects can be repeated, even in those contexts in which omission is the usual tendency, cf. example (2).

References

[1] Baker, Mark (2001) Configurationality and polysynthesis. In Haspelmath, Martin *et al.* (eds.) *Language Typology and Language Universals. An International Handbook*, vol. 2, pp. 1433–1441. Berlin: Mouton de Gruyter.

[2] Barðdal, Johanna (2008) *Productivity. Evidence from Case and Argument Structure in Icelandic*. Amsterdam/Philadelphia: John Benjamins.

[3] Devine, Andrew L. D. and Stephens, Laurence D. (2000) *Discontinuous syntax: Hyperbaton in Greek.* Oxford: Oxford University Press.

[4] Goldberg, Adele E. (1995) *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago: University of Chicago Press.

[5] Harris-Delisle, Helga (1978) Coordination reduction. In Greenberg, Joseph (ed.) *Universals of Human Language*. Stanford: UP, pp. 515–83.

[6] Haug, Dag T.T. (2012) Syntactic conditions on null arguments in the Indo-European Bible translations. In *Acta Linguistica Hafniensia* 44:2, pp. 129–141.

[7] Keydana, Götz and Luraghi, Silvia (2012) Definite referential null objects in Vedic Sanskrit and Ancient Greek. In *Acta Linguistica Hafniensia* 44:2, 116–128.

[8] Luraghi, Silvia (2003) Definite referential null objects in Ancient Greek. In *Indogermanische Forschungen* 108, 169–196.

[9] Sausa, Eleonora (2015) *Argument structure constructions in Homeric Greek. An analysis of bivalent verb.* Ph.D. thesis. Pavia: University of Pavia.

The Digital Fragmenta Historicorum Graecorum and the Ancient Greek-Latin Dynamic Lexicon

Tariq Yousef and Monica Berti

Alexander von Humboldt Lehrstuhl für Digital Humanities
Institut für Informatik - Universität Leipzig
E-mail: name.surname@uni-leipzig.de

Abstract

This paper describes a model that provides training data for a word alignment system that will be used to identify the translation relationships among the words in the parallel texts (Greek/Latin) of the bilingual corpus of the Digital Fragmenta Historicorum Graecorum (DFHG).

1 Introduction

Statistical machine translation uses alignment models to extract and identify translation correspondences between words and phrases in two parallel texts in two different languages. The aligned pairs of words or phrases are used as training data for machine translation systems.

The goal of this paper is to provide training data for a word alignment system and these data will be used to identify the translation relationships among the words in the parallel texts (Greek/Latin) of the bilingual corpus of the Digital Fragmenta Historicorum Graecorum (DFHG). The biggest challenge is that large digitized Ancient Greek/Latin lexica are publicly unavailable. The only available dictionary is Glosbe, which contains about 1600 Ancient Greek/Latin phrases entered by users[1]. Glosbe is an unreliable source, because anyone can update the dictionary without supervision. Moreover, the size of the database is not large enough to build an alignment system only relying on it. This paper investigates a simple and effective method for automatic bilingual lexicon extraction (Ancient Greek/Latin) from the available aligned bilingual texts (Ancient Greek/English and Latin/English) produced by the Dynamic Lexicon project of the Perseus Digital Library.

[1] Glosbe contains thousands of dictionaries for every existing pair of languages: www.glosbe.com/grc/la

2 The DFHG Corpus and the Dynamic Lexicon

The DFHG is a project of the Alexander von Humboldt Chair of Digital Humanities at the University of Leipzig that is producing a digital edition of the five volumes of the Fragmenta Historicorum Graecorum edited by Karl Müller in the 19th century[2]. This corpus includes extracts from ancient sources that preserve quotations and text reuses of Greek authors and works that are now lost. More than 600 fragmentary authors are collected in the volume and the sources range from the 6th century BC through the 7th century CE. The content is arranged by author and the volumes provide scholars with the Greek texts of the fragments (or Latin texts when the witnesses are Latin) and their modern Latin translations produced by Müller. Introductions and commentaries are in Latin [2]. The Dynamic Lexicon is a project of the Perseus Digital Library to automatically create bilingual dictionaries (Greek/English and Latin/English) using parallel texts (source texts in Greek or Latin aligned with their English translations) along with the syntactic data encoded in treebanks[3]. The final goal is to enrich the Perseus Dynamic Lexicon with Greek/Latin pairs and to extend the work also to other sources beyond the fragmentary ones.

3 Previous Work

There are many approaches to construct bilingual lexica by using a third language (usually English). Tanaka and Umemura [8] uses an Inverse Consultation (IC) method to produce a Japanese/French lexicon using English as a bridge language. Ács [1] extends the IC method up to 53 pivot languages to improve the accuracy of the lexicon, which relies on the fact that pairs found via several intermediate languages are more correct. Bond [3] uses semantic classes along with an intermediate language to produce Japanese/Malay dictionary. Paik [7] improves a method (multi-pivot criterion) to produce a Korean/Japanese lexicon using English as an intermediate language and shared Chinese characters among Japanese and Korean words. Noisy translations are a big problem and therefore Kaji [4] introduces distributional similarity (DS) as a measure to avoid noisy translations produced by triangulation. In the next sections we introduce our proposed approach to produce Ancient Greek/Latin lexicon via English as a bridge language, and JACCARD Index as a similarity metric to measure the quality of translation pairs in order to eliminate noisy translations.

4 Proposed Approach

The starting point of our approach is to provide as much parallel texts as possible to extract all possible translation candidates. The Perseus Digital Library con-

[2] http://www.dh.uni-leipzig.de/wo/dfhg/

[3] http://nlp.perseus.tufts.edu/lexicon/

tains approximately 10.5 million words of Latin source texts, 13.5 million words of Greek, and 44.5 million words of English. The texts are all public-domain materials that have been scanned, OCR'd, and formatted into TEI-compliant XML [4]. The Perseus Digital Library contains at least one English translation for most of its Latin and Greek prose and poetry texts. Our Corpus consists of 163 parallel documents aligned at the word level (104 Ancient Greek/English documents and 59 Latin/English documents). The Greek/English dataset consists approximately of 210 thousand sentence pairs with 4.32 million Greek words, whereas the Latin/English dataset consists approximately of 123 thousand sentence pairs with 2.33 million Latin words. The parallel texts are aligned on a sentence level using Moore's Bilingual Sentence Aligner [6], which aligns the sentences with a very high precision (one-to-one alignment). The Giza++ toolkit is used to align the sentence pairs at the level of individual words.

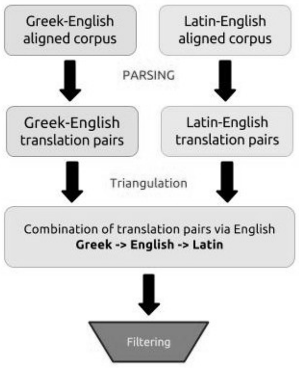
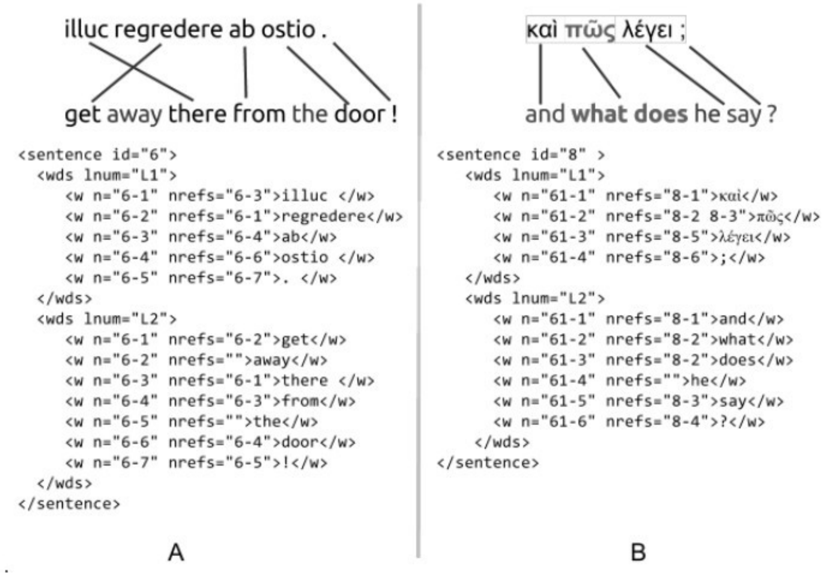

Greek-English aligned corpus
Latin-English aligned corpus
PARSING
Greek-English translation pairs
Latin-English translation pairs
Triangulation
Combination of translation pairs via English
Greek -> English -> Latin
Filtering

Figure 1: Explanation of the method

4.1 Preprocessing

In this stage we are going to parse the data sets we have in XML format (Fig. 2). Each document has a Perseus-id and consists of sentences in the original language



```
<sentence id="6">
  <wds lnum="L1">
     <w n="6-1" nrefs="6-3">illuc </w>
     <w n="6-2" nrefs="6-1">regredere</w>
     <w n="6-3" nrefs="6-4">ab</w>
     <w n="6-4" nrefs="6-6">ostio </w>
     <w n="6-5" nrefs="6-7">. </w>
  </wds>
  <wds lnum="L2">
     <w n="6-1" nrefs="6-2">get</w>
     <w n="6-2" nrefs="">away</w>
     <w n="6-3" nrefs="6-1">there </w>
     <w n="6-4" nrefs="6-3">from</w>
     <w n="6-5" nrefs="">the</w>
     <w n="6-6" nrefs="6-4">door</w>
     <w n="6-7" nrefs="6-5">!</w>
  </wds>
</sentence>
```

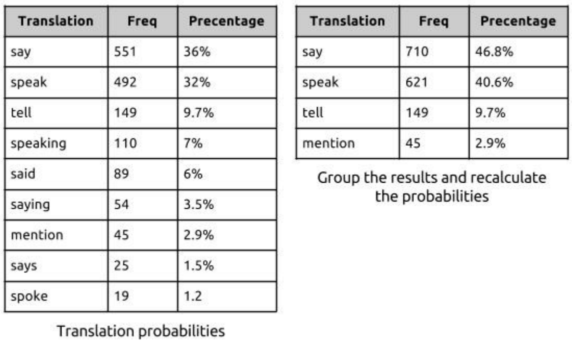
```
<sentence id="8" >
   <wds lnum="L1">
      <w n="61-1" nrefs="8-1">καὶ</w>
      <w n="61-2" nrefs="8-2 8-3">πῶς</w>
      <w n="61-3" nrefs="8-5">λέγει</w>
      <w n="61-4" nrefs="8-6">;</w>
   </wds>
   <wds lnum="L2">
      <w n="61-1" nrefs="8-1">and</w>
      <w n="61-2" nrefs="8-2">what</w>
      <w n="61-3" nrefs="8-2">does</w>
      <w n="61-4" nrefs="">he</w>
      <w n="61-5" nrefs="8-3">say</w>
      <w n="61-6" nrefs="8-4">?</w>
    </wds>
</sentence>
```

Figure 2: The aligned sentences in XML format

[4] https://github.com/PerseusDL

(Ancient Greek or Latin) and its translation in English (Fig. 2A). Each Latin or Greek word is aligned to one word in the English text (one-to-one Alignment), but in some cases a word in the original language can be aligned to many words (one-to-many / many-to-one) or not aligned at all (Fig. 2B).

Lemmatization of English translations will produce better results, because that will reduce the number of translation candidates as we can see in this example: The Greek word λέγειν is translated with "say", "speak", "tell", "speaking", "said", "saying", "mention", "says", "spoke". Many of the translation candidates share the same lemma (*say* for "said", "saying", "says"), (*speak*, "speaking", "spoken"). Before the lemmatization there were nine translation candidates and after the lemmatization there are only four candidates, showing therefore the change of frequencies.



Translation	Freq	Precentage
say	551	36%
speak	492	32%
tell	149	9.7%
speaking	110	7%
said	89	6%
saying	54	3.5%
mention	45	2.9%
says	25	1.5%
spoke	19	1.2

Translation probabilities

Translation	Freq	Precentage
say	710	46.8%
speak	621	40.6%
tell	149	9.7%
mention	45	2.9%

Group the results and recalculate the probabilities

Figure 3: Lemmatization of English translations

4.2 Triangulation

Triangulation is based on the assumption that two expressions are likely to be translations if they are translations of the same word in a third language. We will use triangulation to extract the Greek/Latin pairs via English. In order to do that, we query our datasets to get the Greek and Latin words that share the same English translation along with their frequencies. See Figure 4.

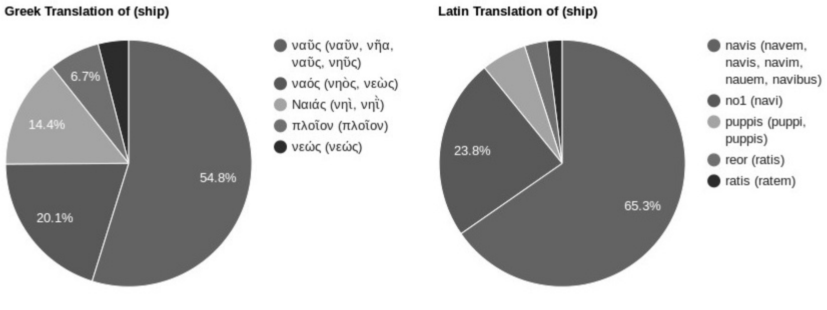


Figure 4: An example of triangulation

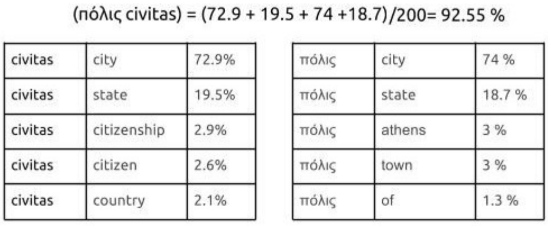
The English word *ship* is translated with the Greek word ναῦς (54.8%), with ναός (21.5%) and so on; the same English word ship is translated with the Latin word navis (65.3%), with no (23.8%), and so on. The extracted pairs via triangulation are the following (ναῦς, navis), (ναῦς, no), (ναός, navis), (ναός, no). These pairs don't have the same level of relatedness, therefore we have to filter the results to keep only strong related pairs.

4.3 Translation-Pairs filtering

The translation pairs are not completely correct, because there are still some translation errors. In order to eliminate incorrect pairs, we will use a similarity metric to measure the similarity or the relatedness between every Greek/Latin pairs. The Jaccard coefficient [5] measures the similarity between finite sample sets (in our case two sets), and it is defined as the size of the intersection divided by the size of the union of the sample sets:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

A and B in equation 2 are two vectors of translation probabilities (Greek/English, Latin/English). For example, the relatedness between the Greek word πόλις and the Latin word *civitas* is reported in figure 5



(πόλις civitas) = (72.9 + 19.5 + 74 +18.7)/200= 92.55 %

civitas	city	72.9%
civitas	state	19.5%
civitas	citizenship	2.9%
civitas	citizen	2.6%
civitas	country	2.1%

πόλις	city	74 %
πόλις	state	18.7 %
πόλις	athens	3 %
πόλις	town	3 %
πόλις	of	1.3 %

Figure 5: Use of Jaccard algorithm

5 Evaluation

The quality evaluation of translations candidates extracted by the proposed method is done manually with the help of humanists. We have randomly selected 200 translation pairs obtained via the proposed method with different frequencies (high and low) and different JACCARD Co values. Each pair should be assigned into one of four categories: Correct, small difference, big difference and incorrect. We employed the mean reciprocal rank (MRR) [9] to assess the performance. We assigned each category a score (Reciprocal Rank) (Table 1).

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} RR_i \tag{2}$$

Category	**Reciprocal Rank (RR)**
Correct	1
Small difference	0.75
Big difference	0.25
Incorrect	0

Table 1: Group the results and recalculate the probabilities

We have to determine a threshold to classify the translation pairs as accepted or not accepted. High threshold yields high accuracy lexicon but with less number of entries, whereas low threshold produces more translation pairs with lower accuracy, as we can see in the table above (Table 2).

Jaccard Co	**0.60 <**	**0.70 <**	**0.80 <**	**0.90 <**
# Pairs	200	150	100	50
MRR	%61.25	%74	%87.50	%94.50

Table 2: MRR Scores

6 Conclusion

The proposed method is language-independent and it can be used to build a bilingual lexicon between any language pairs with aligned corpora that share a pivot language. The accuracy of the method depends on two factors: **1) The size of the aligned-parallel corpora** plays an important role to improve the accuracy of the lexicon: bigger corpora produce better translation probability distribution and more translation candidates which yield a more accurate lexicon, and they also cover more words; **2) The quality of the aligner** used to align the parallel corpora: manually aligned corpora yield more accurate results, whereas automatic alignment tools produce some noisy translations; in our case Giza++ has been used to align the parallel corpora.

References

[1] Judit Ács (2014). Pivot-based Multilingual Dictionary Building using Wiktionary. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland

[2] Berti, M. et al. "The Linked Fragment: TEI and the Encoding of Text Reuses of Lost Authors". In *Journal of the Text Encoding Initiative 8 (2014-2015)* (selected papers from the 2013 TEI Conference) doi 10.4000/jtei.1218

[3] Francis Bond, Ruhaida Binti Sulong, Takefumi Yamazaki, and Kentaro Ogura. 2001. *Design and construction of a machine-tractable Japanese Malay dictionary*. In *MT Summit VIII*, pp. 53– 58, Santiago de Compostela, Spain

[4] Kaji, H., Tamamura, S., and Erdenebat, D. (2008). Automatic construction of a Japanese-Chinese dictionary via English. In *LREC, volume 2008*, pp. 699–706.

[5] Jaccard, P. (1901). Distribution de la flore alpine dans le Bassin des Drouces et dans quelques regions voisines. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37(140), pp. 241–272.

[6] Moore, Robert. 2002. Fast and accurate sentence alignment of bilingual corpora. In *Proceedings of 5th Conference of the Association for Machine Translation* in the Americas, pp. 135–244.

[7] Kyonghee Paik, Francis Bond, and Satoshi Shirai. 2001. Using multiple pivots to align Korean and Japanese lexical resources. In *NLPRS-2001* Tokyo, Japan, pp. 63–70.

[8] Tanaka, K., Umemura, K. 1994. Construction of a bilingual dictionary intermediated by a third language, *Proceedings of COLING- 94*, pp. 297-303.

[9] Voorhees, E.M. 1999. The trec-8 question answering track report. In *Proceedings of the 8 th Text Retrieval Conference*.

[10] Yousef, Tariq. 2015. Word Alignment and Named-Entity Recognition applied to Greek Text Reuse. *MSc's Thesis. Alexander von Humboldt Lehrstuhl für Digital Humanities, Universität Leipzig*.