Volume 26 - Number 2 - 2011 - ISSN 2190-6858

JLCL
Journal for Language Technology
and Computational Linguistics

Annotation of Corpora for Research in the Humanities

Proceedings of the ACRH Workshop, Heidelberg, 5 Jan. 2012

Herausgegeben von / *Edited by*
Francesco Mambrini, Marco Passarotti and Caroline Sporleder

GSCL Gesellschaft für Sprachtechnologie & Computerlinguistik

Contents

Impressum

Herausgeber	Gesellschaft für Sprachtechnologie und Computerlinguistik (GSCL)
Aktuelle Ausgabe	Band 26 – 2011 – Heft 2 "Annotation of Corpora for Research in the Humanities: Proceedings of the ACRH Workshop, 5. January 2012, Heidelberg University, Germany"
Gastherausgeber	Francesco Mambrini, Marco Passarotti and Caroline Sporleder
Anschrift der Redaktion	Lothar Lemnitzer Berlin-Brandenburgische Akademie der Wissenschaften Jägerstr. 22/23 10117 Berlin lemnitzer@bbaw.de
ISSN	2190-6858
Erscheinungsweise	2 Hefte im Jahr, Publikation nur elektronisch
Online-Präsenz	www.jlcl.org

Francesco Mambrini, Marco Passarotti, Caroline Sporleder

Preface

In almost every culture of the world, the sciences of language and the study of cultural-heritage documents are inextricably bound. In Western tradition, as in many others, the need to preserve the literary or historical legacy of the past gave the strongest input to the development of a formalized grammatical speculation.

Since the foundation of linguistics as a discipline, the interaction has proceeded in both directions. Linguistics has profited from the huge amount of material that was gathered by philologists and historians, along with the full apparatus of concepts and problems that originated from their work. Humanities, in their turn, have often seen in linguistics a model of a rigorous scientific approach to a social and historically complex phenomenon like human language.

It is not by chance, thus, that the work of scholars engaged in historical and literary studies was not alien to one of the most original development in contemporary linguistics, namely the creation and use of the first digital corpora. It is worth remembering that the *Index Thomisticus*, which is considered the starting point for both corpus-linguistics and digital humanities, was designed in order to allow a more rigorous approach to the philosophy of Thomas Aquinas.

From the time of the first pioneering projects, the concepts and methodologies of corpus linguistics (including the notion of "corpus" itself) have been widely debated; technologies for storing and processing digital information have also changed radically. Nowadays, computational and corpus-linguistics have grown into autonomous disciplines, with their own set of required expertises. As in many other scientific fields, autonomy means inevitably a certain degree of isolation.

The loss of contact between corpus-linguistics and humanities is particularly visible in one crucial aspect. Although quantitative or stylometric approaches to large collections of documents are increasingly frequent in literary or historical studies, the available resources are not quite at the same level as those used by linguists.

The "Workshop on Annotation of Corpora for Research in Humanities" (ACRH) was held in Heidelberg University on January 5^{th}. The event was co-located to the 10^{th} edition of the international workshop on "Treebanks and Linguistic Theories" (TLT-10), also held in Heidelberg on January 6^{th}-7^{th}. The ACRH workshop was conceived to address one special aspect where the aforementioned gap between the two disciplines is particularly visible: the creation and exploitation of annotated corpora for the needs of research in the different fields of humanities (philology, literary studies, history, philosophy etc.).

The availability of annotated corpora is indeed an area where humanities and corpus-linguistics are more distant. Many corpora that play a relevant role for research in humanities are today available in digital format (theatrical plays, contemporary novels, critical literature, literary reviews etc.). Yet, only a few of them are linguistically tagged, while most still lack any linguistic annotation. Standards for the encoding of a vast

number of information on texts used in the humanities and social sciences, such as the TEI, are becoming increasingly popular and are adopted for the most recent efforts of digitization of collections. But although there is an agreement on the meta-language for the description of texts, what features an annotated corpus for research in the humanities must have and how annotation must be performed in order to conform to the strict requirements described by corpus-linguistics is a question that is still not sufficiently debated. Fostering this discussion was precisely the aim of our workshop.

As the work for the creation of annotated resources is in most cases still in a very early stage, it is only natural that most of the attention is focused on the annotation process. There are a number of peculiarities that distinguish corpora created for the special use of research in humanities from the standard linguistic corpora, such as the Brown Corpus or the Penn Treebank. These special features, which affect both the theoretical debate and the practical task of developing tools for the work of corpus creation, are well reflected by the papers that are collected here.

Two sets of problems arise from the peculiar nature of the documents used by scholars in the humanities, and these issues affect both the work of manual and (semi-)automatic annotation.

On the one hand, even the preliminary task of selecting and digitizing the relevant materials for the corpus must tackle peculiar problems, that can be generally ignored in the case of standard linguistic corpora. Often, as in the case of ancient or badly preserved documents, even the "raw" materials are controversial or need special philological care. Different kinds of expertise are required in order to build a corpus of such documents. This multidisciplinary program, though, can be too vast and too difficult to handle on a large scale. In such situations, the general competence of a native speaker with linguistic training (which is typically sufficient for the annotation of standard linguistic corpora) is inadequate; in most cases linguistic competence must be complemented by a strong training on the particular problems, such as philological, historical, epistemological issues, raised by the materials. New standards and models of manual annotation must be devised in order to integrate the different competences of corpus linguists and specialists for the texts. The challenge of adapting tools and concepts for manual annotation is manifest in several papers presented at ACRH (see especially the contributions by: Manca *et al.*; Bech and Eide; Kaplan *et al.*; Korkiakangas and Passarotti).

On the other hand, automated and semi-automated annotation is difficult due to another special feature of corpora of cultural-heritage documents. The high level of intra-linguistic variation, that is observed, especially in diachronic corpora, is clearly a crucial factor. Almost all the available NLP tools are developed for modern corpora and tested on standardized varieties of a languages. Yet the documents collected for their cultural significance often predate the establishment of modern writing conventions. In other cases, historical documents can reflect a large spectrum of regional dialects that are not yet unified in a national standard language; other times, literary documents are deliberately violating the rules of ordinary languages. How NLP tools (like parsers or pos-taggers) can be effectively adapted to this situation is a crucial practical question, which in its turn can provide new evidence to scholars working in historical linguistics or

literary studies. This is perhaps the most debated topic of the workshop (see especially the papers by: Dipper; Hendrickx and Marquilhas; Sukhareva *et al.*; Rögnvaldsson *et al.*; Ekbal *et al.*; Piotrowski and Höfler; Skjærholt).

The motivation for creating and annotating the collections is also a factor that must be kept in mind. The need of preserving endangered cultural-heritage documents often precedes (and determines) the scientific motivation of creating resources for scholars. The representativeness of those corpora is another difficult question which is tied to that of the motivations. Very often, for corpora of historical languages or corpora designed for specific research purposes on narrow domains in the humanities, there is not much to be chosen; corpus designers have rather to deal with the scanty material that is available. This is another important difference with standard linguistic corpora, where the size of the collection can be determined in advance and adjusted to the needs of researchers. The small extension of the corpora used for research in the humanities is a big challenge both for the quantitative analysis of data and for the performances of the NLP tools, of which many of the papers presented here are well aware. There are many other questions that remain open to discussion. What and how to annotate is another crucial problem in a pluralistic and often very controversial area like the different fields of the humanities. The dilemma between the adherence to a strict theoretical frame and the needs of portability of theoretically independent resources is of course not unknown to linguists, but is even more challenging for scholars in the humanities.

A typical example of resources discussed in many papers of ACRH is a corpus created by digitizing collections of manuscripts (e.g. see the paper by Dipper), archives of hand-written papers (e.g. Hendrickx and Marquilhas), or ancient printed books. In such cases, the textual content that must be tagged is only a part of information that define these cultural artifacts. Linguistic analysis must therefore be associated with other types of annotation, that should encode aspects such as description of e.g. the medium, the script, the collocation of the text in the page, and so on. The interaction of these different levels of information and with the standard markup languages whose aim is the comprehensive description of textual artifacts (of which TEI is perhaps the most widespread) is a problem which requires further attention.

The quantity and quality of the submitted papers are an unmistakable proof of the great interest in the area of annotated digital corpora for humanities.

The call for papers for ACRH requested unpublished, completed work. We received 23 submissions. The submissions were authored by researchers from 15 different countries in America, Asia and Europe. Each submission was evaluated by three reviewers. The Programme Committee consisted of 18 members (including the 3 co-chairs) from 8 different countries. They all worked as reviewers. Based on their scores and comments on the content and quality of the papers, 14 papers were accepted for presentation and publication, which corresponds to an acceptance rate of 58.3%. 8 papers were presented orally, 5 as posters; finally, 1 paper, which was proposed and accepted to both workshops (ACRH and TLT), was considered by the two Programme Committees to be more appropriate for presentation at TLT10.

As we have mentioned, the accepted submissions cover a wide range of topics. The languages studied are: English, French, German, Greek, Icelandic, Italian, Japanese, Latin, Portuguese and Spanish. Some of the corpora that were discussed in the paper cover a significant variety of diachronic phases of these languages; the corpora presented at ACRH extend over a very broad time span (from Antiquity, to Early and Late Middle-ages, to Modern era and our days) and come from different domains (religious texts, private letters, epic poems and chronicles etc.).

The ACRH Workshop was introduced by a keynote lecture by Professor Gregory R. Crane, editor in chief of the Perseus Project and chair of the Classics Department of Tufts University (Medford, USA). Professor Crane has long distinguished himself in pursuing a high-level scholarly activity in the field of Greek and Latin literature; but, especially, the Perseus Project has been providing classicists of all levels and from all over the world with access to the primary sources for their work (both texts and archaeological artifacts) within a digital cyberinfrastructure that relies largely on the implementation of specific NLP technologies. Recently, the Perseus Project has undertaken the creation of the first treebank for Classical Latin and Ancient Greek (The Greek and Latin Dependency Treebank). While actively promoting the work of annotation to the community of scholars and students in Classics, Professor Crane is particularly engaged in showing how the inclusion of a treebank into the larger context of a cultural heritage library could benefit both the digital library and the annotated corpus itself.

The event was made possible by the work and generosity of the local organizers and hosts at the University of Heidelberg. Our gratitude and acknowledgements go to them. We would also like to thank to board of the Journal for Language Technology and Computational Linguistics for accepting the publications of the ACRH proceedings.

The ACRH Co-Chairs and Organizers
Francesco Mambrini, Università Cattolica del Sacro Cuore, Milan, Italy
Marco Passarotti, Università Cattolica del Sacro Cuore, Milan, Italy
Caroline Sporleder, Saarland University, Saarbrücken, Germany

Gregory R. Crane

Linguistic annotation, the reunification of linguistics and philology, and the reinvention of the Humanities for a global age

This paper addresses the critical role that treebanks in particular and linguistic annotation in general must play if the Humanities are to advance the intellectual life of society as a whole. During the twentieth century we saw a rise in specialization that not only separated the practices of philology and linguistics among different researchers but wholly separate (and sometimes conflicting) departments. The reunification of linguistics with philology is an essential element in the evolution of the Humanities and serves three critical functions.

First, linguistic annotation, both machine generated and human curated, is an essential element both for large scale analysis of topics that cross more languages than any research can study, much less master, and for the intensive analysis of individual source and topics. Second, the associated changes in the scale of research demand that we draw upon more cultural and linguistic expertise than the established universities of North America and Europe can offer – we must enlist new collaborators in nations such as Egypt, India, and China, whom boundaries of language and of culture have often kept isolated. Third, even a global network of advanced scholars and library professionals is not sufficient to analyze sources in thousands of languages produced over thousands of years. We must develop student researchers and citizen scholars and a new participatory of scholarship. The potential consequences of these three changes are immense and each depends upon contributions by members of this workshop.

Kristin Bech and Kristine Gunn Eide

The annotation of morphology, syntax and information structure in a multilayered diachronic corpus

This paper describes the annotation scheme we use for old Germanic and Romance languages, with particular focus on syntax and information structure, and the issues of economy and disambiguation. We also discuss some of the annotation problems we have had to solve, in order to demonstrate the complexity of this kind of linguistic annotation.

1 Introduction

This paper reports on annotation work in the project Information Structure and Word Order Change in Germanic and Romance Languages (ISWOC[1]). In the project, we annotate morphology, syntax and information structure in a corpus consisting of texts from old Germanic (English, Norse/Norwegian, German[2]) and Romance (French, Spanish, Portuguese) languages. The purpose of the project is to study the relation between word order change and information structure in these languages, as all these languages had verb-second structures at some stage, but have developed in different directions. Multilayered means that the application is designed in such a way that the different levels build on each other, i.e. the application makes guesses for syntax on the basis of the morphological annotation, and the syntactic annotation in turn provides suggestions for which elements should be included in the information structure annotation (cf. HAUG ET AL. 2009). The aim of this paper is to describe the annotation principles, both from a theoretical and practical point of view, and discuss some of the problems we encounter in the annotation. We focus particularly on syntactic annotation and information structure annotation, and on how to balance the sometimes conflicting requirements of economy and disambiguation.

2 Economy vs disambiguation

Ideally, an annotation would enable us to retrieve *all* the examples of the particular phenomenon we are looking for as well as *only* the examples we are looking for, with as uncomplicated a search string as possible. However, considering the formal and structural ambiguities that we find in all languages, this ideal is not easily attainable. Also, since the annotation is a tool to enable linguistic research rather than a linguistic analysis in itself, it is not the ultimate goal, nor is it feasible, to disambiguate every ambiguous expression. When we evaluate whether or not to disambiguate, we must ask ourselves (1) to what extent that expression or construction will be retrievable through other means, (2) how much it will cost in terms of working time (disambiguation is very time-consuming) and (3) whether or not the ambiguity is a reflection of accidental formal likeness (for instance the Portuguese homonyms (a) reflexive pronoun *se* and (b) subjunction *se* 'if'), or whether it reflects a "constructed" ambiguity, perhaps because the grammatical categories that we use are in themselves less clear-cut than we would like them to be (cf. the discussion concerning demonstratives in section 3). Since we are constructing a multilayered corpus, we have to ask ourselves if, for instance, a certain formal morphological ambiguity will be disambiguated in the syntax layer, or the other way around, if syntactic ambiguity can be resolved in the

morphology layer, since there is little gain in meticulously disambiguating something in one layer that can easily be retrieved in another. We also have to consider whether it is at all possible to disambiguate; in many cases it is not possible.

3 Morphological annotation

The first level of annotation is morphology. In some cases we manage to import morphological annotation from other electronic corpora, but the annotation still involves a considerable amount of manual work. The challenges we experienced with regard to morphology typically had to do with what word classes to include, and with the fact that some words are ambiguous, or not easily classifiable, in terms of word class. In addition, since we wish to compare Germanic and Romance languages, it was important to keep the morphological specifications as consistent as possible. However, in some cases, we had to differentiate. One example is the distinction between demonstrative pronouns and demonstrative determiners. In Old Norse and Old English, there is no morphological distinction between these two categories, and demonstratives are always tagged as determiners, whether they are used as attributes or pronouns. Hence, there is no morphological disambiguation between the pronominal use of *þá* in (1) and the determiner in (2), both examples from Old Norse.

(1) ec vil tala við *þá* (*Strengleikar*, 13th c.)
I will speak with *them*.ACC.PL.M/*her*.ACC.SG.F

(2) um *þá* daga var þar iafnan ufriðr ok bardagar (*Strengleikar*, 13th c.)
in *those* days was there often unrest and war
'In those days there was often unrest and war'

The difference between *þá* in (1) and *þá* in (2) will rather appear in the syntactic annotation, where *þá* in (1) belongs to an argument category and *þá* in (2) is analyzed as an attribute.

Portuguese and Spanish have the same formal identity between demonstrative pronouns and determiners, but in these languages there is also a class of non-inflectional demonstrative pronouns that never modify nouns: *isto/esto* 'this' (1st person), *isso/eso* 'that' (2nd person), *aquilo/aquello* 'that' (3rd person). They are usually classified as neuter pronouns, as opposed to their inflected counterparts, which are used either pronominally or as determiners. We have chosen to keep these non-inflected pronouns together with the inflected pronouns/determiners in the same combined pronoun/determiner class that we use for Old English and Old Norse. In our view, the advantage of maintaining a common analysis for all the languages and the advantage of keeping the demonstratives in one group outweigh the problem of a tagging of three pronouns that is strictly speaking not correct. The non-inflected pronouns are retrievable because they are morphologically marked as uninflected.

4 Syntactic annotation

4.1 Some aspects of the syntactic model

Our syntactic model is based on dependency grammar (DG), as laid out by TESNIÈRE (1959).[3] The main idea of DG is that syntactic structure consists of lexical elements which are linked by asymmetrical relations called dependencies (see NIVRE 2005, chapter 3, for a nice overview of DG). Thus, DG, unlike phrase structure grammar, does not operate with phrasal nodes, as it is the relation between the head and its dependent(s) that determines

what the structure is. A dependent does not have to be adjacent to its head; dependency relations have to do with structural order rather than linear order. In our annotation application, word order is not modelled, but word order is retrievable since each token is indexed with a number showing its position in the sentence. Thus, we can search for features including word order, e.g. "all initial adverbial prepositional phrases followed by the verb and the subject". In the annotation of older languages (and certainly some modern languages as well), it is an advantage not to be constrained by word order, since word order was freer, i.e. used to mark information structure relations rather than grammatical relations, and discontinuous dependency frequently occurred. The tree diagram below (Figure 1) represents the syntactic annotation of the Old English sentence in (3).

(3) ac he ne mæg for scame in gan buton scrude (*Apollonius of Tyre*, 11th c.)
but he not may for shame in go without clothing
'but for shame, he may not enter without clothing'

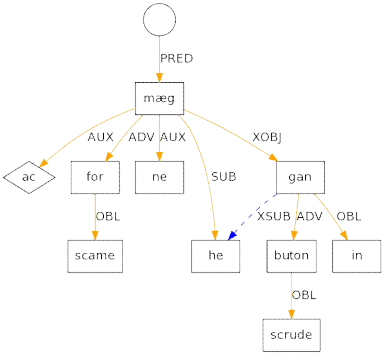


Figure 1: the dependency tree of example (3).

Three features of the tree structure should be noted. The first is the lack of empty nodes; each node corresponds to a word. It happens, however, that we have to insert an empty node, i.e., we may have to insert an empty conjunction in cases of asyndetic conjunction, or an empty verbal node, and the application is designed for and allows such cases. The second feature is the XOBJ relation, which may be unfamiliar from a traditional grammar point of view. This relation is used for predications that have external subjects (they get the subject via coreference relations within the sentence), and are governed by another verb. Non-finite verbs are typically involved, but nouns and adjectives may also be in an XOBJ relation to the matrix verb, in predicative clauses. The XOBJ elements have to be selected and demanded by the matrix verb and are thus arguments. There is also another relation (not shown in Figure 1) which involves adverbials with an external subject: XADV. This relation is typically found with present participles, as in the Old English sentence in (4), where *bysmriende* 'mocking' is XADV to *cwædon* 'said' and has the external subject *heahsacerdas* 'chief priests', which it shares with the matrix verb.

(4) Eallswa þa heahsacerdas bysmriende betwux þam bocerum cwædon
(*Old English Gospels*, 11th c.)
likewise the chiefpriests mocking between the scribes said
'Likewise also the chief priests mocking said among themselves with the scribes'
(*Authorized King James Version*)

This brings us to the third feature: The syntactic model we have adopted differs markedly from DG in one particular respect, namely in allowing structure-sharing, similar to Lexical-Functional Grammar. In DG, a word can only have one head, but this principle leads to problems in the treatment of non-finite verbs in particular. In (3), *he* 'he' is the subject of both *mæg* 'may' and *gan* 'go'. In our graph, this is represented with the slashed arrow, which shows that *he* is the external subject, XSUB, to *gan*. We also use these secondary edges to show other types of shared arguments, most notably in the very frequent case of an ellipted subject in coordinate constructions: *He came, saw and conquered*, where *saw* and *conquered* would both slash to the subject *he*.

4.2 Some syntax annotation challenges

4.2.1 Complex verb phrases

One conflict between the languages in our corpus relates to the status of auxiliary[4] verbs. While the old Germanic languages have a rather limited set of auxiliaries, the Romance languages have a much larger group of verbs that may function as auxiliaries in verbal periphrases. For all the languages, we analyze auxiliary verbs that are used to mark tense as an AUX[5] element attached to the head (and not the other way around) and attach the arguments directly onto the non-finite verb, as in the Old Norse example in (5). Furthermore, Old Norse combines modality and tense, using modal verbs to form futures and conditionals. Consequently, temporal and modal verbs are therefore not distinguished, since disambiguation is not possible. Old Norse therefore has the same annotation for the tense-marking verb *hafa* 'to have' (5) and modal verbs, e.g. *vilia* 'to wish to' in (6).

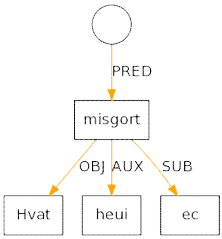


Figure 2: the dependency tree of example (5).

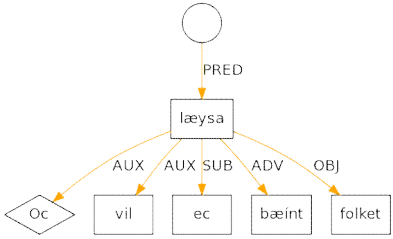


Figure 3: the dependency tree of example (6).

(5) Hvat heui ec misgort (*Strengleikar*, 13th c.)
what have I misdone
'What have I done wrong?'

(6) Oc vil eg beínt læysa folket (*Óláfs saga hins helga*, 13th c.)
and will I straightaway release people.DEF.ART
'And I will straightaway release the people'

In Old English, on the other hand, examples equivalent to the Old Norse example in (6) receive a different analysis, i.e. an analysis with XOBJ (see example (3)), since it can be argued that the verbs that today are considered as modal auxiliaries still had at least some of their lexical force in Old English (TRAUGOTT 1992). As regards the syntactic annotation, then, the Old Norse structure is flat, whereas the analysis of Old English is biclausal.

We analyze the Romance tense and modal auxiliaries in the same way as their Old English counterparts, but the Romance languages also have a range of more or less grammaticalized aspectual auxiliaries that combine with the main verb, and are often preceded by a preposition. These aspectual auxiliaries have given us some headache. There is variation as to what extent these aspectual auxiliaries have been grammaticalized, from the almost completely grammaticalized verb (in Portuguese) *estar* in *estar fazendo* or *estar a fazer* 'to be doing something', through semi-grammaticalized verbs such as *passar* in *passar a fazer* 'to pass on to doing something (else)' to non-grammaticalized finite verbs like *começar* in *começar por fazer* 'start by doing'. In the case of *estar*, which is similar to its English counterpart *be* in progressive constructions, it is relatively easy to argue that the verb is no more than an auxiliary in periphrastic constructions; it has no semantic content other than tense and aspect. With the other verbs, however, we are dealing with a grammaticalization continuum, so if we were to analyze the periphrases according to their grammaticalization status, with the finite verb either as an AUX or as a head, the choice would in most cases be rather arbitrary. We therefore chose to use the same analysis for all these periphrases, regardless of grammaticalization status. This means that we annotate both the subject and the non-finite verb as arguments to the finite verb, and indicate the subject of the infinitive through a slash annotation as described in 4.1.

Our analysis of modals has has consequences for retrievability later on. In Old Norse, modals are analyzed in the same way as other auxiliaries, whereas in our other languages, modals are analyzed as heads in an XOBJ relation, as we do with other verbal periphrases. Hence, the modals will only be retrievable through a search on language-specific lemmas or through relatively complicated syntactic and morphological search strings. A search asking for a particular syntactic structure, e.g. "find a finite verb with an XOBJ that is an infinitive" would render not only the modal verbs, but also other periphrases with the same structure, e.g. the Romance aspectual verbal periphrases described above, or the Old English causative constructions with an accusative + infinitive described in 4.2.2. Also, whether we search for lemmas or structures, the search strings will have to be language-specific. In other words, though some serious thinking will be required in the search stage, the annotation stage is kept simple and manageable for the annotators.

4.2.2 Extended Case Marking and other infinitival clauses

In section 4.1, we described the XOBJ relation where the finite and the non-finite verb share a subject, and this is tagged through slash annotation. The subject of the non-finite verb does not have to be the same as that of the finite verb, but it must be external for a relation to be analyzed as XOBJ; if it is internal we analyze the sentence as a complementizer clause through the relation COMP.

Accusative with infinitive (AcI) constructions are typically COMP, since the accusative + infinitive unit is regarded as an argument of the finite verb, and the subject gets its semantic (theta-) role from the non-finite verb. Example 7 is a case in point, with the syntactic representation shown in Figure 4.

(7) oc hævir hann hœyrt [fugla syngia]... (*Strengleikar*, 11th c.)
and has he heard birds.acc sing.inf...
'and he has heard birds sing...'

(8) oc bauð [þeim] [at biða sin] (*Strengleikar*, 13th c.)
and bid them.dat to wait her.refl
'and bid them wait for her'

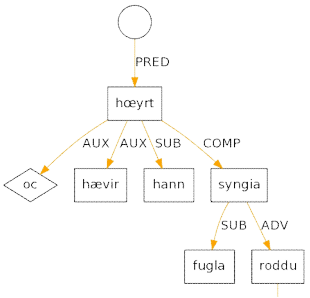


Figure 4: the dependency tree of the accusative with infinitive in example (7).

In Old Norse, there is a distinction between accusative with infinitive after verbs of perception (example 7), and dative with infinitive after verbs of causing and commanding such as 'to order', 'to bid', 'to let', 'to make' (someone (to) do something) (example 8). The latter is analyzed as an XOBJ construction in which the dative is an argument of the finite verb and is tagged as an external subject to the non-finite verb through a slash annotation.

In Old English, both perception verbs and verbs of causing and commanding are followed by accusative + infinitive. It is therefore difficult to distinguish between the constructions; i.e. which ones are true AcI constructions in which the AcI as a unit is dependent on the finite verb, and which ones are constructions in which the accusative and the infinitive are separate arguments of the finite verb. MITCHELL (1985 II) discusses AcI extensively, and makes repeated use of the 'jungle' metaphor to describe the difficulty of the task. In any case, we have decided to analyze Old English in the same way as Old Norse, with perception verbs taking AcI, and thus COMP, and causative verbs and verbs of commanding taking an XOBJ construction; in other words, the accusative element is in the latter case analyzed as an argument of the finite verb. The motivation for this is the existence of various constructions for a verb such as *hatan* 'to order, command, bid', such as the ones exemplified in (9) and (10). Especially the latter construction suggests that the accusative element is in fact an argument of the finite verb.

(9) and het [hine] [in gan] (*Apollonius of Tyre*, 11th c.)
and bade him.ACC in go
'and bade him enter'

(10) þa het ic [eallne þone here] [þæt he to swæsendum sæte]
(*Letter of Alexander the Great to Aristotle*, 10th-11th c.,
from MITCHELL 1985 vol. II: 871)
then bid I all.ACC the.ACC army.ACC that it to meal sit.SBJV
'then I bid the entire army to be seated for the meal'

As regards the Romance languages, the distinction between the COMP and XOBJ construction is not easily made, because the dative/accusative distinction is less clear. We have therefore chosen to analyze both constructions with perception verbs and constructions with causative verbs as COMP constructions. At first sight, a consistent COMP analysis may seem like the poorer alternative, because the syntactic relation between the finite verb and the subject of the infinitive is lost with this annotation. There are, however, other reasons for opting for the COMP analysis.

While in most of our languages, there is a distinction between accusative and dative subjects of the infinitive, these constructions appear in three different structures in Portuguese (MARTINS 2006): non-inflected infinitive with accusative subject (Extended Case-Marking (ECM) construction), as in (11), inflected infinitive with nominative subject, as in (12), and non-inflected infinitive with dative subject, as in (13). All three sentences mean 'I ordered the students to make the cake'.[6]

(11) Mandei os alunos fazer o bolo
I ordered the students make.INF the cake

(12) Mandei os alunos fazerem o bolo
I ordered the students make.INF.3PL the cake

(13) Mandei aos alunos fazer o bolo
I ordered to-the students make.INF the cake

The example in (12) does not have the structural ambiguity of the two others, because there is agreement between the (nominative) subject of the infinitive and the infinitive itself. Consequently, (12) must be analyzed as [mandei [os alunos fazerem o bolo]], with the subject case-marked by the infinitive and not by the finite verb (MARTINS 2006). For (11) and in particular for (13) there are two possible analyses: [mandei [(a)os alunos] [fazer o bolo]] and [mandei [(a)os alunos fazer o bolo]]. The easiest way to capture this structural ambiguity, would be to use an XOBJ relation for the infinitive, annotate the subject of the infinitive as an argument to the finite verb and use a slash to indicate the subject relation. Examples (11) and (13) would then be distinguished from (12), where the infinitive clause is analyzed as COMP. The main reason why we have chosen not to do so is that the third person *singular* form of the inflected infinitive is formally identical to the non-inflected form. In the singular, both (11) and (12) would be *mandei o aluno fazer o bolo* 'I asked the student to make the cake'. Because the third person singular is the most common form, there will be a large

number of ambiguous contexts where it is impossible to distinguish between a nominative + inflected infinitive and an accusative with infinitive.

A disadvantage of the consistent COMP analysis for Portuguese is the irretrievable status of the accusative or dative NP as object/indirect object to the finite verb, once the NP is analyzed as a subject of the infinitive. One consequence of this is that when we carry out the information structure analysis (as described in section 5), the information structure annotation on these arguments will be retrievable as information structure on subjects, not on objects. For example, if we search for "main clause objects that constitute given information", none of the AcI NPs will be included in the result. On the other hand, this can also be an advantage, since we will get fewer examples that may have to be removed anyway because of their ambiguous nature. Another advantage of the COMP analysis is that these constructions are the only COMPs with an infinitive as head. A simple search asking for a "COMP with a head that is infinitive" will render all and only these structures. Given their particular ambiguous nature, in Portuguese at least, it is a good idea to keep them apart from other structures. In addition, there is the additional advantage of keeping the annotation work simple and economical.

The fact that the subcorpora of different languages are annotated in basically the same way facilitates searches to a certain extent. However, in sections 3 and 4 we have seen that one and the same phenomenon may receive different analyses in different languages due to language-internal classification problems, and thus language-specific search strings are required. The information structure annotation, on the other hand, is not language-specific, as the categories are not tied to language-specific properties such as e.g. word order or definiteness. Rather, the scheme for information structure annotation is a method of textual analysis that may be applied to texts in any language, without the language-specific rules needed for morphology and syntax, as will be shown in section 5.

5 Information structure annotation

5.1 The information structure annotation scheme

The information structure annotation aims to represent the writer's assumptions about what is in the mind of the addressee (cf. e.g. CHAFE 1976, PRINCE 1981), and we are interested in how these assumptions contribute to linguistic structure, in particular word order. We use the PROIEL annotation scheme (cf. PROIEL guidelines,[7] and also NISSIM ET AL. 2004, and RIESTER ET AL. 2010) and annotate noun phrases only, distinguishing between old, new and accessible information on one level, and specific, non-specific and generic information on another.

An element is tagged as OLD information if it is co-referential with an element in the preceding discourse. If the old element is outside the limit of the annotation window (13 sentences), it is tagged as OLD-INACTIVE. An element is tagged as NEW if it is mentioned for the first time. The category Accessible subsumes three types of accessible information, namely situational (ACC-SIT), inferable (ACC-INF) and general (ACC-GEN) information. ACC-SIT is used for referents that are available in the discourse situation (mostly relevant in direct speech contexts), ACC-INF is used when the referent is inferable from the preceding discourse, and ACC-GEN refers to world knowledge, i.e. what we think was the world knowledge of the readers of the texts at the time when they were produced. OLD and ACC-INF elements must be textually licensed; they receive an anaphoric link which points back to the licensing element.

The OLD, ACC and NEW categories were the original ones in the annotation scheme, and they all have specific reference. However, the need arose to tag specificity vs. non-specificity as well, and thus the economy–disambiguation scale had to be tilted somewhat in favour of disambiguation. Annotators now have more categories to consider, but the advantage is a more satisfactory and, it is hoped, useful end product. Non-specific referents are basically divided into quantifier restrictions, QUANT, and all other types of non-specific referents. Furthermore, non-specific referents can be old or inferred. In the sentence *All people have nails and have to cut them*, *people* is quantified and receives the QUANT tag, *nails* is NON-SPEC and *them* is NON-SPEC-OLD (cf. PROIEL guidelines for a discussion of these categories and some other special cases). There is also a tag, KIND, for generic expressions, such as *lions* in *Lions are dangerous*.

Before we look at some examples of actual annotation, it should be mentioned that we also annotate NPs which constitute the complement of a preposition, e.g. *that house* in the prepositional phrase *in that house*. In addition, we annotate pro-dropped arguments, which are most commonly subjects, but pro-dropped objects and obliques also occur. Thus, in the information structure layer of the annotation application, we insert a marker for pro. Elements can refer back to pro, and pros can thus be elements in an anaphoric chain.

In the information structure annotation, we assume that certain textual properties, such as givenness, and certain semantic properties, such as specificity, are related to information structure properties such as topic, background and focus, and that the annotation will enable us to detect the information structure system in a given language.

5.2 Some information structure annotation challenges

In this section, we focus on two annotational issues that tend to cause a certain degree of frustration for the annotators: the Accessible category, in particular how to recognize inferables (ACC-INF), and complex noun phrases, i.e. noun phrases that contain more than one NP. One of the main questions as regards inferables is what we can assume was inferable for the intended readers of the text; it may very well be that contemporary readers of the old text inferred different things from the discourse than we do today. What makes this task a little less difficult is that the text guides us; the inferred elements have to be textually licensed from some element in the text, though not necessarily a co-referential element. As a warm-up exercise, let us first consider the entire information structure annotation of (14), which is the first verse of chapter 15 in the Old English Gospel of Mark. To provide some context, the last verse of the previous chapter is given (in Modern English (*King James*) for the sake of simplicity). The annotatable elements are marked in grey.

(14) [And the second time the cock crew (OE: *þa eft sona creow se hana*). And Peter called to mind the word that Jesus said unto him, Before the cock crow twice, thou shalt deny me thrice. And when he thought thereon, he wept.]
þa sona on mergen worhton þa heahsacerdas hyra gemot mid ealdrum and bocerum and eallum werodum and PRO-SUB læddon þone hælend gebundenne and PRO-SUB sealdon him Pilato. (*Old English Gospels*, 11th c.)

then immediately in morning held the chiefpriests their council with elders and scribes and all council and led the saviour bound and delivered him Pilate.DAT
'then straightway in the morning the chief priests held a consultation with the elders and scribes and the whole council, and bound Jesus and delivered him to Pilate'
(*King James Authorized Version*)

First, the prepositional complement *mergen* is tagged as ACC-INF. This is because in the last verse of the previous chapter, it is mentioned that the cock crowed: *þa eft sona creow se hana*. Consequently, the morning can be inferred, since the crowing of the cock signals morning; hence the anaphoric link goes from *mergen* back to the verb *creow*. This is a relatively straightforward example of an inferable relation. *Heahsacerdas*, *ealdrum*, *bocerum*, and *werodum* are all tagged as OLD-INACTIVE. They were mentioned in the previous chapter, so the information is old, but outside the annotation window of 13 sentences. *Hyra gemot* is a complex NP consisting of a possessive pronoun and a noun. The pronoun is tagged as OLD, with an anaphoric link back to *heahsacerdas* and *gemot* is tagged as NEW. Then we have a new main clause without an expressed subject, and in the information structure annotation, we insert a PRO-SUB before the verb *læddon*. The question is what tag this pro element should get. Is it old, i.e. coreferent with *heahsacerdas*, meaning that the chief priests led Jesus, or should it rather be interpreted as inferable, ACC-INF, from the chief priests, since it is probable that the priests had some minions to do the binding and leading for them? Here we choose the former solution, since we do not want to overinterpret the text. The next taggable element is *hælend*, which is OLD information here, because it links back to *me* in the previous chapter. Then we get another PRO-SUB, which is tagged as OLD, linking back to the previous PRO-SUB. *Him* is also OLD, with a link to *hælend*, and finally *Pilato* is tagged as ACC-GEN, since we assume that Pilate was generally known to the readers of the Bible.
With the exception of the pro element mentioned, the annotation of this sequence is straightforward. In (15), however, things get a little more complicated.

(15) and hi tosomne eall werod clypedon ... and beoton hine on þæt heofod mid hreode and spætton him on and heora cneow bigdon. (*Old English Gospels*, 11th c.)
and they together all band called ... and beat him on the head with stick and spit him on and their knees bent
'and they call together the whole band ... and beat him on the head with a stick and spit on him and bent their knees' (*King James Authorized Version*)

There are several annotatable elements here, but we will focus on *þæt heofod* and *heora cneow*, both NPs consisting of a head and a dependent. In *þæt heofod* we annotate *heofod* as ACC-INF, since the inference here is from humans to body parts; we can infer that people have heads. The demonstrative pronoun *þæt*, used in an article-like way here, is not referential, and thus not annotated. *Heora cneow* also refers to a body part, but this NP consists of the possessive determiner *heora*, which is referential, and the noun *cneow*. Here, the referent can be identified by means of an element within the noun phrase, namely *heora*, which gets the OLD tag, and *cneow* is therefore annotated as NEW, even though it, like *heofod* is inferable. The point is to distinguish between NPs which contain the element we need to identify the referent, and NPs that do not contain any such element. In *heora cneow*, we get that information through *heora*, which refers back to *hi* 'they', and hence *cneow* is tagged as

NEW, whereas in *þæt heofod*, we do not get that information, and hence the head gets the ACC-INF tag.

Another example is found in (16). Again there are several annotatable elements, but we will focus on two of them.

(16) Þa wæs underntid and hie ahengon hine. And ofergewrit his gyltes wæs awriten, Iudea cyning (*Old English Gospels*, 11th c.)
Then was third-hour and they crucified him. And superscription his.GEN accusation.GEN was written, Jews.GEN king
'And it was the third hour, and they crucified him. And the superscription of his accusation was written over, THE KING OF THE JEWS'
(*King James Authorized Version*)

Here, *ofergewrit* 'superscription' is tagged as ACC-INF, with a link back to *ahengon* 'crucified', the inference being that superscriptions saying what the crime was were sometimes attached to the cross, and the readers would know this. *Gyltes* is also inferable, but this NP also contains a possessive determiner *his*, which identifies the referent; the link is back to the pronoun *hine*. *Gyltes* is therefore tagged as NEW; in other words, this NP is tagged in the same way as *heora cneow* in (16).

This type of annotation causes some problems for the annotators, because it is easy to overlook such inferences, it may be difficult to know which inferences are valid, and the tagging with old and new within the same NP (as in *his gyltes*) is not particularly intuitive. In other words, here economy has been sacrificed on the altar of disambiguation.

6 Summary

In this paper we have presented some aspects of our work on a corpus consisting of several languages in their older stages. We have given a brief description of the annotation scheme, focusing on syntax and information structure, and discussed some of the problems we come across in the actual annotation, with particular reference to the challenge of balancing out the principles of economy and disambiguation.

In the end, when we get to the search and research stage, we will be able to combine the annotated data of the three levels of morphology, syntax and information structure, and trace not only what types of arguments appear in which linear positions, but also the distribution of information structural properties within and between sentences, the research aim being to study the relation between information structure and word order change.

[1] *http://www.hf.uio.no/ilos/english/research/projects/iswoc/index.html*.

[2] We do not annotate German ourselves, since much work has been done on Old High German by our partners at the Humboldt University, Berlin, in the project *Informationsstruktur und Wortstellung im Germanischen*: *http://www2.hu-berlin.de/sprachgeschichte/forschung/informationsstruktur/index.php*.

[3] The model and application we use were developed by the PROIEL project at the University of Oslo, and we have adapted it for our languages. The PROIEL annotation guidelines can be found at: *http://folk.uio.no/daghaug/syntactic_guidelines.pdf*.

[4] We use the term 'auxiliary verb' here simply to refer to more or less grammaticalized verbs that modify the main verb in complex verb phrases. There is a gradience of auxiliary-hood among the verbs discussed here.

[5] AUX is the tag we use for grammatical words, i.e words that give additional information about their head. The tag is not only used for auxiliary verbs, but also marks modal particles, focus particles, negation, and coordinating conjunctions.

[6] The case of the infinitive subject is perhaps easier to see if we substitute *(a)os alunos* with pronouns: (11') mandei-os.ACC fazer o bolo; (12') mandei eles.NOM fazerem o bolo; (13') mandei-lhes.DAT fazer o bolo.

[7] *http://folk.uio.no/daghaug/info_guidelines.pdf*.

References

CHAFE, W. (1976). "Givenness, contrastiveness, definiteness, subjects, topics, and point of view." In LI, C. (ed.). Subject and Topic. New York: Academic Press, 25-55.

NIVRE, J. (2005). Inductive Dependency Parsing. Dordrecht: Springer.

HAUG, D.T.T, JØHNDAL, M.L., ECKHOFF, H.M., WELO, E, HERTZENBERG, M.J.B., MÜTH, A. (2009). "Computational and linguistic issues in designing a syntactically annotated parallel corpus of Indo-European languages." In Traitement Automatique des Langues, vol. 50, 17-45. Retrievable at *http://www.atala.org/Computational-and-Linguistic*.

MARTINS, A.M. (2006). "Aspects of infinitival constructions in the history of Portuguese." In GESS, R.S. & D. ARTEAGA (eds.). Historical Romance Linguistics: Retrospective and Perspectives. Amsterdam & Philadelphia: John Benjamins, 327-355.

MITCHELL, B. (1985). Old English Syntax, vol. II. Oxford: Clarendon Press.

NISSIM, M., DINGARE, S., CARLETTA, J, STEEDMAN, M. (2004). "An annotation scheme for information status in dialogue." In LINO M.T ET AL. (eds.) Proceedings of the 4th Interna tional Conference on Language Resources and Evaluation, Lisbon 2004. Retrievable at *http://homepages.inf.ed.ac.uk/jeanc/nissim-lrec2004.pdf*.

PRINCE, E. (1981). "Toward a taxonomy of given-new information." In COLE, P. (ed.). Radical Pragmatics. New York: Academic Press, 223-255.

RIESTER, A., LORENZ, D, SEEMANN, N. (2010). "A recursive annotation scheme for referential information status." In CALZOLARI, N. ET AL. (eds.). Proceedings of the 7th International Conference on Language Resources and Evaluation, Valletta 2010, 717-722. Retrievable at: *http://www.lrec-conf.org/proceedings/lrec2010/index.html*.

TESNIÈRE, L. (1959). Eléments de syntaxe structurale. Editions Klincksieck.

TRAUGOTT, E. C. (1992). "Syntax." In HOGG, R. (ed.) The Cambridge History of the English Language, volume I. Cambridge: Cambridge University Press, 168-289.

Stefanie Dipper

Morphological and Part-of-Speech Tagging of Historical Language Data: A Comparison

This paper deals with morphological and part-of-speech tagging applied to manuscripts written in Middle High German. I present the results of a set of experiments that involve different levels of token normalization and dialect-specific subcorpora. As expected, tagging with "normalized", quasi-standardized tokens performs best. Normalization improves accuracies by 3.56–7.10 percentage points, resulting in accuracies of $> 79\%$ for morphological tagging, and $> 91\%$ for part-of-speech tagging. Comparing Middle with New High German data of similar size, the evaluation shows that part-of-speech tagging, but not morphological tagging, is clearly easier with modern data.

1 Introduction[1]

This paper deals with automatic analysis of historical language data, namely morphological and part-of-speech (POS) tagging of texts from Middle High German (1050–1350). Analysis of historical languages differs from that of modern languages in two important points. First, there are no agreed-upon, standardized writing conventions. Instead, characters and symbols used by the writer of some manuscript in parts reflect impacts as different as spatial constraints (parchment is expensive and, hence, use of abbreviations seems favorable) or dialect influences (the dialect spoken by the author of the text, or the writer's dialect, who writes up or copies the text, or even the dialect spoken by the expected readership). This often leads to inconsistent spellings, even within one text written up by one writer. Second, resources of historical languages are scarce and often not very voluminous, and manuscripts are frequently incomplete or damaged.

These features—data variance and lack of large resources—challenge many statistical analysis tools, whose quality usually depend on the availability of large training samples. Automatic taggers have been used mainly for the annotation of English historical corpora. The "Penn-Helsinki Parsed Corpora of Historical English" (Kroch and Taylor, 2000; Kroch et al., 2004) have been annotated with POS tags in a bootstrapping approach, which involves successive cycles of manual annotation, training, automatic tagging, followed by manual corrections, etc. Rayson et al. (2007) and Pilz et al. (2006) map historical word forms to the corresponding modern word forms, and analyze these by state-of-the-art POS taggers. The mappings make use of the Soundex algorithm,

[1] I would like to thank the anonymous reviewers for helpful comments. The research reported here was supported by Deutsche Forschungsgemeinschaft (DFG), Grant DI 1558/1-1.

Edit Distance, or heuristic rules. Rayson et al. (2007) apply this technique for POS tagging, Pilz et al. (2006) for a search engine for texts without standardized spelling.

Morphological tagging has received far less attention than POS tagging, presumably because English, which is the most researched language in computational linguistics, does not have rich morphology, and, furthermore, a considerable amount of (overtly marked) morphological information is in fact recorded by common English POS tagsets, e.g. for nouns: singular vs. plural form, for verbs: uninflected base form vs. third-singular present tense vs. past tense vs. participle, etc. Similar coarse-grained distinctions have been transferred to languages with rich(er) morphology, such as German. For instance, in the de-facto standard tagset for modern German corpora, the STTS (Schiller et al., 1999), all finite verb forms receive the tag VVFIN ("full verb, finite"), infinitives the tag VVINF ("full verb, infinitive"), etc. However, in contrast to English, the tag VVFIN covers up to five differently-inflected verb forms; similarly, the tag NN ("common noun"[2]) also corresponds to up to five different forms. Hence, full morphological tagging, which would differentiate between the different forms, could provide valuable information in languages with rather free word order: morphological information can help in determining constituents and grammatical functions. POS and morphological tagging thus represents important preprocessing steps, e.g., for treebanking or natural language processing of such languages.

This paper reports on experiments in applying a state-of-the-art tagger, the TreeTagger (Schmid, 1994), to a corpus of texts from Middle High German (MHG).[3] The tagger is used for both morphological and POS tagging. My approach is similar to the one by Kroch et al. in that I train and apply the tagger to historical rather than modern word forms. The tagging experiments make use of a balanced MHG corpus that is created and annotated in the context of two projects, the projects "Mittelhochdeutsche Grammatik" and "Referenzkorpus Mittelhochdeutsch".[4] The corpus has been semi-automatically annotated with morphology, POS tags, lemma, and a normalized word form, which represents a virtual historical standardized form. The corpus is not annotated with modern word forms.

I present the results of a set of experiments that involve different types of tokens (original and normalized versions) and dialect-specific subcorpora. Sec. 2 gives detailed information about the corpus and its annotations, Sec. 3 addresses the tagging experiments and results. In many places, I contrast the historical data with a modern corpus, the TIGER corpus (Brants et al., 2004). Sec. 4 presents a summary.

[2] Tags for nouns in German tagsets are usually unspecified for number.

[3] In a recent evaluation of part-of-speech taggers on German web data, Giesbrecht and Evert (2009) found that the Stanford tagger (Toutanova et al., 2003) performed best (97.63%) while the TreeTagger only achieved an accuracy of 96.89%. On the other hand, training the taggers took 10 seconds (TreeTagger) vs. 5.5 hours (Stanford). Another important advantage of the TreeTagger is the fact that its model can be inspected and easily interpreted (the options "-print-prob-tree" and "-print-suffix-tree" print out the decision tree for ngrams and the suffix lexicon, respectively). Moreover, training the TreeTagger is straightforward and does not require any specific preprocessing, in contrast, e.g., to the RFTagger (Schmid and Laws, 2008), which presupposes the definition of a finite-state automaton for the tag labels.

[4] `http://www.mittelhochdeutsche-grammatik.de` and `http://www.linguistics.rub.de/mhd/`.

DIPL	ich	dir	gelobe	.	dar	zů	ne	helbē	ich	dir
NORM	ich	dir	gelobe	.	dar	zuo	ne	hilfen	ich	dir
LEMMA	ich	dû	ge-loben		dâr	zuo	ne	hëlfen	ich	dû
MORPH	*.Nom.Sg	*.Dat.Sg	1.Sg.Pres.*	–	–	–	–	1.Sg.Pres.Ind	*.Nom.Sg	*.Dat.Sg
POS	PPER	PPER	VVFIN	$.	ADV	ADV	NEG	VVFIN	PPER	PPER
GLOSS	I	you	promise		there	to	not	help	I	you

Figure 1: A line from Eilhart's *Tristrant* (Magdeburg fragment), along with a diplomatic transcription, normalized word forms, and linguistic annotations. The complete sentence is: *vil ernirsthafte ich dir gelobe. dar zuo ne helben ich dir niet* 'Very seriously I promise you: I do not help you with this'.

2 The Corpus

The corpus is a collection of texts from the 12th–14th centuries, including religious as well as profane texts, prose and verse. The texts have been selected in a way as to cover the period of MHG as optimally as possible. The texts distribute in time, i.e. over the relevant centuries, and in space, coming from a variety of Central German (CG) and Upper German (UG) dialects. CG dialects were spoken in the central part of Germany; examples are Franconian or Thuringian. UG dialects were (and are still) spoken in Southern Germany, Switzerland, and Austria, e.g. Swabian, Alemannic, or Bavarian.

The corpus provides two different versions of "word forms": the diplomatic transcription and a normalized form. Figure 1 presents an example fragment encoded in the different versions.[5] Below the lines with the word forms, linguistic annotations are displayed: lemma, morphology, parts of speech (POS).

Lines DIPL and NORM The texts are *diplomatic* transcriptions, i.e., they aim at reproducing a large range of features of the original manuscript or print, such as large initials, superscribed letters (e.g. ů), variant letter forms (e.g. short vs. long s: <s> vs. <ſ>), or abbreviations (e.g., the superscribed "nasal bar" <‾> substitutes *n*).[6]

[5] The manuscript screenshot has been taken from `http://www.hs-augsburg.de/~harsch/germanica/Chronologie/12Jh/Eilhart/eil_tmma.html`

[6] Internally, I use an isomorphic ASCII-encoded representation of the diplomatic transcription. Instead of letters with diacritics or superposed characters (*ö*, ů), it uses ASCII characters combined with the backslash as an escape character (*o\"*, *u\o*). Ligatures (*æ*) are marked by an underscore (*a_e*), *&* is mapped to *e_t*, *þ* to *t_h*.

Corpus Dialect (#Texts)	Tokens	Types and TTR *dipl*	*norm*
total (51)	211,000	40,500 .19	20,500 .10
CG (27)	91,000	22,000 .24	13,000 .14
UG (20)	67,000	15,000 .22	8,500 .13
mixed (4)	53,000		

Corpus	Tokens	Types and TTR
TIGER	1,000,000	81,000 .09
	210,000	30,000 .14
	90,000	16,000 .18

Table 1: Number of tokens and types in the Middle High German corpus (left) and in differently-sized subcorpora of the TIGER corpus (right). Below each type figure, the type-token ratio (TTR) is given.

The *normalized* version is an artificial standard form, similar to the citation forms used in lexicons of MHG, such as Lexer (1872).[7] The normalized form abstracts away completely from dialectal sound (grapheme) variance. It has been semi-automatically generated by a tool developed by Thomas Klein (Klein, 2001) within the project "Mittelhochdeutsche Grammatik". The tool exploits lemma and morphological information in combination with symbolic rules that encode linguistic knowledge about historical dialects. The user has to provide information about the dialect of the text, and to correct intermediate results interactively. No information about overall accuracy or inter-annotator agreement is available.

Table 1 displays some statistics of the current state of the corpus (left table). The first column shows that there are currently 51 texts in total, with a total of around 211,000 tokens. The shortest text contains only 51 tokens, the longest one 25,000 tokens. 27 texts are from CG dialects and 20 from UG dialects. 4 texts are classified as "mixed", because they show mixed dialectal features, or are composed of fragments of different dialects. Due to their nature, the mixed texts have been excluded from detailed consideration.

The table shows that the numbers of types are considerably reduced if diplomatic word forms are mapped to normalized forms. This benefits current taggers, as it reduces the problem of data sparseness to some extent. The question is, however, how reliably the normalized form can be generated automatically. The current tool requires a considerable amount of manual intervention during the analysis.

[7] Internally, I use a simplified ASCII version of the normalized form, with the following modifications: Umlaut has been replaced by the corresponding voyel + e (e.g. "ä" becomes "ae"); other accents or diacritics have been removed.

CG texts seem more diverse than UG texts: despite the fact that the CG subcorpus is larger than the UG subcorpus, it has a higher type/token ratio (TTR). Usually longer texts tend to have lower TTR values. This is shown by the right table of Table 1: The entire TIGER corpus (1,000,000 tokens) has a TTR of .09, i.e., there are 11.1 corpus instances of each word (type) on average. Taking into account only the first 210,000 tokens of the TIGER corpus, TTR goes up to .14; this corresponds to 7.1 instances of each word on average. The TTR of the 90,000 TIGER subcorpus, which is comparable in size with the CG subcorpus, shows that New High German (NHG, i.e. newspaper texts from the 1990s) has a more diverse vocabulary than the MHG texts.

Judging from these figures, one could predict the following outcomes:[8]

1. Normalized vs. diplomatic: Tagging **normalized** data should be easier
2. CG vs. UG vs. NHG data:
   a) Tagging **CG** should be easier that UG, because more training data is available
   b) Alternatively: tagging **UG** is easier than CG, because it is less diverse (has a lower TTR)
   c) Tagging (equally-sized subsets of) **MHG** should be easier than NHG, because it has lower TTRs

Line MORPH In addition to normalized word forms, the texts have also been annotated with morphological and part-of-spech (POS) tags, by the tool by Klein (2001). The original morphological tagset consists of around 430 tags. The large number of tags is partly due to the fact that inherent gender of nouns was not yet as fixed as it is nowadays. That is, many nouns could be used, e.g., with masculine or feminine articles (or with all three genders). In all cases where the context does not allow for gender disambiguation, ambiguous tags have been annotated, as in Ex. (1). "MascFem.Nom.Pl" means nominativ plural, masculine or feminine. "*" means that a feature is entirely underspecified, such as gender with the plural pronoun *sie* 'them', which is therefore tagged as "*.Acc.Pl".

(1)

daz	si	slangen	bizzen
—	*.Acc.Pl	MascFem.Nom.Pl	3.Pl.Past.*
that	them	snakes	bit

'that snakes bit them'

Moreover, properties such as postnominal position, e.g., of adjectives or possessive determiners, or morphological unmarkedness, have also been recorded by the original tagset. For the experiments described in this paper, these morphology tags were mapped automatically to a slightly modified version of the STTS morphological tagset. (If the value of a specific slot could not be determined automatically, it was also filled by "*".)

[8] Of course, the outcomes also depend on properties of the tagsets, see below.

Corpus	Morphology			Part of Speech		
	#Tags	Tags/Word	x̃ (max)	#Tags	Tags/Word	x̃ (max)
CG *norm*	245	Ø1.40 ± 1.16	1 (23)	44	Ø1.10 ± 0.37	1 (7)
UG *norm*	219	Ø1.46 ± 1.28	1 (33)	41	Ø1.10 ± 0.35	1 (6)
TIGER						
1,000 K	270	Ø1.48 ± 1.22	1 (40)	54	Ø1.05 ± 0.25	1 (7)
210 K	230	Ø1.37 ± 0.97	1 (26)	53	Ø1.05 ± 0.23	1 (6)
90 K	205	Ø1.32 ± 0.86	1 (18)	51	Ø1.04 ± 0.21	1 (6)

Table 2: Sizes of the tagsets and average number of tags per word (with standard deviation), as occurring in the normalized training data, along with the median (x̃) and maximum.

Line POS The original POS tagset comprises more than 100 tags and, similarly to the morphological tagset, encodes very fine-grained information. For instance, there are 17 different tags for verbs, whose main purpose is to indicate the inflection class that the verb belongs to. For the experiments described in this paper, these POS tags were mapped automatically to a modified version of the STTS POS tagset (for a description of the modifications, see Dipper (2010, Fn.5)).

Table 2 presents relevant statistical information about the resulting STTS-based tagsets. One can see that the sizes of the tagsets are similar with CG, UG, and NHG data. Morphological tagsets are 4–5.5 times larger than POS tagsets. Historical data in general seems more ambiguous than modern data, on average. The figures have to be interpreted with care, though, because the tagsets cannot be directly compared: there is no isomorphic mapping between the information encoded by the original MHG tagsets and the STTS tagsets, and underspecified tags have to be used in the MHG data rather often.

The figures also confirm that the sizes of the corpora are rather small: numbers calculated from the TIGER subcorpora show that adding more data increases the number of tags occurring in the data, especially in the case of morphological tags. That is, even in the complete TIGER corpus, not all available (morphological) tags do occur at least once.[9]

Despite these caveats, we could add the following predictions, based on the figures in Table 2:

3. Morphology vs. POS:
   Tagging of **POS** information should be easier (due to a lower ambiguity rate)

[9] As defined in the header of the TIGER corpus, the total number of morphological STTS tags is 585. Presumably, however, a good amount of them are theoretically possible tags but without any actual instance in the language.

4. CG vs. UG vs. NHG data:
   a) Tagging **NHG** data should be easier (due to a lower ambiguity rate) — this is contrary to the expectation formulated above (see Prediction 2c).
   b) Results for CG and UG should be comparable (almost identical average of ambiguity rates). — The situation here is similar to above: no clear advantage emerges (cf. Predictions 2a and 2b).
   c) However, UG has a higher maximum with ambiguous morphology tags, CG with ambiguous POS tags. Hence, **CG** could perform better with morphological tagging than UG, and **UG** could perform better with POS tagging than CG.

3 Experiments and Results

For the experiments with the historical data, I performed a 10-fold cross-validation. The split was done in blocks of 10 sentences (or "units" of a fixed number of words, if no punctuation marks were available[10]). Within each block, one sentence was randomly extracted and held out for the evaluation.

For the analysis, I used the TreeTagger. It takes suffix information into account so that it can profit from units smaller than words. This seems favorable for data with high variance in spelling. Moreover, the TreeTagger allows the user to inspect the ngram and suffix models acquired during training (see Fn. 3).

In the experiments, I varied two parameters concerning the input data ("dialect, word forms") and one parameter concerning training ("tagger"):

1. Dialect: *CG, UG*

2. Word forms: *dipl, norm*
   For instance, in one setting input data consists of normalized data from Central German (CG-*norm*).

3. Tagger: *gen*(eric), *spec*(ific). In the generic setting, the tagger is trained on the entire corpus (210,000 tokens) and then evaluated on the CG and UG subcorpora. In the specific setting, the tagger is trained and evaluated on the subcorpora only (e.g., the tagger is trained and evaluated on CG-*norm* data). This allows us to evaluate whether a larger set of training data is favorable to a set that is smaller but more homogeneous.

Furthermore, as I have discussed in Sec. 1, POS tags already encode a considerable amount of morphological information. Hence, to improve accuracy with morphological tagging, I also fed the tagger with preprocessed data, which contained POS annotations, so that the morphological tagger could profit from that information.

[10] Punctuation marks in historical texts do not necessarily mark sentence or phrase boundaries. Nevertheless, they probably can serve as indicators of unit boundaries at least as well as randomly-picked boundary positions.

Since I wanted to use the TreeTagger in all experiments, there were two options to integrate POS information in the input data. First, morphological and POS tags can be presented in turn, as shown in (ii) below. Second, POS tags could be appended as suffixes to wordforms, as in (iii). With the first option, the TreeTagger would make use of POS information in its ngram model; with the second option, the suffix lexicon would record POS-morphology dependencies. (i)–(iii) show example input for all three scenarios, for the sequence *werde disemo* 'would this'.

(i) *No use* of POS; input example:

```
werde    3.Sg.Pres.Subj
disemo   Neut.Dat.Sg
```

(ii) *Successive pairs* of <word, morph><word, POS>:
(or vice versa: <word, POS><word, morph>):

```
werde    3.Sg.Pres.Subj
werde    VAFIN
disemo   Neut.Dat.Sg
disemo   PD
```

(iii) *Merged pairs* of <word.POS, morph>:

```
werde.VAFIN    3.Sg.Pres.Subj
disemo.PD      Neut.Dat.Sg
```

The task based on successive pairs seems harder than the task with merged pairs: Successive pairs involve learning POS and morphology assignments simultaneously. With merged pairs, in contrast, the POS tags are given (as part of the word forms). However, to make the scenario realistic, the POS tags of the evaluation data have been assigned automatically and, hence, are incorrect to a certain extent. To assess the impact of incorrect POS tags, I repeated the evaluation of Scenario (iii) with gold POS annotations, which gives us an upper bound of the approach.

The results of the different scenarios are summarized in Table 3. For each scenario, mean and standard deviation of per-word accuracy across the 10 folds are given.[11] I now check the predictions from Sec. 2 against the figures in Table 3.

Prediction 1: Tagging normalized data should be easier Tagging with normalized word forms turns out better, as expected. This holds for both morphological and POS tagging.[12] Improvements are more pronounced with CG data (4.74–7.10 percentage points) than with UG data (3.56–5.36). There is no obvious explanation for this

[11] Evaluation of Scenarios (ii) and (iii) only considers morphological tags. Reordering the pairs as POS > morph resulted in slightly lower accuracy (< 1.6 percentage points). A more detailed evaluation of tagging POS can be found in Dipper (2010).

[12] Normalization resulted in a highly significant increase of accuracy in all scenarios (paired t-test; $p<.001$).

Morphology Scenario	Dialect	Tagger	Word Forms *diplomatic*	Word Forms *normalized*
(i) No use	CG	*gen*	73.91 ± 0.51	***79.70*** ± 0.36
		spec	72.64 ± 0.54	78.43 ± 0.53
	UG	*gen*	73.85 ± 1.16	78.28 ± 1.71
		spec	73.23 ± 1.02	78.15 ± 1.28
	TIGER 1,000 K		—	79.08
	210 K	≈ *gen*	—	76.95
	90 K	≈ *spec*	—	75.71
(ii) Successive pairs (morph > POS)	CG	*gen*	74.23 ± 0.51	**80.84** ± 0.55
		spec	72.37 ± 0.51	79.47 ± 0.50
	UG	*gen*	74.17 ± 1.10	79.11 ± 1.51
		spec	73.27 ± 0.96	78.63 ± 1.30
(iii) Merged pairs	CG	*gen*	74.39 ± 0.50	***79.81*** ± 0.42
		spec	72.86 ± 0.36	78.48 ± 0.53
	UG	*gen*	74.07 ± 0.88	77.63 ± 1.99
		spec	73.14 ± 0.85	77.02 ± 1.69
(iv) Gold POS (with (iii))	CG	*gen*	77.14 ± 0.47	*82.19* ± 0.39
		spec	75.54 ± 0.40	80.80 ± 0.49
	UG	*gen*	76.79 ± 0.87	80.83 ± 1.56
		spec	75.79 ± 0.86	80.26 ± 1.22

Part of Speech	Dialect	Tagger	Word Forms *diplomatic*	Word Forms *normalized*
	CG	*gen*	86.92 ± 0.64	91.66 ± 0.47
		spec	86.62 ± 0.63	91.43 ± 0.39
	UG	*gen*	88.88 ± 0.68	92.83 ± 0.39
		spec	89.16 ± 0.75	**92.91** ± 0.29
	TIGER 1,000 K		—	95.81
	210 K	≈ *gen*	—	95.67
	90 K	≈ *spec*	—	94.39

Table 3: Results of different test runs for morphological tagging (table on top) and POS tagging (table at the bottom), based on different types of word forms, dialect subcorpora, and taggers. For each scenario, mean and standard deviation of per-word accuracy across the 10 folds are given (all values are percentages). The overall best results for morphological and POS tagging of MHG data are indicated in bold, best results for other scenarios in bold italics. Results of Scenario (iv) represent an upper bound. Selected results from simple training (no cross-validation/standard deviation) on NHG (TIGER) are added for comparison. Training data of 210 K corresponds to the training data of the generic tagger, 90-K-training data corresponds to the data of the CG-specific tagger.

difference — with both dialect subcorpora, the type-token ratios are almost cut in half with normalized data.

Comparing the two types of taggers, generic vs. specific, the tables show that the generic taggers almost always perform better than the specific ones (the exception is POS tagging of UG). This seems to indicate that enlarging the training set is favorable even if the input becomes more heterogeneous. However, the differences in accuracy are rather small in general, and not significant in some of the scenarios.[13]

Predictions 2a / b: CG data / UG data is easier to tag Judging from the morphological top results, performance on CG data is slightly superior to performance on UG data (Prediction 2a). However, most of the differences are not significant.[14] On the other hand, UG data yields the best result with POS tagging (Prediction 2b; highly significant differences). Maybe this "contradiction" can be attributed to the fact that the morphological ambiguity rate is more favorable for CG data (lower mean and smaller standard deviation and maximum than UG data), while the opposite is true of the POS ambiguity rate.

Predictions 2c / 4: Tagging MHG / NHG should be easier Looking at the morphology table, we see that tagging of MHG data indeed outperforms tagging of NHG data (thus confirming Prediction 2c). Turning to the morphology table, the picture is, again, reversed (thus confirming Prediction 4): NHG tagging is well above MHG tagging. When the training size is reduced, accuracy of NHG degrades to a certain extent, but clearly remains superior. As above, the discrepancy can be traced back to ambiguity rates, which favour morphology tagging of MHG data, and POS tagging of NHG data.

Prediction 3: POS tagging should be easier Prediction 3 is clearly borne out. The gap between morphological and POS tagging is more than 10 percentage points:

- Morph (Scenarios (i)–(iii)): $> 79\%$ (CG-*norm*), $> 77\%$ (UG-*norm*)
- POS: $> 91\%$ (CG-*norm*), $> 92\%$ (UG-*norm*)

Interestingly, Scenario (iii) is not superior to Scenario (i), which makes no use of POS tags at all. This seems to suggest that automatically-assigned POS tags could not improve morphological tagging. However, the results from Scenario (ii) show that some improvement can indeed be achieved.

[13] The differences between the generic taggers and the corresponding specific taggers are *not* significant when they are evaluated on data from UG-*norm* (morphology Scenario (i) and POS), and UG-*dipl* (POS) (paired t-test).

[14] The differences between CG and UG taggers *are* significant with the generic taggers applied to normalized data, in all scenarios (paired t-test; $p<.01$ to $p<.05$).

4 Summary

I presented a set of experiments in morphological and POS tagging of historical data. The aim of this enterprise is to evaluate how well a state-of-the-art tagger, such as the TreeTagger, performs in different kinds of scenarios. The results cannot directly compared to results from modern German, though: The corpora are rather small; historical data is considerably more diverse than modern data; and I used modified versions of the STTS.

To summarize the main results from the set of experiments: Simple training on historical data results in satisfiable results of $> 91\%$ accuracy for POS tagging. In contrast, morphological tagging ($> 79\%$ accuracy) needs more sophisticated methods. For instance, the RFTagger (Schmid and Laws, 2008) is able to analyze and decompose complex morphological tags and, thus, to reduce the problem of data sparseness that arises especially with large, fine-grained tagsets (but see Fn. 3). Normalization increases accuracy by 3.56–7.10 percentage points.

The evaluations show that fully-automatic annotations (without subsequent manual corrections) currently only make sense with POS taggers, but not (yet) with morphological taggers. Assuming that automatic annotations would be checked manually, it is interesting to know how many correct tags are among the top n most probable tags. If most of the time, the correct tag is easy to select, in an efficient way, the current performance of the taggers might not be such a problem, after all.

I computed the ranks of all correct tags for a CG-*norm* sample, tagged with morphology, Scenarios (iii), and POS, see Table 4. The morphology table shows that in 87.1% of the cases, the correct tag is among the top-3 ranks (POS: 96.2%).[15] This means that it would probably speed up the annotation process if human annotators were presented the first three most probable tags to choose from.

As a next step, I want to evaluate the RFTagger for tagging of historical data. In addition, I plan to perform a detailed analysis with the goal of relating the tagging results to linguistic features of the different dialects.

References

Brants, S., Dipper, S., Eisenberg, P., Hansen, S., König, E., Lezius, W., Rohrer, C., Smith, G., and Uszkoreit, H. (2004). TIGER: Linguistic interpretation of a German corpus. *Research on Language and Computation*, 2(4):597–620.

Dipper, S. (2010). POS-tagging of historical language data: First experiments. In *Proceedings of the 10th Conference on Natural Language Processing (KONVENS-10)*, Saarbrücken.

Giesbrecht, E. and Evert, S. (2009). Is part-of-speech tagging a solved task? an evaluation of pos taggers for the German Web as Corpus. In *Proceedings of the Fifth Web as Corpus Workshop (WAC5)*, pages 27–35.

[15] I set the probability threshold to .1, i.e., all tags with a probability higher than 10% of the probability of the best tag are output. Scenario (ii) cannot be easily evaluated in this respect, because the probabilities are distributed over both morphological and POS tags.

Morphology (iii)		
Rank	# Word forms	
1	7467	79.6%
2	600	6.4%
3	99	1.1%
None	1122	12.0%

Part of Speech		
Rank	# Word forms	
1	8160	92.0%
2	370	4.2%
None	303	3.4%

Table 4: Ranks of the correct tags, which have been sorted according to their probabilities (left: morphology, Scenario (iii), right: POS). Absolute and relative frequencies are given (no cross-validation). Rank "None" shows the number of word forms whose actual tag is not among the automatically-proposed tags. Ranks with less than 1% instances are not displayed.

Klein, T. (2001). Vom lemmatisierten Index zur Grammatik. In Moser, S., Stahl, P., Wegstein, W., and Wolf, N. R., editors, *Maschinelle Verarbeitung altdeutscher Texte V. Beiträge zum Fünften Internationalen Symposion Würzburg 4.-6. März 1997*, pages 83–103. Tübingen: Niemeyer.

Kroch, A., Santorini, B., and Delfs, L. (2004). Penn-Helsinki parsed corpus of Early Modern English. `http://www.ling.upenn.edu/hist-corpora/PPCEME-RELEASE-1/`.

Kroch, A. and Taylor, A. (2000). Penn-Helsinki parsed corpus of Middle English. Second edition, `http://www.ling.upenn.edu/hist-corpora/PPCME2-RELEASE-2/`.

Lexer, M. (1872). *Mittelhochdeutsches Handwörterbuch*. Leipzig. 3 Volumes 1872–1878. Reprint: Hirzel, Stuttgart 1992.

Pilz, T., Luther, W., Ammon, U., and Fuhr, N. (2006). Rule-based search in text databases with nonstandard orthography. *Literary and Linguistic Computing*, 21:179–86.

Rayson, P., Archer, D., Baron, A., Culpeper, J., and Smith, N. (2007). Tagging the Bard: Evaluating the accuracy of a modern POS tagger on Early Modern English corpora. In *Proceedings of Corpus Linguistics 2007*, University of Birmingham, UK.

Schiller, A., Teufel, S., Stöckert, C., and Thielen, C. (1999). Guidelines für das Tagging deutscher Textcorpora mit STTS (kleines und großes Tagset). Technical report, University of Stuttgart and University of Tübingen.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*.

Schmid, H. and Laws, F. (2008). Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging. In *Proceedings of COLING 2008*.

Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*.

Asif Ekbal, Francesca Bonin, Sriparna Saha, Egon Stemle, Eduard Barbu,Fabio Cavulli, Christian Girardi, Massimo Poesio

Rapid Adaptation of NE Resolvers for Humanities Domains using Active Annotation

The entities mentioned in collections of scholarly articles in the Humanities (and in other scholarly domains) belong to different types from those familiar from news corpora, hence new resources need to be annotated to create supervised taggers for tasks such as NE extraction. However, in such domains there is a great need for making the best use possible of the annotators. One technique designed for this purpose is **active annotation**. We discuss our use of active annotation for annotating corpora of articles about Archaeology in the Portale della Ricerca Umanistica Trentina.

1 Introduction

Many of the entities mentioned in collections of scholarly articles in subjects such as Archaeology, History, or History of Art do not belong to the types found in the news corpora on which Computational Linguistics work has focused, such as the MUC and ACE corpora. For instance, the most important entity types found in archaeological texts are `Culture`, `Site`, and `Artefact`. In some such domains, even if more familiar types such as `Person` play an important role, it is essential to distinguish between their subtypes. E.g., in History of Art articles, it is not enough to classify an entity as a `Person`; it is also crucial to recognize if a particular individual was a `Painter`, a `Sculptor`, an `Architect`, etc. Hence, dedicated resources need to be created to train Named Entity (NE) recognizers for these domains; training on news corpora is of limited use to extract semantic content from such articles.

However, creating resources is always expensive, and Humanities projects tend not to have lots of funding for these purposes. In addition, collections of articles in the Humanities tend to be fairly small.It is therefore essential to use the limited funding available wisely, and to maximise the benefit to be obtained from the data. In other words, this is a domain for which **active learning** techniques (Settles, 2009), already used for NE tagging by, e.g., Vlachos (2006), seem ideally suited.

In this paper we discuss our work on using active learning for NE annotation of a corpus of scholarly articles in the Humanities being created in support of the creation of the Portale della Ricerca Umanistica Trentina, whose aim is to give scholars and the general public entity-, spatial-, and temporal-indexing based methods to access the many different collections of scholarly articles in the Humanities held by private and public collections in Trentino. After a brief introduction to the Portale della Ricerca Umanistica Trentina and the corpus under creation in Section 2, we introduce our

approach to combining active learning with CRF-based NE tagger in Section 3, and the results obtained in Section 4.

2 The Portale della Ricerca Umanistica Trentina

2.1 Aims

The Portale della Ricerca Umanistica Trentina (Humanities Research Portal, PRU) (Poesio et al., 2011a) is a pilot project to set up a one-stop search facility for repositories of scholarly articles and other types of publications in the Humanities held by digital libraries, museums and archives in Trentino. The portal will use content extraction techniques to automatically extract citations and semantic metadata including temporal, spatial, and entity references from the publications in those repositories. This information will then be used to offer visitors to the portal two main functionalities: **content-based search and browsing** and **semantic uploading**.

Besides standard keyword-based search, the PRU will also offer **entity-based search**. Two types of browsing will be possible: **spatial** and **temporal** browsing. Entity search allows users to retrieve all documents that discuss a particular entity irrespective of the way it's called–e.g., all Archaeological documents that discuss sites in which a particular shellfish was found irrespective of whether it's called in the document *Spondylus sp.* or *Spondilo*. Spatial browsing allows users to retrieve the publications that mention a particular locality in Trentino by visualizing a map of Trentino and clicking on the appropriate location. Temporal browsing (currently under development) will allow users to retrieve all historical articles discussing a particular period.

These novel types of searching and browsing will be supported by a **semantic upload function**: registered scholars and / or curators of the collections will be able to upload publications that will then be processed by the PRU pipeline discussed below to automatically extract both metadata and information about the publication to be inserted in the catalogue of the repository after being checked by the curator.

The first repository whose documents have been made accessible through the PRU is the collection of articles in the Archaeological domain in the APSAT / ALPINET digital library. We are currently working on indexing other repositories as well.

2.2 The apsat / alpinet Portal and Collection

The APSAT / ALPINET portal is a pilot Spatial Humanities project developed by the University of Trento's "B. Bagolini" Lab and allowing scholars to visualize Archaeological sites in the Alps through a Web GIS interface, through which Scholars can examine an area in general to find which sites are present, or look in detail at the features of a particular site. Through the portal, scholars can also access Archaeological articles about these sites, either through keywords or through the Web GIS interface.

Among the holdings of the portal is the complete collection of the journal *Preistoria Alpina* published by the Museo Tridentino di Scienze Naturali. We will focus on this collection in the present work. The collection is multilingual, containing articles written in English, French, German and Italian; in fact, as typical of the Humanities, many

NE type	Details
Culture	Artefact assemblage characterizing a group of people in a specific time and place
Site	Place where the remains of human activity are found (settlements, infrastructures, cimiteries, production site, ...)
Artefact	Objects created or modified by men (tools, vessels, ornaments, ...)
Ecofact	Biological and environmental remains different from artefacts but culturally relevant (e.g., *Spondylus*)
Feature	Remains of construction or maintenance of an area related with dwelling activities (fire places, post-holes, pits, channels, walls, ...)
Location	geographical reference
Time	historical periods
Organization	association (no publications)
Person	human being discussed in the text (e.g., Ötzi the Iceman, Pliny the Elder, Caesar)
Pubauthor	author in bibliographic references
Publoc	publication location
Puborg	publisher
Pubyear	publication year

Tabelle 1: Annotation scheme for Named Entities in the Archaeology Domain

articles are themselves multilingual, in that they contain, in addition to text in the main language, an abstract, keywords, and occasionally captions in a second language, often but not always English.

2.3 A Structure-Sensitive, Multilingual Pipeline

The articles to be made accessible through the PRU are processed by a pipeline that tokenizes, POS-tags, and NE tags the text in order to extract semantic indices (Poesio et al., 2011b). The pipeline, accessible as a Web service, is based on the TEXTPRO pipeline[1] (Pianta et al., 2008), and has two distinguishing features.

First, it is **structure sensitive**, in the sense that it includes a module that identifies the structure of a document to find citations and the like, in the manner of the FlyBase pipeline (Briscoe, 2011). Second, it is **constituent-level multilingual**, in that each constituent of the document structure is first run through a language identifier in order to find which version of the TEXTPRO system should be run on that constituent. (English and Italian are supported at the moment.) The first version of the pipeline included the default TEXTPRO NE tagger, ENTITYPRO, trained to recognize the standard ACE entity types. The objective of this work was to create a corpus that could be used to train a new NE tagger able to recognize the relevant entities in the APSAT / ALPINET collection.

2.4 Annotation Scheme for the apsat / alpinet collection

The most important NE types for the domain, identified in collaboration with the domain experts from the Bagolini Lab, are shown in Table 1.

Two broad classes of entities were identified on the basis of the types of queries that may be performed: entities that are part of what may be considered the content matter of the article (sites, cultures, individuals, names of ecofacts found in sites such as *Spondylus*), and entities that are part of the bibliographical references (e.g., authors of papers cited, year of publication, etc.). One of the most interesting aspects of these

[1] `http://textpro.fbk.eu/`

data is the prevalence of underspecified references. For instance, the term *Fiorano* refers to a culture from the Ancient Neolithic, that takes its name from the site *Fiorano*, which in turn is named from *Fiorano* Modenese in Emilia; for many uses of this term, it is impossible to tell which sense is intended. Possible solutions to this problem are to develop a system for underspecified typing like the `GPE` type in the ACE annotations[2] or guidelines forcing one interpretation. For the moment, coders have been asked to tag such cases as `underspecified`; we intend to return to the issue discussing options with the Archaeology experts, and develop a scheme / carry out agreement studies then.

3 Active Annotation and Conditional Random Fields

In this Section we first briefly review the notion of active annotation and the Conditional Random Fields approach to supervised learning we used to train our NER system, before introducing the approach to selecting the most informative samples we adopted.

3.1 Active Annotation

Active annotation–the term introduced by Vlachos (2006) to refer to the application of active learning (Settles, 2009) to corpus creation–is becoming a popular annotation technique because it can lead to drastic reductions in the amount of annotation that is necessary for training a highly accurate statistical classifier. In the traditional, random sampling approach, unlabeled data is selected for annotation at random. In contrast, in active learning, the most useful data for the classifier are carefully selected. In a typical active learning setup, a classifier is trained on a small sample of the data (usually selected randomly), known as the **seed** examples. The classifier is subsequently applied to a pool of unlabeled data with the purpose of selecting additional examples that the classifier views as informative. The selected data is annotated and the cycle is repeated, allowing the learner to quickly refine the decision boundary between the classes.

The key question in this approach is how to determine the samples that will be most useful to the classifier. A number of techniques have been proposed, ranging from choosing the sample on which the classifier trained on the seeds is less certain, to a variety of entropy-based approaches (Vlachos, 2006; Settles, 2009). We discuss our approach after first introducing the supervised training method we chose.

3.2 Conditional Random Fields

Conditional Random Fields (CRFs) (Lafferty et al., 2001) are undirected graphical models, a special case of which corresponds to conditionally trained probabilistic finite state automata. Being conditionally trained, these CRFs can easily incorporate a large number of arbitrary, non-independent features while still having efficient procedures for non-greedy finite-state inference and training, and are fast becoming the preferred method for NE tagging.

[2] (Buitelaar, 1998) is the earliest and possibly one of the most developed versions of this approach.

CRFs are used to calculate the conditional probability of values on designated output nodes given values on other designated input nodes. The conditional probability of a state sequence $s =< s_1, s_2, \ldots, s_T >$ given an observation sequence $o =< o_1, o_2, \ldots, o_T >$ is:

$$P_\wedge(s|o) = \frac{1}{Z_o} \exp(\sum_{t=1}^{T} \sum_{k=1}^{K} \lambda_k \times f_k(s_{t-1}, s_t, o, t)),$$

where $f_k(s_{t-1}, s_t, o, t)$ is a feature function whose weight λ_k, is to be learned via training. The values of the feature functions may range between $-\infty, \ldots + \infty$, but typically they are binary. To make all conditional probabilities sum up to 1, we must calculate the normalization factor,

$$Z_o = \sum_{s} \exp(\sum_{t=1}^{T} \sum_{k=1}^{K} \lambda_k \times f_k(s_{t-1}, s_t, o, t)),$$

which as in HMMs, can be obtained efficiently by dynamic programming.

To train a CRF, the objective function to be maximized is the penalized log-likelihood of the state sequences given the observation sequences:

$$L_\wedge = \sum_{i=1}^{N} \log(P_\wedge(s^{(i)}|o^{(i)})) - \sum_{k=1}^{K} \frac{\lambda_k^2}{2\sigma^2},$$

where $\{< o^{(i)}, s^{(i)} >\}$ is the labeled training data. The second sum corresponds to a zero-mean, σ^2 -variance Gaussian prior over parameters, which facilitates optimization by making the likelihood surface strictly convex. Here, we set parameters λ to maximize the penalized log-likelihood using Limited-memory BFGs (Sha and Pereira, 2003), a quasi-Newton method that is significantly more efficient, and which results in only minor changes in accuracy due to changes in λ.

When applying CRFs to the NER problem, an observation sequence is a token of a sentence or document of text and the state sequence is its corresponding label sequence. A feature function $f_k(s_{t-1}, s_t, o, t)$ has a value of 0 for most cases and is only set to be 1, when s_{t-1}, s_t are certain states and the observation has certain properties. We have used the C^{++} based CRF^{++} package, version 0.54[3], a simple, customizable, and open source implementation of CRF for segmenting or labeling sequential data.

3.3 Active Annotation with CRF

The main steps of the active annotation approach we followed are shown in Figure 1.

A feature vector consisting of the features described in the following Section is extracted for each word in the NE tagged corpus. Now, we have a training data in the form (W_i, T_i), where, W_i is the i^{th} word and its feature vector and T_i is its out-

[3] http://crfpp.sourceforge.net

Step 1: Evaluate the system on the gold standard test data.
Step 2: Test on the development data and calculate the conditional probabilities of all the output classes.
Step 3: Compute the confidence interval (CI) between the two most probable classes for each token.
Step 4: If CI is below the threshold value (set to 0.1 and 0.2) then
  Step 4.1: Add the NE token along with its sentence identifier and CI in a list of effective sentences, selected for active annotation (named as EA).
Step 5: Sort EA in ascending order of CI.
Step 6: Select the top most 10 sentences.
Step 7: Remove the 10 sentences along with the preceding one and following one sentences from the development set.
Step 8: Add the sentences to the training set.
Step 9: Retrain the CRF classifier and evaluate with the test set.
Step 10: Repeat steps 2-9 until the performance in two consecutive iterations be same.

Abbildung 1: Main steps of the proposed active learning technique

put tag. We consider various combinations from the set of feature templates specified by:

F_1 ={$w_{i-m}, \ldots, w_{i-1}, w_i, w_{i+1}, \ldots, w_{i+n}$; Combination of w_{i-1} and w_i; Combination of w_i and w_{i+1}; Feature vector consisting of root word, prefix and suffix, PoS, first word, infrequent word, digit, content words, and capitalization of w_i; B}

where B denotes the bi-gram template that calculates all the feature combinations of the current and previous tokens. The CRF is trained with the above-mentioned feature set and evaluated on the gold standard test set. For CRF training, we set the following parameter values: regularization parameter (a): default setting, i.e. L2; soft-margin parameter (c): trades the balance between overfitting and underfitting (default value); and cut-off threshold for the features (f): uses the features that occurs no less than its value in the given training data (set to 1, i.e. all the features that appear at least once in the training dataset is considered). We varied the context within the previous two and next two words. New sentences are chosen from the development set and added to the initial training set using the following selection method.

For each token of the dataset containing additional data to annotate, our CRF classifier outputs the confidence values (conditional probabilities) of each class. Our proposed selection criterion is to choose the token for which the differences between the confidence values of the most probable two classes is smaller– the hypothesis being that items for which this difference is smaller are those of which the classifier is less certain. A threshold on the confidence interval is defined, and at each iteration we select for further annotation the sentences in the 'extension' dataset containing such items, have the annotators label them, and add them to training.

We tested two ways of adding to the training set: either (i) add only the current sentence that contains the most informative example, or (ii) add the current sentence

Set	# token	# NEs
Training	20,739	2,611
Development	5,292	622
Test	11,534	1,582

Tabelle 2: Statistics about the training, development and test sets

along with the previous one and next one sentences. Thus, in each iteration, we add either 10 or 30 sentences to the training set. The iteration stops when the performance in two consecutive iterations doesn't change.

4 Annotation Experiments

4.1 Datasets

In order to train and evaluate NE taggers for the domain, a small collection of papers from the journal *Preistoria Alpina* was annotated. 11 articles from the journal, for a total of around 50,000 tokens, were annotated according to the scheme in Section 2.4. Of these, five articles were randomly chosen as training set, three as test set, and three articles for active annotation and development. Some statistics about the training, development and test tests are shown in Table 2.

Basic NE tags were converted into the BIO format, where B–, I– and O– denote the beginning, inside and outside tokens of NEs. For example, the name *le conchiglie* gets tagged as *le*/B-Ecofact *conchiglie*/I-Ecofact.

4.2 Named Entity Features

We use the following main set of features, which are domain as well language independent in nature, and automatically extracted without the help of any domain dependent resources and/or language specific rules. We also compared these results with the results obtained by adding information extracted from a gazetteer.

1. Context words: These are the preceding and succeeding words of the current word. This is based on the observation that surrounding words carry effective information for the identification of NEs.

2. Word suffix and prefix: Fixed length (say, n) word suffixes and prefixes are very effective to identify NEs and work well for the highly inflective Indian languages. Actually, these are the fixed length character strings stripped either from the rightmost or from the leftmost positions of the words. If the length of the corresponding word is less than or equal to $n-1$ then the feature values are not defined and denoted by ND. The feature value is also not defined (ND) if the token itself is a punctuation symbol or contains any special symbol or digit. This feature is included with the observation that NEs share some common suffixes and/or prefixes. Here, we consider prefixes and suffixes of length upto 3 characters.

3. First word: This is a binary valued feature that checks whether the current token is the first word of the sentence or not. We consider this feature with the observation that the first word of the sentence is most likely a NE.

4. Word length: We define a binary valued feature that fires if the length of w_i is greater than a pre-defined threshold. Here, the threshold value is set to 5. This feature captures the fact that short words are likely not to be NEs.
5. Infrequent word. A list is compiled from the training data by considering the words that appear less frequently than a predetermined threshold. The threshold value depends on the size of the dataset. Here, we consider the words having less than 10 occurrences in the training data. Now, a feature is defined that fires if w_i occurs in the compiled list. This is based on the observation that more frequently occurring words are rarely the NEs.
6. Capitalization: This is a binary valued feature that determines whether the word starts with a capital letter or not. This feature captures the fact that capitalized words are most likely NEs.
7. Part-of-Speech (PoS) information: PoS information of the current and/or the surrounding tokens(s) extracted using TextPro were used for NE identification.
8. Word normalization: We use a normalization feature clustering the words that have similar structures. This feature indicates how a target word is orthographically constructed. Word shapes refer to the mapping of each word to their equivalence classes. Here each capitalized character of the word is replaced by 'A', small characters are replaced by 'a' and all consecutive digits are replaced by '0'. For example, *Dalla* is normalized to *Aaaaa*, *123* is normalized to *0* and *1993* is also normalized to *0*.
9. Root word: The stems of the wordforms, extracted using TextPro.
10. Digit features: Several digit features are defined depending upon the presence and/or the number of digits and/or symbols in a token. These features are digitComma (token contains digit and comma), digitPercentage (token contains digit and percentage), digitPeriod (token contains digit and period), digitSlash (token contains digit and slash), digitHyphen (token contains digit and hyphen) and digitFour (token consists of four digits only).
11. Content words in global context: This feature is based on global contextual information. We consider all unigrams in contexts $w_{i-3}^{i+3} = w_{i-3} \dots w_{i+3}$ of w_i (crossing sentence boundaries) for the entire training data. We convert tokens to lower case, remove stopwords, numbers and punctuation symbols. We define a feature vector of length 10 using the 10 most frequent content words. Given a classification instance, the feature corresponding to token t is set to 1 iff the context w_{i-3}^{i+3} of w_i contains t.

4.3 Results

We trained a CRF model with the feature set mentioned in Section 4.2. We conducted a number of experiments with the various context sizes within the context window of $w_{i-2}, \dots, w_{i+2}$, and the feature template as mentioned in Section 3.3. We observed the best performance with the context of w_{i-1}, w_i, w_{i+1}, and thus only report its results. The best configuration is obtained by tuning the system on the development data. The system is evaluated using the evaluation metrics of standard recall, precision and F-measure. We used strict matching criteria, i.e. the system is given full credit only if the predicted labels of all the tokens of a NE is same as that of the gold labels.

Iteration number	Threshold=0.1			Threshold=0.2			Baseline (random)		
	r	p	F	r	p	F	r	p	F
1	63.02	65.48	64.23	64.32	67.83	66.03	64.64	66.35	65.47
2	64.73	67.11	65.90	65.84	68.81	67.29	64.21	65.99	65.09
3	65.08	67.92	66.47	66.10	69.6	67.81	65.40	66.90	66.14
4	65.66	68.41	67.01	66.80	70.09	68.41	65.86	67.73	66.78
5	66.82	69.62	68.19	67.68	70.92	69.27	65.54	67.25	66.39
6	67.31	70.06	68.66	68.26	70.26	69.24	65.66	67.25	66.44
7	67.63	70.31	68.94	68.26	70.54	69.38	65.77	67.41	66.58
8	67.63	70.31	68.94	68.26	70.54	69.38	66.90	68.56	67.72
9	67.86	70.57	69.19	68.83	70.99	69.89	67.19	68.90	68.04
10	67.86	70.57	69.19	68.83	70.99	69.89	67.19	67.90	68.04

Tabelle 3: Evaluation results of active learning with (a) threshold=0.1 (b) threshold=0.2 (c) random selection. Here, 'r': recall, 'p': precision, 'F': F-measure (we report percentages)

Iteration number	Threshold=0.1					Threshold=0.2				
	r	p	F	#sentence added	#NE added	r	p	F	#sentence added	#NE added
1	67.51	66.93	67.18	27	113	65.52	68.93	67.18	27	113
2	66.08	67.29	65.65	23	115	66.08	69.29	67.65	23	115
3	66.46	69.36	67.88	24	118	66.46	69.36	67.88	24	118
4	67.29	70.08	68.66	25	123	67.29	70.08	68.66	25	123
5	68.87	71.24	70.04	19	68	68.87	71.24	70.04	19	68
6	69.19	71.19	70.18	8	16	68.86	71.57	70.19	17	35
7	69.19	71.19	70.18	1	3	69.51	71.47	70.48	3	5
8	69.19	71.19	70.18	0	0	69.51	71.47	70.48	0	0
9	69.19	71.19	70.18	0	0	69.51	71.47	70.48	0	0
10	69.19	71.19	70.18	0	0	69.51	71.47	70.48	0	0

Tabelle 4: Evaluation results of active learning with (a) threshold=0.1 (b) threshold=0.2 by including gazetteer based features (we report percentages)

We experimented with the selection criteria that not only adds the current sentence but also adds the surrounding sentences (the preceding and the following sentences). We experiment with this selection with the intuition that wider context could give more useful information to the statistical classifier. For selecting the candidates of annotation, we determine the appropriate confidence thresholds from the development set.

The results of the proposed active learning technique with the confidence threshold of 0.1 are presented in Table 3. Here, the 10 most effective sentences and their preceding one and following one sentences are removed from the development set and added to the training set. The highest performance obtained with this method are recall, precision and F-measure values of 67.86%, 70.57% and 69.19%, respectively. This result is obtained at the ninth iteration and does not improve in the next iteration.

The results with a threshold of 0.2 are also shown in Table 3. The table shows that this threshold results in a better performance than with a threshold of 0.1: we obtained recall, precision and F-measure values of 68.83%, 70.99% and 69.89%, respectively.

The results of the baseline model, where in each iteration 10 sentences with their preceding and following ones are randomly chosen from the development set and added to training set, are shown in Table 3. Recall, precision and F-measure values of 67.19%, 67.90%, and 68.04%, respectively. This is lower in comparison to our proposed approach by 1.64, 2.09 and 1.85 percentage of recall, precision and F-measure values, respectively.

In our next experiment we used two gazetteers for the types `SITE` and `CULTURE` extracted from the ALPINET / APSAT database and containing 2,078 and 98 wordforms,

Class	Bound. Id. Error %
Artefact	0.08
Location, Site, Culture	0.05
Ecofact	0.3
Time	0.01
Pubauthor, Publoc, Pubyear	0
Feature, Person, Puborg	-

Tabelle 5: Bound[ary] Id[entification] Error out of the total of NE (both B- and I-) per category

respectively. These gazetteers were used to compute two binary valued features included into CRF. The features fire iff the current token matches with any element of the gazetteers. The system is retrained by including this feature to the previous feature set (c.f. Section 4.2) and keeping all other parameters unaltered. Overall evaluation results with two different thresholds 0.2 and 0.1 are reported in Table 4. We here again experimented with the selection criteria that not only adds the current sentence but also adds the surrounding sentences (preceding one and following one sentences). We have also shown in Table 4 that the number of sentences and number of named entities added from the development set to the training set in each iteration. At the end of 10th iteration with threshold equals to 0.1, it shows the overall recall, precision and F-measure values of 69.19%, 71.19%, and 70.18%, respectively. Again at the end of 10th iteration with threshold equals to 0.2, it shows the overall recall, precision and F-measure values of 69.51%, 71.47%, and 70.48%, respectively. Comparisons between Table 3 and Table 4 suggest that gazetteers help to improve the performance. The baseline model (based on random selection) showed the recall, precision and F-measure values of 68.66%, 70.51% and 69.57%, respectively. In table we also show the number of sentences and NEs that are added to the initial training data in each iteration. The instances of B- and I- are treated as two different counts for NEs.

4.4 Error analysis

We carried out two types of analysis: of the ability of the system to identify named entity boundaries (here called **identification problem**), and of its ability to correctly classify the mentions (**classification problem**).

To evaluate identification, we calculated the amount of mismatches between B-subtype and I-subtype for every class: those cases in which the system succeeds in recognizing the NE class, but fails to identify the correct bound. We only considered correctly identified entities (e.g. a true positive Artefact), calculating, among these, the error rate due to border mismatches (e.g. a B-Artefact marked as I-Artefact or viceversa).

In Table 5 we report the boundary identification error out of the total amount of NE per class.[4] In most cases, the problem of border identification lies in the ability of the system of incorporating the complex preposition which opens the mention; the lack of a

[4] Given the classes B-Artefact and I-Artefact, we calculated the ratio between the *FN*s and the population(B-artefact+I-artefact). Since we consider only cases in which the entity is correctly identified, we end up having a binary situation (either *b-entity* or *i-entity*); thus, FPs are not relevant as they overlap with the FNs of the other class.

Class	TP	FP	FN	Tot Retr	Total	P	R	F-M
B-Artefact	26	70	21	96	47	0.27	0.55	0.36
B-Culture	12	34	17	46	29	0.26	0.41	0.32
B-Ecofact	164	37	107	201	271	0.82	0.61	0.69
B-Feature	0	9	0	9	-	0	-	
B-Location	117	78	52	195	169	0.6	0.69	0.64
B-Person	0	20	0	20	-	0	-	
B-Pubauthor	380	23	55	403	435	0.94	0.87	0.91
B-Publoc	2	1	3	3	5	0.67	0.4	0.5
B-Puborg	1	0	7	1	8	1	0.13	0.22
B-Pubyear	265	20	10	285	275	0.93	0.96	0.95
B-Site	57	64	66	121	123	0.47	0.46	0.47
B-Time	97	14	44	111	141	0.87	0.69	0.77
I-Artefact	70	76	27	146	97	0.48	0.72	0.58
I-Culture	20	48	26	68	46	0.29	0.43	0.35
I-Ecofact	232	40	121	272	353	0.85	0.66	0.74
I-Feature	0	0	14	0	14	-	0	-
I-Location	262	164	66	426	328	0.62	0.8	0.69
I-Person	0	0	24	0	24	-	0	-
I-Pubauthor	64	9	40	73	104	0.88	0.62	0.72
I-Publoc	6	0	30	6	36	1	0.17	0.29
I-Puborg	13	1	24	14	37	0.93	0.35	0.51
I-Pubyear	0	0	2	0	2	-	0	-
I-Site	168	98	95	266	263	0.63	0.64	0.64
I-Time	400	40	66	440	466	0.91	0.86	0.88
Total	2356	817	946	3173	3302	-	-	-
O	11703	126	38	11829	11741	0.99	1	0.99

Tabelle 6: Precision and Recall per class

consistent number of these mentions in the training set can be behind this difficulty.

Classification accuracy is a measure of the system w.r.t. its ability to correctly assign the exact class to the identified NE. As shown in Table 6, there are categories in which the NE tagger obtains very good results, such as `Pub-year`, `Pub-author`, and `Time`. (Not surprisingly, these are among the classes most frequently studied in HLT.) On the other hand, categories such as `Artefact`, `Culture` and `Site` are more difficult to classify. These classes are also difficult for coders, which suggests that in part the problem may be that this new domain still isn't well understood.

More specifically, the most frequent confusions are a) `Culture` vs `Site` and vs `Time` b) `Site` vs `Location` and c) `Ecofact` vs `Artefact`. The confusions under a) were expected, because the classes `Culture` and `Site`, and `Culture` and `Time`, are systematically correlated: e.g., many cultures such as *Starcevo* are so-named from a so-called **type site**. As a result, whereas 55% of `Culture` NEs are correctly identified, 20% are marked as `Site`. To study this issue we asked annotators to mark mentions they felt could instantiate to different classes with two labels, *label 1* (the more likely) and *label 2*, and set an `underspecification` attribute (see (Poesio and Artstein, 2005) for a more extensive study of this type of annotation), and we found that the cases of confusion had often been marked as `underspecified`.[5] In class b), `Site` vs `Location`, 70% of `Site` NEs

[5] Though human annotators mark these entities with two labels, during training, the NE tagger choses only the first one, the one considered most likely; for this reason the underspecification issue does not affect the evaluation phase.

are correctly identified, but 14% is marked as `Location`. In this case we have a semantic ambiguity between classes that share similar context: e.g. *nella vicina Alta Valtrompia* vs *il sito nei pressi di Bressanone*. As expected, the introduction of the Gazetteer reduced the distance in particular in this case. Finally, for class c), `Ecofact- Artefact`, 65,5% of `Ecofact` NE a are correctly identified, but 19% are marked as `Artefact`, while only 5% is confused with `Location`, which is the second most confused class. This case also concerns a critical distinction, often marked as `underspecified` by our coders, and the focus of ongoing discussions in the domain experts community.[6]

5 Conclusions

Our results suggest, first of all, that active annotation does lead to better results than random sampling; and second, that our approach leads to reasonable results with relatively small amounts of trained data. Our future work will include, first of all, revising the coding scheme for the Archaeology domain in collaboration with the Archaeology experts, in particular developing a solution to the Underspecification problem and carrying out agreement tests; and testing the generality of our results by incorporating a new domain.

Acknowledgments

This work was in part supported by the LiveMemories project.

Literatur

Briscoe, E. e. a. (2011). Intelligent information access from scientific papers. In et al, J. T., editor, *Current Challenges in Patent Information Retrieval*. Springer.

Buitelaar, P. (1998). *CoreLex : Systematic Polysemy and Underspecification*. PhD thesis, Brandeis University.

Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *ICML*, pages 282–289.

Pianta, E., Girardi, C., and Zanoli, R. (2008). The textpro tool suite. In *Proc. of 6th LREC*, Marrakech.

Poesio, M. and Artstein, R. (2005). The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In Meyers, A., editor, *Proc. of ACL Workshop on Frontiers in Corpus Annotation*, pages 76–83.

Poesio, M., Barbu, E., Bonin, F., Cavulli, F., Ekbal, A., Saha, S., Stemle, E., Girardi, C., Nardis, D., and Nardelli, F. (2011a). The humanities research portal: Human language technology meets humanities publication repositories. In *Proc. of SDH*, Copenhagen.

Poesio, M., Barbu, E., Stemle, E., and Girardi, C. (2011b). Structure-preserving pipelines for digital libraries. In *Proc. of LaTeCH*, Portland, OR.

[6] In this domain there is also a systematic ambiguity between `Culture` and `Artefact`, because of the tradition to name other cultures according to their most distinctive artefact, as in: *Cultura dei Vasi a bocca quadrata*.

Settles, B. (2009). Active learning literature survey. In *In Computer Sciences Technical Report 1648 University of Wisconsin-Madison.*

Sha, F. and Pereira, F. (2003). Shallow Parsing with Conditional Random Fields. In *Proceedings of NAACL '03*, pages 134–141, Canada.

Vlachos, A. (2006). Active annotation. In *Proc. EACL 2006 Workshop on Adaptive Text Extraction and Mining*, Trento.

Voula Giouli

Annotating corpora from various sources in the humanities domain: shortcomings and issues

In this paper, we present work aimed at the linguistic annotation of Greek corpora that belong to the humanities domain, the focus being on the methodological principles as well as the implementation framework adopted. This framework builds on an existing XML annotation platform that was initially developed in an Information Extraction setting and in order to cope with texts that pertain to domains such as news, administrative, economics, etc.; we elaborate on the initial steps taken towards customization of the tools.

1 Introduction

Over the last years, there has been a significant effort in creating various annotated corpora that have been made available in order to serve as training and evaluation benchmarks for Natural Language Processing (NLP) tasks even for the so-called "less-resourced" languages. These corpora are meant to model language used in specific domain-oriented applications. In this paper we present work aimed at the development and annotation of text corpora pertaining to disciplines in the humanities. Annotations have been carried out with the use of existing generic NLP tools that are currently customized so as to handle older and/or dialectical language varieties depicted in the corpus.

The paper is organized as follows. In section 2 we provide an overview of the corpus collection, regarding composition, size, features and the metadata schema employed for the representation of the digital content. In the following section, we present the levels of annotation that have been implemented so far, along with a suite of corresponding generic NLP modules for the Greek language, which have been used to initiate the annotation process. In section 4, after commenting on problems arising from the language varieties at hand, we present the steps taken so far for the customization of the afore-mentioned generic tools. The multilevel annotation tool has been employed for the extensive manual annotation of the data is presented in section 5, whereas initial remarks are discussed in section 6. Finally, conclusions are outlined in section 7.

2 Corpus description

The corpus hereby presented was initially developed in order to be integrated into a platform aimed to promote and highlight the cultural heritage of the Northern areas of Greece the focus being on literature and folklore. The intended infrastructure (corpora and platform) were targeted to a rather diverse audience ranging from students to teachers and the general public alike. Corpus collection followed a two-steps procedure thus reflecting the modular approach taken within the project it was initially intended for. At the first stage, texts were selected adhering to the genres of (a) literature, (b) folktales and legends, and (c) folklore texts (i.e., those commenting on and/or depicting a wide range of aspects of everyday human

activity such as traditions, customs, practices, spiritual beliefs in past eras). A set of pre-defined criteria conforming to specifications in the axis of place and time guided the design and content of the intended textual resource.

However, literature cannot be set apart from its era and the historical settings. Moreover, interpretations made by specialists (i.e., literary critics) are always sought for. To make therefore the platform as complete as possible, and a useful tool to prospect users, the initial collection was further coupled with texts that were intended to serve as accompanying material to the literary texts, namely authors' biographies, commentaries and literary criticism; on top of that, historical texts depicting the background of literary texts were also added to the collection. Within the intended project, the afore-mentioned criticism and historical texts had a two-fold purpose: (a) to be used as accompanying material, and (b) to guide the extraction of indexing terms, the ultimate goal being the development of a thesaurus that will enhance access to and retrieval of the primary data.

The so-collected corpus amounts to 304K words, and it covers a time span from the 19th century till the present day. Moreover, it represents a range of language varieties in the axes of time (i.e., contemporary, non-contemporary language) and place. More precisely, texts dated prior to 1976 (when *Dimotiki* was declared the official language of Greece) depict the non-contemporary language variety of "katharevousa", whereas, a number of literary and folklore texts depict the language variety spoken in the northern areas of Greece (*northern* dialect). Corpus composition and language coverage are depicted in Table 1 below:

	contemporary	***non-contemporary***	***dialectical***	***total***
literature	54K	57K	49K	160K
folktales	18K	-	23	41K
folklore	19K	22K	-	41K
historical	-	62K	-	62K
total	91K	141K	72K	304K

Table 1: Corpus composition

2.1 The metadata schema

To ensure easy access and re-usability of the corpus, a metadata scheme compliant with state-of-the-art standards was adopted, with certain modifications that cater for the peculiarities of the texts. The encoding scheme is compliant with the specifications of the Text Encoding Initiative (*http://www.tei-c.org*) (TEI Guidelines for Electronic Text Encoding and Interchange). To this end, metadata elements have been deployed reflecting bibliographical information that is primarily important for text identification with respect to text title, author, publisher, publication date, etc. (bibliographical information). Additionally, information on certain characteristics of the texts, such as language variety or sublanguage (contemporary/non-contemporary/idiomatic) was also added to the metadata descriptions manually.

In order to ensure documentation completeness and facilitate the inter-relation among primary data (i.e., literary texts) and the accompanying material (biographies, commentaries, criticism, etc), the documentation scheme has been extended accordingly so as to include such descriptive elements. Information regarding text type/genre and topic (where applicable) was also added manually on the grounds of generally accepted standards. To this end, *folklore* texts have been classified in accordance with the Classification scheme developed and used by the Library of Congress *(http://www.loc.gov/catdir/cpso/lcco/)*, whereas folktales categorization is conformant with the widely established Aarne-Thompson classification system (Aarne, 1961).

To keep track of the status and management of Intellectual Property Rights of the selected documents, appropriate metadata elements have been employed too.

From another perspective, the metadata scheme implemented in this project caters for the linguistic annotations that were provided for. The scheme employed builds on XCES, the XML version of the Corpus Encoding Standard (*XCES, http://www.cs.vassar.edu/XCES/* and CES, *http://www.cs.vassar.edu/CES/CES1-0.html*), which has been proposed by EAGLES (*http://www.ilc.cnr.it/EAGLES96/home.html*). This standard is compatible with TEI and can be mapped if considered appropriate. It has also been favored due to the fact that it is more appropriate for linguistically annotated corpora. On top of that, metadata elements inspired by the Dublin Core Metadata Initiative (DCMI) standard, including among others, *Annotator* (an entity responsible for providing the annotation content), *Subject* (what the annotation is about), *Resources* (the resources and tools that have been used in the annotation session), *Language* and *Date* (a date associated with the current session) are also included in the metadata headers.

Most of the aforementioned metadata descriptions were added manually to the texts and are kept separately from the primary data in an xml header that is to be deployed by the text management system for search and retrieval purposes. Moreover, metadata are stored separately from the raw data.

3 Corpus annotation

After text selection, digitization and extended manual validation (where appropriate) were performed. Normalization of the primary data was kept to a minimum so as to cater, for example, for the conversion from the Greek polytonic to the monotonic encoding system.

However, corpus development within the NLP community is meaningless unless appropriate encodings or annotations are included that are designed to support different views of the language. To this end, to further enhance the textual collection, rendering it, thus, a useful resource to prospective users and researchers, further annotations at various levels of linguistic analysis were integrated into the primary textual material. These annotations served a two-fold purpose, that is, to enhance efficient indexing and retrieval of the textual documents, and to further facilitate the study of textual data and the elicitation of meaningful observations over the data.

3.1 Linguistic Annotation of texts

Linguistic annotation involves the following levels of analysis: (a) Part-of-Speech (POS) tagging and lemmatization, (b) surface syntactic analysis (chunking), (c) indexing with terms/keywords and Named Entities (NEs), (d) coreferece encoding, and (e) dependency annotation. These layers of linguistic annotations will be further elaborated in the remaining of this section. More precisely:

Part-of-Speech tagging (POS-tagging) is the first stage of linguistic analysis and involves the assignment of word class (part of speech) information coupled with more fine-grained morphosyntactic characteristics to every token in the text. Our scheme employs a PAROLE-compliant tagset. Surface syntactic analysis (chunking) consists in the recognition of non-recursive phrasal categories: adjectives, adverbs, prepositional phrases, nouns, verbs (chunks). Main as well as subordinate clauses were also recognized and labeled as appropriate.

Building on existing schemes developed for the annotation of NEs in texts, namely MUC-7 (Message Understanding Conference) and ACE (Automatic Content Extraction)), annotation at this level of linguistic analysis caters for the recognition and classification of the following types of entity names: person (PER), organization (ORG), location (LOC) and geopolitical entity (GPE). The generic schema also caters for the identification of numerical values: *(MONEY), (PERCENT),* and certain time expressions: *(DATE)* and *(TIME)* – yet, only the former was retained. Moreover, NE's of the type *(LOC)* were also assigned a subtype value, namely: geographical region (GEO) and facility (FAC). Though compatible in form with ACE, in that it retains most of the types and subtypes provided for by ACE, our classification schema differs in that disambiguation between *(LOC)* and *(GPE)* uses of names is being attempted.

At the next stage, terms were spotted and recognized. Conceived as the linguistic representation of concepts pertaining in a certain subject field, and being characterized by special reference "as opposed to words that function in general reference over a variety of codes" (Sager, 1980), terms and their identification were deemed meaningful only for the more "technical" texts in the collection, namely those pertaining to the domain folklore. Annotation at this level consisted in the selection of the word or word group that form a *simple-word* or *multi-word term in the given domain,* and their association with a pre-defined hierarchical list of topics. This list was created on the basis of the Library of Congress Classification scheme *(http://www.loc.gov/catdir/cpso/lcco/)*, and augmented on an as needed basis (see Fig. 1 below):

At the level of dependency annotation, the head-dependent relations among syntactic constituents were encoded, for representing the syntactic structure of a sentence. Grammatical roles that are identified and annotated include subjects, predicative complements, direct and indirect objects, prepositional phrases functioning as arguments or modifiers, and clausal arguments. Guidelines for the annotation of Modern Greek (MG) (Prokopidis, et al. 2005) and ancient Greek (Bamman et al. 2008) were taken into account. The latter were of particular importance for the annotation of texts written in the older language variety (*katharevousa*).

From the broad set of referential phenomena that characterize Greek language, we have focused on NP co-reference. In our work, two forms of co-reference have been accounted for: intra-sentential, in which case the co-referring expressions occur in the same sentence,

and inter-sentential, where a nominal expression refers to an entity mentioned in a previous sentence. The annotation involves: (a) the identification of *markables* in a sentence, that is, definite, indefinite and bare NPs, (b) the assignment of values to a set of attributes corresponding to their form (definite/indefinite/bare) and function (apposition, argument, etc), and (c) the identification of their antecedent. The interlinking of mentions of the same entity in a text results in the creation of *co-referential chains*. The anaphoric relation treated is that of identity. Our encoding scheme builds on the guidelines provided by the MATE project with certain modifications so as to cater for the particularities of the Greek language, as for example the fact that Greek is a pro-drop language.

3.2 Natural Language Processing tools

Annotations at almost all levels of linguistic analysis were performed semi-automatically (except for the last one that was applied manually), using a NLP pipeline developed at the Institute for Language and Speech Processing. The tools have been trained on Greek textual data from various sources (newspapers, internet, etc.) that cover domains such as finance, politics, sports, travel, etc. The main modules of this pipeline include a *tokenizer*, a *POS tagger and lemmatizer*, together with tools that recognize NEs and non-recursive syntactic units.

More precisely, at the first stage, handling and tokenization was performed using a Greek tokenizer that employs a set of regular expressions, coupled with precompiled lists of abbreviations, and a set of simple heuristics for the recognition of word and sentence boundaries, abbreviations, digits, and simple dates. To accomplish this task, we used the POS-tagger developed in-house (Papageorgiou et al. 2000) that is based on Brill's Transformation Based Learning architecture (Brill, 1997). Following POS tagging, lemmas retrieved from a Greek morphological lexicon were assigned to every word form.

A maximum-entropy Named Entity recognizer (Giouli et al. 2006) trained on financial and travel data identifies NEs of the afore-mentioned types (cf. above).

A term detection module (Georgantopoulos et al 2000) was then employed to identify terms in Greek text. It is a hybrid system comprising a regular expression-based term pattern grammar, and a statistical filter, used for the removal of terms lacking statistical evidence. Term Extractor functions in three pipelined stages: (a) POS annotation of the domain corpus, (b) corpus parsing based on a pattern grammar endowed with regular expressions and feature-structure unification, and (c) lemmatization. Candidate terms are then statistically evaluated with an aim to skim valid domain terms and lessen the over-generation effect caused by pattern grammars.

In parallel, a module responsible for the automatic identification of grammatical relations has been employed that works on the basis of a pattern matching mechanism. The main resource used at this stage is a sub-categorization frames lexicon. The entries have been retrieved from a database containing sub-categorization information for the 5927 most frequent verbs, 4950 most frequent nouns, and 375 most frequent adjectives of a general purpose corpus.

4 Validation of the automatic processing

As it has already been mentioned, annotation was performed in most cases semi-automatically, that is, automatic processing using the afore-mentioned NLP tools followed by human validation. To minimize the effect of error transferring from previous levels to consecutive ones, the output of each processing component was manually validated prior to being fed to the next processing module. Moreover, due to the fact that the POS-tagger has been reported to achieve high accuracy levels (F-score 0.97) on standard texts, manual annotation was performed by two expert linguists only on the sub-corpus that deviated from the norm, that is, the non-contemporary and dialectical texts. Accuracy at the levels of NE and term annotation was even lower ranging from 0.21-0.63 depending on the text type. Moreover, the initial NE annotation schema was proved to be inadequate for the texts at hand.

For each annotation level, initial guidelines were provided to the linguists in charge of each annotation task. These guidelines were initially developed by expert linguists on the basis of existing encoding specifications in view of training generic NLP tools for a certain domain/text type. Within the current project, however, the initial specifications were appropriately revised so as to accommodate the peculiarities of the data at hand. For example, at the POS-tagging level, the dative case or the morphologically distinct subjunctive mood of the *katharevousa* (see below) should be accommodated for, the ultimate goal being the efficient description of the language variety used in the older texts. Similarly, NE annotation was meaningless in the current setting (literature, legends, and folktales) if it was intended for entities of the type (ORG). To this end, only NE's of the type PER and LOC were retained, and the specifications were modified so as to also include entities that are of interest in the texts at hand. The following new entity (sub-) types were defined, and our initial annotation scheme was revised accordingly:

- PER.human: Names of people, either dead or alive were further classified as human
- PER.animal: Names of animals fell into this subtype
- PER.fictional: Names of fictional characters were also tagged
- PER.other: All other animate entities that do not fall into the above subtypes were tagged as PER.other.

After a brief testing period of the new schemas/guidelines, and following the amendments or clarifications that were considered appropriate, samples by the annotators were collected and inter-annotator agreement was examined. Labels assigned by the two annotators were compared, and if the same label was assigned to the same spans of text by both annotators, it was counted as a match, otherwise not. By this measure, the average agreement score was counted around c. 90% for all levels of linguistic analysis.

As it has already been said, the major shortcoming in this procedure consisted in that the automatic processing of the textual data yielded very poor results especially in the cases of texts depicting language varieties that deviate from the norm. Manual annotation, on the other hand, aimed at re-training the tools is costly and time-consuming. To reduce manual effort, human validation has been performed on half of the data.

A close inspection over the data helped us to identify errors, and also to classify the sources of erroneous output so as to find appropriate solutions. A close inspection over the data has revealed the following as the main error-baring cases:

- **problematic/erroneous output from the Optical Character Recognition (OCR) module** resulting into various misspellings or even into a intelligible output;
- **encoding problems,** resulting from the conversion of initial documents to a format that is appropriate for the processing tools;
- **various misspellings or variant spellings;**
- **non-standard orthography and spelling variation** due to the language variety depicted in the literary works and/or the non-contemporary and idiomatic texts in the folklore domain.
- **word-formation and or declension** in accordance with the paradigm of the older/regional language variety.

4.1 Annotating non-contemporary and idiomatic words

The texts collected do not depict or represent a uniform variety of the Greek language. Instead, depending on the text type and the date of publication three main varieties are depicted: (a) *Modern Greek,* the official language of Greece, (b) *katharevousa*, and (c) a language exhibiting features of the *Northern Greek dialects.*

The situation of *diglossia*, i.e. the simultaneous existence of a vernacular and a high variety of the Greek language, was prominent from the birth of the new country until practically the end of the 20th century. Shortly after Greece was declared independent in 1830, the language issue was raised. The traditionalist, influenced by the Enlightenment ideal for a national language "argued for the resurrection of the classical Greek, uncontaminated by 'impure' admixtures with which it had been 'polluted' during its contacts" (Dendrinos, 2007). Their opponents, on the other hand, favored over the usage of the language actually spoken by the people. In between the two options, a third one advocated the use of the current language, 'purified' through its infusion with classical Greek in terms of morphology, syntax and vocabulary. The latter, which bore also the symbolic charge of continuation of Ancient Greek, prevailed, leading to this situation of diglossia. The high variety, *Katharevousa* (from *katharo,* meaning "clean"), an imitation of classical Greek was used in administration, education, science while the low variety, *Dimotiki,* was used in everyday informal communication, literature (although not by all authors) and primary education. This situation is reflected in those texts in the collection which were dated prior to 1976, and the prevalence of the (mainly the historical ones, most of the folklore texts and a few literary ones)

Literary and folklore texts, on the other hand, depict a language variety exhibiting all the characteristics that are present in the *Northern Greek dialects* (roughly covering the areas of central Greece, Thessaly, Macedonia, Epirus, Thrace, Euboea, and some islands in the Ionian and NE Aegean). The peculiarities of this language variety of Greek consist in deviations from the norm with respect to: (a) phonological features (the characteristic process of high-vowel (i, u) deletion in unstressed syllables leading to the creation of various consonant clusters), (b) syntactical features (the use of 'object' pronoun forms as indirect objects), and (c) morphological features.

All texts in the collection are appropriately encoded with respect to the language variety used. However, to keep track of the lexical entries deviating from the norm, words/word forms were appropriately tagged with this respect. This was especially important for texts which depict more than one language varieties shifting one language variety to another (norm, "*katharavousa*", and the *regional variety* with all possible combinations).

4.1 Towards resource customization

Annotations applied to the texts automatically were checked manually by expert linguists using a graphical user interface suitable for manual annotation, verification and correction on the processed texts. It should be noted, however, that this was not a trivial task and posed many difficulties even to trained annotators, due to the fact that we had to cope with a number of phenomena not present in Modern Greek. After multiple passes over the data and the identification of the errors in the corpus the following preliminary actions were taken towards the customization of the POS-tagger:

- tagset expansion (in compliance to the PAROLE specifications) so as to capture the morpho-syntactic characteristics of the "katharevousa". To this end, the extended tagset caters characteristics such as the dative case in nouns, adjectives, articles, pronouns and in all the three genders, as well as the existence of participles, infinitives, and of the morphologically distinct subjunctive mood for verbs. This has resulted to the increase in the numbers of the allowed tags from 584 initially used for Modern Greek to 625 for the older language variety.
- enrichment of the morphological lexicon employed by the POS-tagger and/or revision in two axes: (a) inclusion of pronouns, adverbs, prepositions of the "katharevousa", etc. These were extracted from the validated material and further enhanced with entries from various sources (i.e., grammars, etc)
- word lists were further enriched with ambiguous words and wordforms that are specific to the language variety at hand.
- revision of the annotation specifications at POS-tagging level so as to capture the peculiarities of the language variety at hand.
- Revision of the specifications set by MUC/ACE for the identification of NEs that are more relevant to the text types (cf. above). Additionally, new trigger words were manually selected for inclusion in the relevant tool.

All the afore-mentioned lexical resources (lexicons, wordlists) will be added to the resources employed by the tagger and formal validation performed.

5 A graphical user interface: Marker

An important component in the whole process of annotation was the usage of a flexible annotation environment called *Marker*. Marker is a Graphical User Interface that allows annotators to have simultaneous views of all levels of previous annotations, while working at a particular task. It supports annotations at the following levels of linguistic analysis: (a) morpho-syntax; (b) chunk and recursive phrases; (c) Named Entities; (c) term spotting and annotation; (d) coreference annotation, and (e) annotation of grammatical relations.

Classes of XML annotations that share a common vocabulary and structure (morphology, syntax, etc.) are described in DTD's. The Marker looks for the relevant DTD when initiating

an annotation session and configures the GUI appropriately by providing the needed functionality to the annotator. This dynamic process of building and customising a GUI on the fly (based on external DTD files) is currently restricted to simple elementary structures which however fulfill most of our current annotation needs. Additionally, a validation step is being performed ensuring that a particular instance is compliant with the pre-specified constraints in the DTD's. This environment also encompasses an editor which was extensively used for the editing/modification of the initial metadata and/or the rearrangement of their hierarchy in the schema. New annotation schemas were also implemented using the functionalities provided by the tool.

Within the current project, the tool has also served as an aid in our lexicographic work. Although it is not a proper lexicographic environment, it allows, at the level of term annotation for the inclusion of other information, such as definition, and reliability. The former was automatically retrieved along with lemma information and domain type/subtype facilitating, thus the population of the glossaries that were developed for the specific collection and domains covered.

6 Discussion

The textual collection that has been described above was primarily intended for laymen. As it has already been pointed out, the ultimate goal of the whole project was to create a set of language resources along with an infrastructure targeted to a wide and rather diverse audience. The application was aimed to serve as a teaching aid either in the domain of literature and folklore, or even in language teaching and learning. However, we argue that it can also be perceived as a pilot work that may guide future large-scale endeavors aimed equally at researchers as well. In this respect, a more ambitious target of the project was to familiarize scholars in the humanities with applications assisting their research, and to raise awareness amongst scholars and researchers in the humanities with respect to the digital resources and capabilities offered by NLP. The intended tools would be useful for a number of applications ranging from automatic indexing and retrieval of documents in specialized digital libraries, to the extraction of glossaries and the (comparative) study of word usage across writers, local communities, etc. to mention but a few.

However, the tendency for creating textual collections coupled with metadata for a variety of languages and language varieties, and the resulting need for portability and customization of generic NLP tools, has brought about the issue of a basic research infrastructure that goes beyond the needs of customary language technology (LT) applications. This need guides us to the notion of the Basic Language Resource Kit (BLARK), which refers to a core set of language resources and LT tools that are deemed essential not only to basic research in LT but also to the development of a variety of applications for a particular language (i.e., linguistically annotated text corpora, lexical resources, tools for linguistic annotation of tools, etc). And although this notion usually refers to modern standard languages and state-of-the-art applications, researchers are know starting to argue in favor of the idea of a BLARK that goes beyond the standard or modern usage of language, extending itself to one or more of the following axes of language

variation: (a) community (languages, dialects, sociolects), (b) subject, purpose or medium (topics, genres), (c) time (historical language stages) (Borin et al. 2010). Indeed, this need is increasingly recognized by the language resource community and research funding agencies alike, and to this respect, the work presented here was conceived from the beginning as a contribution to a BLARK for the Greek language extended in the axes of time and community, the focus being at present on the creation of annotated corpora, the elaboration of annotation schemes, and the development/modification of accompanying lexical resources.

7 Conclusions

We have hereby presented work aimed at the annotation of specialized corpora comprising texts from the humanities disciplines. We have described the methodology adopted and the tools used, elaborating on the annotation schemas and the initial steps towards tool customization. Manual validation of the output of the automatic processing was intended for training the respective tools so as to handle texts in the domains and language varieties at hand.

References

AARNE, A. (1961). The Types of the Folktale: A Classification and Bibliography. Translated and Enlarged by Stith Thompson. 2nd rev. ed. Helsinki: Suomalainen Tiedeakatemia / FF Communications.

ACE. http://www.itl.nist.gov/iaui/894.01/tests/ace/

BAMMAN, D., AND CRANE, G. (2008). Guidelines for the Syntactic Annotation of the Ancient Greek Dependency Treebank (1.1). The Perseus Project, Tufts University. September 1, 2008.

BONTCHEVA, K., D. MAYNARD, H. CUNNINGHAM, AND H. SAGGION (2002). Using Human Language Technology for Automatic Annotation and Indexing of Digital Library Content. Lecture Notes In Computer Science, Vol. 2458. In Proc. of the 6th European Conference on Research and Advanced Technology for Digital Libraries. 613–625.

BORIN, L., AND M. FORSBERG (2008b). Something old, something new: A computationalmorphological description of Old Swedish. In LREC 2008 Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008). 9-16. Marrakech: ELRA.

BORIN, L., D. KOKKINAKIS, AND L. J. OLSSON (2007). Naming the past: Named entity and animacy recognition in the 19th century Swedish literature. In *Proc. of the ACL Workshop: Language* Technology for Cultural Heritage Data (LaTeCH.). 1-8. Prague: ACL.

BORIN, L., M. FORSBERG, AND D. KOKKINAKIS (2010). Diabase: Towards a diachronic BLARK in support of historical studies. In *Proc. of LREC 2010*.

BROWNING, R. (1991). Medieval and Modern Greek. 1st edition 1962, 2nd edition 1983; Greek edition 1991 Athens: Papadima Publications.

CHRISTIDIS, A.F., ed. (2000). La Langue Grecque et ses Dialectes. Thessaloniki: Centre de la Langue Grecque.

CRANE, G. (2002). Cultural Heritage Digital Libraries: Needs and Components. In Research and Advanced Technology for Digital Libraries. Lecture Notes in Computer Science. Vol. 2458. 51-60.

DAVIES, S., POESIO, M., BRUNESEAUX, F., ROMARY, L. (1998). Annotating coreference in dialogues: proposal for a scheme for MATE. First draft. Available at http://www.hcrc.ed.ac.uk/~poesio/MATE/anno_manual.html

DENDRINOS B., THEODOROPOULOU, M. (2007). Language issues and language policies in Greece. EFNIL, Riga http://www.efnil.org/documents/conference-publications/riga-2007/Riga-06-Dendrinos-Mother.pdf

GENEREUX, M. (2007). Cultural Heritage Digital Resources: From Extraction to Querying. In *Language Technology for Cultural Heritage Data (LaTeCH 2007), Workshop at ACL 2007*, June 23rd–30th 2007, Prague, Czech Republic.

GEORGANTOPOULOS, B, PIPERIDIS, S. 2000. *Term-based Identification of Sentences for Text Summarization*. In *Proceedings* of LREC 2000

GIOULI,V., KONSTANDINIDIS, A., DESYPRI, E., PAPAGEORGIOU, H. 2006. Multi-domain Multi-lingual Named Entity Recognition: Revisiting & Grounding the resources issue. In: *Proceedings of the 4th International Conference on Language Resources and Evaluation, Geneva, Italy*.

NISSIM, M., C. MATHESON, AND J. REID (2004). Recognizing Geographical Entities in Scottish Historical Documents. In *Proc. of the Workshop on Geographic Information Retrieval at SIGIR 2004.*

PAPAGEORGIOU, H., PROKOPIDIS, P., GIOULI, V., PIPERIDIS, S. *2000*. A unified tagging Architecture and its Application to Greek. *In: Proceedings of the 2nd International Conference on Language Resources and Evaluation, Athens, Greece.*

PROKOPIDIS, P., DESIPRI, E., KOUTSOMBOGERA, M., PAPAGEORGIOU, H., PIPERIDIS, S. (2005). Theoretical and Practical Issues in the Construction of a Greek Dependency Treebank. In *Proceedings of The Fourth Workshop on Treebanks and Linguistic Theories* (TLT 2005). Barcelona, Spain.

RAYSON, P., D. ARCHER, A. BARON, AND N. SMITH (2007). Tagging historical corpora – the problem of spelling variation. In *Proc. of Digital Historical Corpora, Dagstuhl-Seminar 06491*. 3-8. International Conference and Research Center for Computer Science, Schloss, Dagstuhl, Wadern, Germany.

RAYSON, P., D. ARCHER, A. BARON, J. CULPEPER, AND N. SMITH (2007). Tagging the bard: Evaluating the accuracy of a modern POS tagger on Early Modern English corpora. In *Proc. of Corpus Linguistics* 2007. Birmingham: University of Birmingham.

SAGER, J.C., DUNGWORTH, D., MCDONALD P. F. (1980). English Special Languages. Oscar Brandstetter Verlag KG - Wiesbaden.

Iris Hendrickx, Rita Marquilhas

From old texts to modern spellings: an experiment in automatic normalisation

We aim to tackle the problem of spelling variations in a corpus of personal Portugese letters from the 16th to the 20th century. We investigated the extent to which the task of normalising Portuguese spelling can be accomplished automatically. We adapted VARD2 (Baron and Rayson, 2008), a statistical tool for normalising spelling, for use with the Portuguese language and studied its performance over four different time periods. Our results showed that VARD2 performed best on the older letters and worst on the most modern ones. In an extrinsic evaluation, we measured the usefulness of automatic normalisation for the linguistic task of automatic POS-tagging and showed that automatic normalisation of spelling helps improve the performance of the POS-tagger.

1 Introduction

Our goal was to reduce the problem of spelling variations in the Portuguese CARDS-FLY corpus of personal letters written over a period of approximately 500 years, from the 16th to the 20th century. As the letters were written by a diverse group of authors, some of whom were semi-illiterate, and most of the manuscripts predate the first standardisation of Portuguese spelling which only took place in 1911, they contain many spelling variations. We wanted to make the corpus available for further linguistic research and also make it accessible to a wider community.

As the corpus is being prepared for research into language change, given the exceptional, near-spoken status of the texts, it was advisable to preserve the original spellings in a paleographic edition appropriate for research into sound variations and change, and for discourse analysis focussing on the specific behaviour of the social agents. However, a standardised corpus is essential for lexical and grammatical research, since this is the only format that can be given the necessary mark-up, such as POS tagging, semantic tagging, and syntactic parsing, enabling it to be used as an empirical tool for testing theories of lexical and syntactic change.

On the other hand, the rarity of the documents in question also make them valuable to the lay public, since they represent part of the country's cultural heritage, reflecting the everyday lives of ordinary people, especially those who suffered hardship. The 16th to 19th century corpus comprises original epistolary texts produced by servants, children, wives, lovers, thieves, soldiers, artisans, priests, political campaigners and many other social agents who fell foul of the Inquisition and the civil courts, two institutions that habitually seized personal correspondence for use as evidence. The letters in the

20^{th} century corpus come from personal collections compiled by the families of former soldiers, emigrants, and political prisoners. Editions intended for the lay public have the understandable obligation to provide readers with a clean text, free of distracting variations in spelling.

Our aim was therefore to add an additional orthographic component to the historical documents, which involved modernising the spelling. Modernising or normalising spelling is not an arbitrary step: if done manually it involves an enormous workload that can only be carried out by a specialist in historical linguistics and if done automatically, errors inevitably occur.

Here we investigate the extent to which the task of modernising Portuguese spelling can be accomplished automatically. We adapted VARD2 (Baron and Rayson, 2008) a well-studied statistical tool for normalising spelling, for use with the Portuguese language and studied its performance over four different time periods. The performance of this automatic normalisation tool was evaluated using intrinsic and extrinsic methods. Firstly, the automatically normalised text was compared with manually normalised text. Secondly, as an *evaluation of use*, the effect and usefulness of automatic spelling normalisation was evaluated in terms of the task of automatic grammatical tagging.

The performance of a POS-tagger used on a data set with the original, non-standardised spelling was compared with the effect of automatically normalised text and, as the upper bound, on manually normalised text. This research is similar to the work of Rayson et al. (2007) who investigated the usefulness of pre-processing with VARD2 for POS-tagging historical English and showed that normalisation does help to improve the performance of the POS-tagger.

The paper is structured as follows. In the next section we first discuss previous approaches to historical spelling normalisation. In Section 3 the experimental setup is explained, in particular the adaptation of the VARD2 tool to Portuguese in Section 3.1. Section 3.2 describes the historical corpus and provides additional background information on historical spelling changes in Portuguese. In Sections 3.3 and 5 the results of the experiments in normalisation and POS-tagging are presented and the conclusion appears in 6.

2 Related work

The problem of spelling variations in historical texts has been investigated from different perspectives and with different aims. Automatic text retrieval on historical data suffers severely from spelling variation and a common approach to this problem is not to modernise the full text collection, but to expand the search query to cover lexical variants (e.g. (Koolen et al., 2006; Hauser and Schulz, 2007; Ernst-Gerlach and Fuhr, 2007; Gotscharek et al., 2011)).

Other approaches attempt to modernise the spelling in the historical documents themselves. The VARD tools were developed for corpus linguistic research into Early Modern English. The original VARD tool consisted of a list of manually created mappings between historical variants and their modern versions. VARD2 (Baron and

Rayson, 2008) has an additional module that can search for variants and mappings. In the work of Rayson et al. (2005) the ability of VARD to detect spelling variants and suggest the correct modern spelling is compared with two commercial spelling correctors, MS-Word and Aspell, showing that VARD works better for historical texts since it detects fewer false positives. VARD2 will be discussed in more detail in Section 3.1.

Kestemont et al. (2010) describe an automatic approach to normalise the spelling in Middle Dutch Text from the 12th century. In this case however, they chose not to convert historical word forms to their modern counterparts, but to their modern lemma. They used machine learning to discover how to transform one spelling variant into another to resolve intra-lemma variation.

Several studies of spelling variation in Portuguese historical documents have been conducted and we were grateful to be able to re-use some of the resources already developed for historical Portuguese. We will briefly discuss this previous work and, in Section 3.1, explain how we used the available resources.

The Historical Dictionary of Brazilian Portuguese (HDBP) is constructed on the basis of a historical Portuguese corpus of 1,733 texts and approximately 5 million tokens. As there was no standard spelling at the time (16th to 19th century), it is not easy to create lexicographic entries on the basis of the corpus or produce reliable frequency counts. A rule-based method was therefore developed for the automatic detection of spelling variation in the corpus (Giusti et al., 2007). The HDBP researchers compared their automatic variant detection method with Agrep[1] using a small test set and showed that their method was more precise, whereas Agrep had much better recall. These experiments also led to a spelling variants dictionary containing approximately 30K clusters of variants.

Another resource available for Portuguese is the Digital Corpus of Medieval Portuguese (CIPM- Corpus Informatizado do Português Medieval)[2] which covers the 12th to the 16th century and counts around 2 million words. Rocio et al. (2003) describe how they annotated part of the CIPM with linguistic information such as POS-tags, morphological analysis and partial parse information. They did not proceed with modernisation but used automatic tools on the historical data as such, followed by a manual correction phase.

The Tycho Brahe Parsed Corpus of Historical Portuguese[3] is an electronic corpus of historical texts with prose from different text genres from the Middle Ages to the Late Modern era. The TBCHP contains 52 source texts but not all of them are annotated in the same way. Some of the texts maintain the original spelling variations, while in other texts, intended for part-of-speech and syntactic annotation, the spelling was standardised.

[1] Agrep: http://www.tgries.de/agrep/
[2] CIPM corpus: http://cipm.fcsh.unl.pt/
[3] TBCHP: http://www.tycho.iel.unicamp.br/~tycho/corpus/en

3 Data and Methods

This section describes the adaptation of the VARD2 spelling standardisation tool for use with Portuguese, the corpus in question, which is a historical corpus of private letters written in Portuguese, and the experimental setup for normalisation and POS-tagging.

3.1 VARD2 for Portuguese

The aim of this study was to evaluate the performance and usefulness of the VARD2 tool for historical Portuguese. As mentioned in Section 2, VARD2 was developed for historical forms of English. It combines several resources to detect and replace spelling variants with a normalised form and, where possible, we adapted every module for use with the Portuguese language.

VARD2 uses a modern lexicon, a spelling variants dictionary list that matches variants against their modern counterparts, a list of letter replacement rules and a phonetic matching algorithm to predict normalised candidates for each variant detected, using an edit distance algorithm to determine the most likely candidate. The tool can be configured, since each module can be assigned a certain weight and can be individually configured in favour of recall or precision. When training the tools on a specific data set, new words and variants are added to the lists and the module weights are adapted accordingly. We replaced the modern frequency lexicon, spelling variants dictionary and letter replacement rules with Portuguese versions.

The following Portuguese resources were used to convert VARD2 to Portuguese. The Multifunctional Computational Lexicon of Contemporary Portuguese[4] contains 26,443 lemmas corresponding to 140,315 tokens. Only lemmas with a minimum lemma frequency of 6 were extracted from a sample from the contemporary Portuguese corpus CRPC (a corpus sample of 16,210,438 words) for inclusion in the lexicon. We filtered this lexicon to suit our purpose and removed all multi word expressions (for example "sem abrigo") and words with non-conventional spellings. The frequency counts for homonyms were reduced to one count for each particular word form. The word frequency list that we used has a total of 127,891 word forms.

We created a list of letter replacement rules based on the rule set described in detail by Giusti et al. (2007) and developed for historical Brazilian Portuguese. The rules encode accent changes, spelling changes, such as 'x' to 'ch', and letter combinations that are no longer used in modern Portuguese, for example 'th', 'ph', 'aes', or double consonants such as 'dd', 'ff' etc.

As mentioned in Section 2 a spelling variants dictionary[5] was created based on the Historical Corpus of Brazilian Portuguese. Giusti et al. (2007) have created a corpus-based tool to automatically generate and test rewrite rules that cluster spelling variants together. These groups are clustered around one common word form, the so-called head

[4] Lexicon is available for download at: http://www.clul.ul. pt/en/resources/88 -project- multifunctional- computational -lexicon- of-contemporary- portuguese-r

[5] BP spelling variants dictionary is available: http://www.nilc.icmc.usp.br/nilc/projects/hpc/

word of the cluster. As the original dictionary consists of clusters of spelling variants, and we needed a list of one-to-one mappings between variants and their modernized counterparts to integrate into the VARD2 tool, this variants dictionary had to be converted to meet our needs. As a logical choice, we initially mapped each variant in a cluster to the head word of the cluster. However, the head word is not always the modern or most frequent word form, although this is usually the case, and this implies that these automatic mappings sometimes lead to errors. Here is an example of a cluster from the spelling variants dictionary:

tambem	(12211)	
	tambem	(9002)
	também	(3160)
	tanbem	(47)
	ttambem	(1)
	ttanbem	(1)

The modern word form of this cluster is the accented version *também* (En: “also") and the cluster head *tambem* does not occur in the current modern lexicon. To prevent mappings between variants and non-modern word forms, every head word was checked to determine whether it occurred in our modern lexicon. If it did not, the most frequent word form in the cluster that did occur in the modern lexicon was selected. Closer inspection of the resulting variants list showed that this automatic mapping of variants and head words can still lead to some errors in cases where the head word occurs in the modern lexicon but is not the most obvious candidate, for example the “aviam -haviam" cluster. A manual correction phase of (at least the most frequent) variant clusters would certainly improve the variant list. We did not alter the phonetic matching algorithm of VARD2, but in future work we would like to evaluate this module for Portuguese.

3.2 The corpus

As mentioned in the Introduction, the CARDS-FLY corpus[6] was compiled from a rare collection of documents written by a variety of social agents living in difficult times. Later, disregarding the authors' intentions, the letters found their way into several archives instead of being destroyed, as might be expected in the case of everyday private papers. The manuscripts from the 1500-1800 period are personal letters that were, unusually, retained as part of religious legal proceedings, as evidence used by the Inquisition in heresy trials. Those from the 19th century were also used as legal evidence, this time in criminal cases heard by the Portuguese Royal Appeal Court (abolished in 1833) and civil cases that appeared before a regional court in the north-east. The 20th century letters date from 1901-1974 and consist mainly of manuscripts, together with some typed scripts, sent or received by soldiers who fought in World War I or in the Portuguese Colonial War, emigrants of Portuguese origin, and prisoners held by the political police. They were mostly kept in family archives and sometimes donated

[6] CARDS-FLY corpus: http://alfclul.clul.ul.pt/cards-fly/

to public documentation centres. A few others were archived by propaganda and censorship institutions. The whole collection is being processed electronically (involving transcription into XML-TEI file format) so that it can function both as a digital archive available to the general public and as a corpus intended for historical, linguistic and sociological research.

For the sake of historical accuracy, the letters were divided into different time periods, taking into account the serialisations already proposed for the history of Portuguese language. Not all Portuguese historical linguists working on serialisations agree on the chronology for milestones in language change as Martins2002 explains. However, they do agree on the convenience of distinguishing between Old Portuguese, Classical Portuguese and Modern Portuguese, following the traditional classification in general history that distinguishes between the Middle Ages (from the end of Antiquity up until the Renaissance), the Early Modern Age (up until the liberal Revolutions), and the Contemporary Age.

Our corpus contains sources for the study of both Classical and Modern Portuguese and a dividing line therefore had to be drawn between the two in the early 19th century. However, a second milestone was needed since a great deal of debate surrounds the Classical Portuguese period with regard to innovations in European Portuguese, especially syntax, vis-à-vis Brazilian Portuguese. In the current state of the art, the beginning of the 18th century represents such an important milestone (Galves and de Sousa, 2005) and we therefore subdivided the Classical letters into those dating from 1500-1700 and those dating from 1701-1800. As for the Modern letters in our corpus from the period 1801-1974, on the one hand we had to take into account the fact that we were dealing with written texts that generally adopted non-standard spellings, but also the fact that the Republican decree of 1911 had instituted the first national spelling agreement (Castro et al., 1987). Prior to this, despite several debates, there was no standard way of spelling Portuguese and the discussion was very much divided between the 'Sonics' and the 'Etymologists'. The Sonics fought for phonographic spelling using diacritics (matéria versus materia) and the absence of learned consonantal clusters of Greek or Latin origin (catedral versus cathedral). The Etymologists advocated the reverse, which had a better established tradition in Portuguese writing.

In terms of the division of our corpus into time spans, we considered that the effects of the 1911 spelling reform would only be evident by 1931, when the children who had started grammar school in 1911 had already become adults using correspondence as a common interactive practice. The Modern letters in our corpus were therefore divided into two groups, namely 1801-1930 and 1931-1974.

The current version of the corpus contains 1,802 letters from which we randomly selected a subset of 200 letters for the experiments, respecting the frequency division of the different centuries. This subset was manually annotated by a linguist to be used as training and evaluation material. The texts were tokenised (punctuation was separated from words) and we converted any names in the text into the string 'NAME'. The names in the modern letters were already anonymous and, following this conversion, all the documents had the same form of representation for names. For the purposes of our

experiments, this data set was split into 100 letters for training the VARD tool, and 100 for the evaluation set.

Table 1 presents the statistics for the evaluation set, showing the number of letters, tokens and the average number of spelling normalisations made by the human annotator. As might be expected, more corrections per letter were found in the oldest letters, which were also the shortest. In the modern letters only 4% of the tokens were normalised. The 18^{th} century letters are remarkably long in comparison with the other letters. One possible explanation for this is that, on the one hand, the corpus contains more letters from the 18^{th} than the 16^{th} and the 17^{th} centuries and so there is a greater likelihood of obtaining long letters. In addition, the lower-classes were gradually becoming literate (or semi-illiterate) during the 19^{th} and $2o^{th}$ century and would therefore have tended to restrict themselves to short letters dealing with urgent matters.

Table 1: Statistics for the evaluation set of 100 letters, divided into the four time periods. # Tok/file shows the average number of tokens per letter, '#Norm/file' the average number of manual spelling corrections per letter and '% Norm/tok' is the percentage of all tokens that is normalised.

Period	Files	Tokens	#Tok/file	# Norm/file	% Norm toks
1500-1700	10	2262	226.2	56.9	25.2
1701-1800	28	13913	496.9	120.8	24.3
1801-1930	43	14343,	333.6	60.7	18.1
1931- 1974	19	6817	358.8	16.1	4.2

Even though the letters from the final period 1931-1974 basically use modern spelling, an average of 16 spelling changes per letter can still be observed. The type of spelling changes here are mostly due to hypercorrections associated with the inner logic of Portuguese 'sonic' orthography, which contains many inconsistencies in grapheme -phoneme correlations.

3.3 Experiments in normalisation

After discussing the VARD2 tool used for standardisation and the corpus in detail, we now describe exactly how this tool was applied to the data and explore how well it performed with the Portuguese corpus. Our aim was also to investigate whether it was more practical to have one spelling modernisation tool for the complete Early Modern-Contemporary period, or separate tools for shorter periods and whether the advantage of having specialised tools for each period outweighed the disadvantage of less data, given that each specific tool would be trained using a smaller set.

To evaluate the performance of the tool, we compute accuracy, recall, precision and F-score for the words (excluding punctuation marks) in the test data. True positives (TP) refer to cases in which there was a spelling variant in the text and the modern variant was correctly predicted by the tool. False positives (FP) involve cases in which

the tool erroneously predicted a spelling variant, and false negatives (FN) are the spelling variants that were not detected by the tool. True negatives (TN) are the remaining words correctly predicted as 'not a spelling variant'. We compute accuracy, recall (R), precision(P) and the harmonic mean (F-score) between recall and precision (van Rijsbergen, 1979) as follows:

$$Accuracy = TP + TN/(TP + TN + FP + FN) \quad (1)$$

$$P = TP/(TP + FN), R = TP/(TP + FP), F - score = 2 * P * R/(P + R) \quad (2)$$

As a first step the VARD2 tool was configured for the data set. VARD2 has two parameters that need to be set: the first establishes the weighting given either to recall or to precision, and the second is the replacement threshold which decides whether a potential variant should be replaced with the equivalent modern candidate. We set the first parameter to assign equal weight to recall and precision. To determine the value for the second parameter, we ran a series of experiments with different thresholds. We divided the training set in 80 letters for training and 20 as a development set. We tested the following settings: 1, 5, 10, 20,.. 90 for this threshold. The best F-score, 65.5%, for the development set was obtained with parameter 1. All the parameters tested between 5 and 40 obtained a score of 64% and a gradual decrease in performance was observed when the parameters were increased to values above 50. The best performance was obtained with the threshold set to 1 and this setting was therefore used in all further experiments.

When examining the errors made by VARD2 in the development set, one specific error stood out: the letter q is an abbreviation and is almost always standardised to que. As q itself is listed as a valid word in the modern lexicon, it was never detected as potential spelling variant. Since this q occurs very frequently and many errors were due to this mismatch, a rule was added to the tool to normalise each q to *que*.

In order to investigate whether it was better to have specialised tools trained separately for each time period, or one tool trained using the full training set for the whole period, we trained five versions of VARD2, one for each individual period and one for the whole period. In the first set we applied the VARD2 trained on all 100 training letters to the full test set of 100 letters, and then to the four individual test sets whose properties are listed in Table1.

4 Results of normalisation

The results from these experiments can be seen in Table 2. The second row shows the evaluation when training on 100 letters and testing on 100 letters. The VARD2 tool has a much higher precision than recall. The tool detected around 2/3 of the spelling variations (61% recall), and, if a variant was found, in 97% of the cases it was correctly changed to its modern counterpart. Row 3-6 of the table focuses on the performance of VARD2 in each specific time period. In the latest time period, 1931-1974, high accuracy and a remarkably low recall and F-score can be observed. In calculating accuracy,

true negatives are counted, whilst the other measurements focus solely on the spelling variants, which comprise a much smaller set for this particular period. As shown in Table 1, these letters had the fewest spelling normalisations, amounting to only 16 changes per document on average, whereas the letters from the other periods had an average of at least 50 per document.

Contrary to expectations, the highest precision and F-scores were found in the oldest letters. Better results would have been expected for the period 1801- 1930, since we had the largest amount of training and testing material for this period. One explanation may be found in the nature of the data set, since this was a time when the lower-classes were becoming semi-literate and it therefore contains a set of letters produced by people who would never have put pen to paper in earlier times and who produced many creative misspellings that are extremely difficult to predict.

Table 2: Precision, recall and F-score for VARD2 trained on 100 training letters in the evaluation set of 100 letters, divided over the four time periods.

time period	Acc	R	P	F-score
total	91.92	61.10	97.21	75.03
1500 -1700	91.75	68.72	98.74	81.04
1701 -1800	90.77	65.95	97.72	78.75
1801- 1930	91.06	56.0	96.81	70.96
1930 - 1974	96.46	35.23	87.5	50.24

In the second round of experiments we trained VARD2 on the time-specific subsets of the training set and tested on the same test subsets. The results are shown in Table 3. Again it can be observed that the best F-score performance occurs in the older letters and the worst in the most modern ones. When the results in Table 2 are compared with the results in Table 3, a slight improvement can be seen in the two oldest time periods and a decrease in the more modern periods, both in terms of accuracy and F-score. Recall is mainly affected by the change in training set, showing an increase for the oldest data. For the two more modern data sets precision slightly increases at the cost of the recall.

Table 3: Precision, recall and F-score for specialised VARD2 tools, trained and tested separately for the four time periods.

Period	Acc	R	P	F-score
1500-1700	92.52	71.88	98.55	83.13
1701-1800	91.81	70.24	97.49	81.65
1801-1930	90.61	53.45	97.08	68.94
1931-1974	96.32	31.88	87.96	46.80

In order to investigate the spelling variants that could not be corrected automatically, we analysed the most frequent errors in the final round of experiments for each of the different time periods. The most frequent error was the failure to recognise the traditional spelling of 'um' with 'h-' (43 times for the 1701-1800 period and 52 times for 1801-1930). VARD2 does not recognise this as a spelling variant, since hum is listed in the modern lexicon. For all periods we clearly observe that many spelling variants originate from the fact that writers find it difficult to master the diacritics. Furthermore, the older groups of letters have a large number of archaisms (e.g 'inda' and 'cousa' in the top 10 errors for the period 1500-1700) that are no longer used in current spelling, but are erroneously part of the VARD2 modern lexicon list and are therefore not recognised as spelling variants. The older letters also have a large percentage of abbreviations (e.g. v., va. and etcra.) which are difficult to recognise automatically, unlike the modern ones. Grammar teachers condemn the use of abbreviations, which they label bad style, and this 'lesson' seems to have been learned by the 20th century authors writing between 1931 and 1974.

Major trends involving confusion between different spellings were observed within the different periods. For the period 1500-1700, difficulty in mastering the etymological use of s/c/ss for the single sound [s] was evident, and , for the period 1701-1800, the etymological use of z/s for the single sound [z], whilst in the period 1801-1930 the phonetic spelling of 'i' for 'e' frequently occurs.

5 POS-tagging

Our other goal was to quantify the effect of text normalisation on the application of NLP tools such as a POS-tagger. We trained one POS-tagger on normalised texts from the Portuguese Tycho Brahe corpus and tested it both on non-normalised text and normalised text. The Tycho Brahe corpus contains 19 normalised and POS-tagged texts with a total of approximately 40K sentences and 891K tokens. The POS-tag set contains 280 different tags that express specific information such as gender and number.

We created an automatic POS-tagger by training MBT (Daelemans et al., 2007) on the 19 texts from Tycho Brahe. MBT is a memory-based machine learning system specifically developed to handle sequence labelling such as POS-tagging. When assigning POS tags to words, the previous labelings can be very informative in terms of the current decision: for example if the previous word is labeled as determiner, the current word is likely to be an adjective or noun. MBT takes its previous decisions into account when labelling words.

We tested the POS-tagger on three versions of our corpus of letters: the original unnormalised text, the text automatically standardised using VARD2 (trained on 100 letters), and on the gold standard of manual annotation. The results of these experiments are shown in Table 4. The first column shows the POS-tagging accuracy for all tokens (including punctuation marks) in the 100 letters from the test set.

The major source of errors made by a POS-tagger is unknown words that the tagger has not encountered previously in the training set. In the case of known words, the

tagger assumes that it knows (on the basis of the training set) which labels are applicable for a certain word and chooses one label from this small sub set. For example the word *via* can be a verb or noun and the POS-tagger only needs to choose between these two. However, for an unknown word, the POS-tagger needs to consider all 280 possible tags. Therefore the accuracy rate for unknown words is lower than for known words as demonstrated in the last two columns of Table 4.

Table 4: POS-tagger accuracy for the evaluation set of 100 letters, based on the original non-normalised text, text automatically normalised by VARD2, and the gold standard created by manual annotation.

Type	Total	Unknown	Known
# tokens	37,335	5,869	31,466
Original	76.86	42.34	87.06
VARD2	83.41	47.57	90.1
Gold	86.578	49.11	91.94

6 Conclusions

We have presented an approach to standardising the spelling of historical Portuguese and demonstrated how to adapt the statistical VARD2 normalisation tool for the Portuguese language by re-using several Portuguese resources currently available. Having split the data set into 4 time periods, it was observed that VARD2 performs best on the older letters and worst on the most modern ones. We also investigated whether it was more useful to have specialised normalisation tools for each time period, or whether the tool benefits more from one large training set covering the whole time period 1500 to 1974. The results show that for the Classical period the advantage of a specialised tool outweighs the smaller amount of data. Conversely, for the Modern period a tool trained using a larger, diverse data set works better. In terms of extrinsic evaluation, we measured the usefulness of automatic normalisation in terms of the more complex linguistic task of automatic POS-tagging and showed that automatic normalisation of spelling helps improve the performance of the POS-tagger. In all periods, the letter writers can be seen to struggle with two problems: i) how to master etymological spellings without knowing Latin, Greek, or Old Portuguese, ii) how to master phonographic spellings if they never obey purely phonetic facts, given that phonological (segmental and suprasegmental), morphological and lexical information always influences apparently phonographic principles to some extent. Nevertheless, the effectiveness of the 20$^{\text{th}}$ century Portuguese spelling reform can be clearly observed in our corpus, as well as its 'Sonic' profile: many etymological principles have clearly been abandoned. This trend was recently reinforced when all the Portuguese speaking countries in the world adopted a new more phonographic reform, celebrated in a 1990 treaty and implemented in Portuguese public education in 2011.

References

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

Castro, I., Duarte, I., and Leiria, I. (1987). *A Demanda da Ortografia Portuguesa*. Edições João Sá da Costa, Lisbon, Portugal.

Daelemans, W., Zavrel, J., Van den Bosch, A., and Van der Sloot, K. (2007). MBT: Memory-Based Tagger, version 3.1, Reference Guide. Technical report, ILK Series 07-08.

Ernst-Gerlach, A. and Fuhr, N. (2007). Retrieval in text collections with historic spelling using linguistic and spelling variants. In *Proceedings of the ACM/IEEE-CS conference on Digital libraries*, pages 333–341.

Galves, C. and de Sousa, M. C. P. (2005). *Romance Languages and Linguistic Theory 2003*, chapter Clitic Placement and the Position of Subjects in the History of European Portuguese, pages 97–113. Current Issues in Linguistic Theory 270. John Benjamins, Philadelphia.

Giusti, R., Candido, A., Muniz, M., Cucatto, L., and Aluísio, S. (2007). Automatic detection of spelling variation in historical corpus: An application to build a Brazilian Portuguese spelling variants dictionary. In *Proceedings of the Corpus Linguistics Conference*.

Gotscharek, A., Reffle, U., Ringlstetter, C., Schulz, K., and Neumann, A. (2011). Towards information retrieval on historical document collections: the role of matching procedures and special lexica. *International Journal on Document Analysis and Recognition*, 14:159–171.

Hauser, A. W. and Schulz, K. U. (2007). Unsupervised learning of edit distance weights for retrieving historical spelling variations. In *Proceedings of the First Workshop on Finite-State Techniques and Approximate Search*, pages 1–6, Borovets, Bulgaria.

Kestemont, M., Daelemans, W., and De Pauw, G. (2010). Weigh your words - Memory-Based Lemmatization for Middle Dutch. *Literary and Linguistic Computing*, 25:287–301.

Koolen, M., Adriaans, F., Kamps, J., and de Rijke, M. (2006). A cross-language approach to historic document retrieval. In *Advances in Information Retrieval: Proceedings of ECIR 2006*, volume 3936 of *LNCS*, pages 407–419. Springer Verlag, Heidelberg.

Rayson, P., Archer, D., Baron, A., Culpeper, J., and Smith, N. (2007). Tagging the Bard: Evaluating the Accuracy of a Modern POS Tagger on Early Modern English Corpora. In *Proceedings of the Corpus Linguistics Conference (CL2007)*, University of Birmingham, UK.

Rayson, P., Archer, D., and Smith, N. (2005). VARD versus Word: A comparison of the UCREL variant detector and modern spell checkers on English historical corpora. In *Proceedings from the Corpus Linguistics Conference Series*, volume 1, Birmingham (UK).

Rocio, V., Alves, M. A., Lopes, G. P., Xavier, M. F., and Vicente, G. (2003). *Automated creation of a Medieval Portuguese partial Treebank*, volume 20 of *Language and Speech Series*, pages 211–230. Kluwer, anne abeillé edition.

van Rijsbergen, C. (1979). *Information Retrieval*. Buttersworth, London.

Stefan Höfler, Michael Piotrowski*

Building Corpora for the Philological Study of Swiss Legal Texts

We describe the construction of two corpora in the domain of Swiss legal texts: The DS21 corpus is based on the Collection of Swiss Law Sources and contains historical legal texts from the early Middle Ages up to 1798; the Swiss Legislation Corpus (SLC) is based on the Classified Compilation of Swiss Federal Legislation and contains all current Swiss federal laws. The paper summarizes the key properties of both corpora, discusses issues encountered while building them, and outlines some applications.

1 Introduction

Legal texts are a fruitful object of study for the humanities. For historians, they represent a crucial source of information on the distribution of power in societies past and present, the values holding these societies together, their methods of resolving conflicts, and on the ways in which both the distribution of power and the underlying values have changed over time. For linguists, legal texts constitute a special case of highly conventionalized language use that strives, and often struggles, to find an optimal balance between rigorousness and flexibility, formality and understandability, that is, between expressing authority and yet being grounded in the everyday life of those affected by it. In fact, legislative texts do not merely describe, but create law.

The present paper introduces two annotated corpora of Swiss legal texts that have been compiled to provide scholars in the humanities with additional means to study this domain and to support the development of domain-specific natural language processing (NLP) tools. Both corpora are based on preexisting text collections, which were compiled according to criteria that are specific to this particular domain: The DS21 corpus is based on the *Collection of Swiss Law Sources* and the Swiss Legislation Corpus (SLC) is based on the *Classified Compilation of Swiss Federal Legislation*. The DS21 corpus is a corpus of historical legal texts, while the SLC comprises texts that constitute contemporary statutory law. We describe the range of texts contained in either corpus (sections 2.1 and 3.1 respectively), detail the characteristics of these texts (sections 2.2 and 3.2), and discuss some of the implications that such domain-specific features have for the automatic annotation of the texts (sections 2.3 and 3.3). The presentation of either corpus is followed by a brief survey of the range of humanities research it facilitates (sections 2.4 and 3.4). We conclude with a summary of the most important properties of the two corpora (Table 2).

*Authors are given in alphabetical order. MP is responsible for the DS21 corpus (partly funded under SNSF grant no. 124427); SH for the SLC (funded under SNSF grant no. 134701).

2 A Corpus of Historical Legal Texts

2.1 The Collection of Swiss Law Sources

The *Collection of Swiss Law Sources* is an edition of historical *sources of law* created on Swiss territory from the early Middle Ages up to 1798 (the downfall of the *ancien régime* in Switzerland). The Collection employs a broad definition of *law source* and includes not only acts, decrees, and ordinances, but also indentures, administrative documents, court transcripts, and other types of documents. It is edited by the Law Sources Foundation, which was established in 1894 specifically for this task. Since then the Foundation has edited and published over 100 volumes containing more than 60,000 pages of source material and commentary.

The primary sources are manuscripts in various regional historical forms of German, French, Italian, Rhaeto-Romanic, and Latin, which are transcribed, annotated, and commented by the editors. The critical apparatuses are in modern German, French, or Italian. Each volume contains an index of persons and places and a combined subject index and glossary.

The Collection organizes the sources by cantons and then generally subdivides them by areas of jurisdiction, such as towns or bailiwicks. At the time of this writing, it covers 17 of the 26 Swiss cantons to different extents. The edition of the Collection of Swiss Law Sources is an ongoing project and further volumes are in preparation.

Historians and historians of law are currently the primary users of the Collection, but it is also an important source for the Swiss-German Dictionary ("Idiotikon"). See Gschwend (2008) for a description of the Collection from a historian's point of view.

From 2009 to 2011 the Swiss National Science Foundation funded the digitization the Collection. As a result, the complete Collection is now available online in facsimile form[1]. The tables of contents were digitized using optical character recognition (OCR) with extensive post-editing to create an XML registry of the titles and the dates of creation of all texts in the Collection.

2.2 The Corpus: DS21

The availability of online facsimiles of all the volumes of the Collection represents a significant advance. It would, nevertheless, be desirable to have the full text of the historical sources and the apparatuses available in digital form.

During the retrodigitization project we discovered that digital typesetting data (FrameMaker 3 and 6 files) still exists for the 22 latest volumes. This provides us with a sizeable collection (about 4 million running word forms) of medieval and early modern texts free from errors introduced by digitization (whether OCR or retyping). We refer to this digital subset of the Collection as DS21.

DS21 contains volumes from ten cantons representing most linguistic and geographic regions of Switzerland. The subset also covers the full period of time documented by

[1] `http://ssrq-sds-fds.ch/online/`

Volume ID	Canton	Primary Language(s) of the Sources	Period Covered	Pages	Texts
SSRQ AG II 9	Aargau	German	1301–1798	735	415
SSRQ AG II 10	Aargau	German	1303–1797	735	530
SSRQ AR/AI 1	Appenzell	German	1409–1632	658	10
SSRQ BE I/13	Berne	German	1230–1796	1143	468
SSRQ BE II/9	Berne	German	1267–1797	992	690
SSRQ BE II/10	Berne	German	1277–1797	1191	808
SSRQ BE II/11	Berne	German	1256–1795	1305	894
SDS FR I/2/6	Fribourg	French, German	1296–1795	582	639
SSRQ GR B II/2	Grisons	German	1289–1832	1403	850
SSRQ LU I/1	Lucerne	German	1178–1501	592	436
SSRQ LU I/2	Lucerne	German	1426–1463	481	476
SSRQ LU I/3	Lucerne	German	1425–1489	731	465
SSRQ LU II/2	Lucerne	German	1301–1799	2428	692
SSRQ SG I/2/3	St. Gallen	German	754–1797	1173	518
SSRQ SG II/1/1	St. Gallen	German	1302–1601	492	16
SSRQ SG II/1/2	St. Gallen	German	1452–1701	538	300
SSRQ SG II/2/1	St. Gallen	German	1229–1799	1184	537
FDS TI A/1	Ticino	Italian	1286–1799	401	623
SDS VD B/2	Vaud	Latin, French	1211–1797	622	448
SDS VD C/1	Vaud	French, German	1530–1797	971	537
SDS VD C/2	Vaud	French	1539–1770	925	21
SSRQ ZH NF II/1	Zurich	German	1307–1794	522	318
				Total 19,804	10,691

Table 1: Composition of the DS21 collection. The languages given in the table refer to historical variants of these languages from the periods indicated.

the Collection: with the oldest text being from 754 and the most recent one being from 1832, it spans 1078 years. We therefore believe DS21 to be a good sample of the legal documents from the relevant period preserved in Swiss archives. Table 1 gives details of the composition of DS21.

We are now working on the conversion of the FrameMaker files of DS21 into a form that is usable for historians, linguists, and other researchers; more specifically, we want to create a corpus marked up according to the TEI P5 guidelines[2]. Several steps are required to reach this goal: First, the FrameMaker files must be converted into an open format; then the markup contained in these files must be regularized; the regularized markup subsequently enables the inference of the semantics of marked-up elements and the upconversion into a format that makes the semantics explicit. After this, links between textual elements—in particular between source text and annotations—can be detected and also made explicit using TEI markup.

To this end, we have developed a multi-stage conversion process for automatically converting FrameMaker files into valid TEI documents. The FrameMaker files contain

[2] http://tei-c.org/

only visually oriented markup, i.e., text is marked up as bold, italics, superscript, etc., not as title, date, or apparatus text. The first step is the conversion of the FrameMaker files into valid XHTML files with CSS stylesheets that closely mirror the layout and formatting of the FrameMaker documents (Piotrowski, 2010a). This conversion is more challenging than it may appear at first sight and requires deep processing of the FrameMaker files, in particular the tracking of style inheritance and font changes. Furthermore, while superficially similar, the conceptual models of FrameMaker and CSS differ significantly in details.

The bodies of the books are then converted to TEI, while the indices are converted into an application-specific XML format. At this point, the TEI markup is still similar to the XHTML markup, but the differences between XHTML and TEI require certain structural changes; for example, the `<br/>` element in XHTML marks the end of a line, whereas the `<lb/>` element in TEI marks the beginning of a line.

The upconversion primarily relies on the normalized typographic information produced in the previous conversion steps: it allows the automatic identification of article headings, footnote markers and the corresponding footnote text, and source text and commentary.

2.3 Automatic Annotation

DS21 is based on scholarly editions of historical texts: Each text is accompanied by a modern-language summary or title, the date and place of creation (as far as they are known), a description of the original physical document (archive location, writing material, measurements, etc.), and typically notes and a critical apparatus (describing additions, deletions, alterations, etc.). Figure 1 shows a text from DS21 as it appears in the printed version.

Together, the apparatuses, indices and glossaries form a rich annotation of the source texts. In contrast to summaries and apparatuses, which are located directly before and after each text, indices are stored separately in printed books, and while the index entries point to the relevant points in the text (by giving page and line numbers), there are no pointers *from the text* to the indices. In the TEI-encoded version, however, we want to have the information from the indices integrated into the text.

The FrameMaker files were used for printing the books and thus have the exact same line breaks and pagination. Throughout the conversion process this information is preserved, so that it is possible to identify the locations—with a precision of about one line or 10–15 word forms—to which index entries refer.

We have developed a tool that reads a TEI document and an index of persons and places (in an intermediate XML format), analyzes the index entries and inserts a `<persName>` or `<placeName>` element at the beginning of the corresponding line in the TEI document. For our prototype volume the element is then manually moved to the exact position in the text, so that it encloses the occurrence of the name. As index entries typically list some of the most frequent spelling variants (e.g., “Ausserrhoden: Äussere Rhoden, Außer Rhoden, Ußeren Roden, Ussern Rooden, Usroden, Uß

Roden, Ussroden"), we intend to have the program identify the occurrence in the text automatically.

Glossary entries are linked to the lines to which they refer, as the identification of the exact referent of a glossary entry is often hard, e.g., due to inflectional variation or different word order. All glossary entries referring to a particular text are used to generate keywords for this text, complementing the full-text search that is always possible for electronic texts but difficult for historical texts in different languages. We are currently developing a controlled vocabulary based on the combined glossaries of DS21.

Automatic linguistic annotation of the historical texts in DS21 would be nontrivial; it therefore currently does not contain linguistic annotation beyond what is provided by the glossaries. However, future annotation with further linguistic information is not precluded. Even though not all of existing annotation from the indices and glossaries is directly useful for linguistic purposes, it is nevertheless important to preserve this information, in order not to exclude other uses of DS21. While most historical corpora are based on scholarly editions, they usually do not preserve the critical apparatuses; Boschetti (2007) points out that, for the philologist, even the text of an authoritative edition has "no scientific value without the apparatus."

2.4 Applications

Work on the DS21 corpus is still ongoing, so we cannot report on actual uses of this corpus yet. However, we want to outline some potential uses of the corpus.

First, it will be possible to use the corpus for all types of research for which one would formerly have used the printed volumes or the digital facsimiles of the Collection. Typical research questions come from legal, economic, and social history. Even if a scholar does not use any new research methods, access to the texts will be easier and faster, and digital full text generally offers users more convenience, e.g., text can easily be copied. However, the corpus will also help to investigate research questions which the printed indices were not designed to support, it will facilitate studies that span several of the traditional volumes, and it will make it possible to explore topics orthogonal to the traditional regional organization of the Collection.

DS21 contains much information that is relevant for completely different fields of research besides legal history. For example, it contains many historical documents mentioning animals or animal products, such as regulations of hunting, fishery, butchery, and livestock. For archeozoology, these documents may represent valuable evidence complementing archeological findings and documenting, e.g., the presence of certain animals in a certain region. In fact, an archeozoologist has manually analyzed the indices of one printed volume to find texts to animals and animal products and compiled the species of fish mentioned documented for Lake Murten. However, this was tedious work, as the printed indices were not designed for this type of questions. The DS21 corpus will make such new uses of the texts much easier and will allow a multitude of new, as yet unanticipated, uses of the information it contains.

144. Strassenverordnung

1545 September 7* (menntags an unnser liebenn frouwen abend gepurt). *Rapperswil

Witter der straßen halb, so jn der statt verleitt, ist erkentt, wo stein, misthůffen, schiter oder holz, was da schådlich so jn straßen, alles dannen than und niemands sol verschont werden, sol stattknecht versorgen und sagen, darzů ist wachtmeister geordnot.

Rats-, Gerichtsprotokolle: *StadtARap, Bd. B 1, fol. 82r, Pap. 21,5/22 x 32 cm.*

Bemerkung

1566 Mai 14 (zinstags nach Panngratzi). *Rapperswil:* Joner unnd Bußkilcher sonnd einandren den weg zů Bußkilch uff der allmendt machen, damitt jederman tags und nachts wandlen mögen etc. unnd sonnds angends thůn etc. *(Rats-, Gerichtsprotokolle: StadtARap, Bd. B 2, S. 110). – Vgl. auch Ratsprotokolle: StadtARap, Bd. B 28, S. 204; B 31, S. 420; B 32, S. 515; 516; B 33, S. 151; B 39, S. 110; B 51, S. 205; Nr. 203.*

Figure 1: Example of a text contained in DS21: Ordinance on streets by the city of Rapperswil from September 7, 1545 (reproduced from Rechtsquellenstiftung des Schweizerischen Juristenverbandes, 2007, p. 407)

Second, DS21 will enable new modes of access, for example geographic browsing. We have created a prototype system that offers users an interactive map on which all places mentioned in the texts are marked (see Piotrowski, 2010b). By clicking on a place marker, the titles of the sources associated with a place are listed; clicking on a title brings up the corresponding source for reading. Other non-textual access modes could be based on persons or dates.

Third, the annotated electronic text facilitates interlinking with complementary resources, e.g., the *Deutsches Rechtswörterbuch* (DRW), the Swiss-German Dictionary, `e-codices.ch`, or `monasterium.net`.[3]

Finally, the corpus will be invaluable for research into NLP for historical languages, especially for research on normalization of spelling variants, as the glossaries provide lemmas and glosses for the most important words and because historical spelling variation is not confounded with digitization errors.

3 A Corpus of Contemporary Legislative Texts

3.1 The Classified Compilation of Swiss Federal Legislation

The Classified Compilation of the Federal Legislation (abbreviated SR) is a systematic collection of the contemporary statutory law of the Swiss Confederation. It comprises federal acts, ordinances issued by the federal authorities, the federal constitution, all

[3] We are working with these projects to create linkages between the various resources; for example, the DRW already links evidence from the Collection to the digital facsimiles.

cantonal constitutions, federal decrees, and treaties between the confederation and individual cantons or municipalities.

Each text is published in the three official languages of the Swiss Confederation: German, French, Italian. While the German and the French version of a legislative text are usually drafted and edited in parallel, the Italian version is, in most cases, merely a translation. However, all three official language versions are considered authentic, i.e., they all have equal legal force (see Lötscher, 2009).

As opposed to historical legal texts, contemporary laws are relatively easy to obtain—which greatly facilitates the process of corpus building. Nowadays, most legislative texts are published online, and the texts are usually not subject to copyright provisions that would prevent their use and distribution for research purposes. The Classified Compilation of Swiss Federal Legislation is no different in this regard: The collection can be accessed online at the website of the Swiss federal authorities, where all texts are available in HTML and in PDF format.[4]

3.2 The Corpus: The SLC

We have exploited the Classified Compilation of Swiss Federal Legislation to build an annotated corpus of contemporary legislative texts: the Swiss Legislation Corpus (SLC). This corpus has the following characteristics.

First, the SCL is *domain-complete*. It contains all texts published in the Classified Collection and thus comprises the whole body of contemporary legislative writing of the Swiss Confederation. In total, the SLC consists of 1915 texts per language. The sizes of the individual texts range from roughly 800 words (Federal Decree on the Coat of Arms, SR 111) to over 1.3 million words (Code of Obligations, SR 220).[5]

Second, the SLC is a *parallel corpus*. All texts are available in German, in French and in Italian. The conventions of Swiss legislative drafting ensure that even in their raw form, the texts exhibit a precise alignment of all language versions down to the level of individual sentences and enumeration items. Legal technicalities make it mandatory that each sentence and enumeration item of a legislative text can be identified unequivocally by naming the respective law and the number of the article, the paragraph and, where applicable, the sentence or enumeration item (e.g., *Art. 6 Par. 2 Ltr. b Federal Act on Professional Education, SR 412.10*). This identifier is language independent and thus ensures alignment between the text versions (see Figure 2). Occasionally, translation issues make it necessary that a statement that is expressed in a single sentence in one language has to be rendered as two sentences in another language. In these cases, the two sentences in the latter version are separated by a semicolon rather than a full stop. Thus, it can be guaranteed that the respective passage can still be referred to by one and the same sentence identifier in all language versions of the text.

Third, the SCL exhibits both *inter- and intra-textual time depth*. Despite the fact that all material found in the SLC constitutes contemporary Swiss federal law, there is

[4] `http://www.admin.ch/ch/d/sr/`

[5] The sizes refer to the German versions of the texts.

Art. 6 Verständigung und Austausch zwischen den Sprachgemeinschaften

[1] Der Bund kann Massnahmen im Bereich der Berufsbildung fördern, welche die Verständigung und den Austausch zwischen den Sprachgemeinschaften verbessern.

[2] Er kann insbesondere fördern:

a. die individuelle Mehrsprachigkeit, namentlich durch entsprechende Anforderungen an die Unterrichtssprachen und die sprachliche Bildung der Lehrkräfte;

b. den durch die Kantone, die Organisationen der Arbeitswelt oder die Unternehmen unterstützten Austausch von Lehrenden und Lernenden zwischen den Sprachregionen.

Art. 6 Compréhension et échanges entre les communautés linguistiques

[1] Dans le secteur de la formation professionnelle, la Confédération peut encourager les mesures qui favorisent la compréhension et les échanges entre les communautés linguistiques.

[2] Elle peut notamment encourager:

a. le plurilinguisme individuel, en veillant en particulier à la diversité des langues d'enseignement ainsi qu'à la formation des enseignants sur le plan linguistique;

b. les échanges d'enseignants et de personnes en formation entre les régions linguistiques, s'ils sont soutenus par les cantons, les organisations du monde du travail ou les entreprises.

Figure 2: Example of the alignment between the language versions of contemporary Swiss laws (excerpts from the German and the French versions of the Federal Act on Professional Education, SR 412.10)

a considerable diachronicity both between and within individual texts. As a whole, the corpus exhibits a time depth of 136 years: Its oldest text dates from June 22, 1875, its most recent text from March 30, 2011. Likewise, a time span of up to 122 years can be found within individual texts. Laws are subject to continuous alterations: articles, paragraphs, sentences or enumeration items may be added, changed or removed by the legislator. The Federal Act on Debt Enforcement and Bankruptcy (SR 281.1), for instance, originates from April 11, 1889, but its most recent update—in which an article and an enumeration item were added—only dates from September 1, 2011.

Fourth, the SCL is an *annotated corpus*. The texts have been converted into XML. At present, they are enriched with tags providing meta information (dates of issue and last update, title and number of the text, issuing authority, legal basis of the law), delineate structural units (chapter, section, article, paragraph, sentence and enumeration item boundaries) and indicate parts of speech. The annotation of syntactic structures is in preparation. To facilitate querying, the SCL has been imported into the ILM Corpus Workbench (Christ, 1994).

3.3 Automatic Annotation

The annotation of the SLC is largely determined by the characteristics of the domain of legislative texts. One task that plays a much more central role in the processing of laws than it does in other domains is text segmentation. As we have illustrated in the previous section, laws are heavily structured: They are partitioned into numbered chapters, sections, articles, paragraphs, sentences and enumeration items. Marking the boundaries of these structural units is crucial if one wants to preserve of the alignment between the individual language versions of the texts. Furthermore, the availability of such an annotation is a prerequisite for corpus-based studies into the discourse structure of legislative texts.

We have developed a tool that automatically marks the boundaries of textual units. The method that we employ combines line-based pattern matching with a look-behind strategy. For example, a line is recognized as an enumeration item if (a) it begins with a lowercase character (optionally accompanied by a Latin ordinal such as *bis, ter, quater*, etc.), followed by a full stop and one or more words, and if (b) the previous line has already been tagged either as an enumeration item or as the introductory sentence of an enumeration. The second line of the following excerpt (Article 26 of the Federal Act on Forest, SR 921.0), for instance, will thus be annotated as an enumeration item:

```
1  <paragraph issue_date="04/10/1991"><par_nr>1</par_nr>
   <enum_intro_sentence>Der Bundesrat erlässt Vorschriften über forstliche
   Massnahmen:</enum_intro_sentence>
2  a.  zur Verhütung und Behebung von Waldschäden;
```

Another feature of legislative texts is that each structural unit can be associated with a number of dates: the date of its first publication as part of a decree, the date of its official approval by the parliament or the people, and finally the date of its commencement. If the dates for a specific textual unit differ from those of the text as a whole, they are listed in a footnote attached to that unit (see Figure 3). Text segmentation must therefore also include date stamping.

We use pattern-matching methods to extract the dates from the footnotes and make them explicit in the markup of the corresponding text unit. In the markup, all text segmentation tags are augmented with attributes denoting the dates of the respective unit (e.g. `<paragraph issue_date="04/10/1991">`). By default, they are assigned the dates associated with the whole text. If, however, the respective unit is accompanied by a footnote mentioning a specific date (as it is the case with paragraphs 3 and 4 in Figure 3), that date is extracted from the footnote (e.g. by matching the string `'Angenommen in der Volksabstimmtung vom DATE'`) and inserted in the paragraph tag (thus replacing the default date).

The provision of precise date stamping for each textual unit allows for diachronic analyses of the linguistic material found in a text: We can, for instance, study if the language of earlier passages deviates from the language of passages that were added

Art. 175 Zusammensetzung und Wahl

[1] Der Bundesrat besteht aus sieben Mitgliedern.

[2] Die Mitglieder des Bundesrates werden von der Bundesversammlung nach jeder Gesamterneuerung des Nationalrates gewählt.

[3] Sie werden aus allen Schweizerbürgerinnen und Schweizerbürgern, welche als Mitglieder des Nationalrates wählbar sind, auf die Dauer von vier Jahren gewählt.[89]

[4] Dabei ist darauf Rücksicht zu nehmen, dass die Landesgegenden und Sprachregionen angemessen vertreten sind.[90]

[89] Angenommen in der Volksabstimmung vom 7. Febr. 1999 (BB vom 9. Okt. 1998, BRB vom 2. März 1999 – AS **1999** 1239; BBl **1993** IV 554, **1994** III 1370, **1998** 4800, **1999** 2475 8768).

[90] Angenommen in der Volksabstimmung vom 7. Febr. 1999 (BB vom 9. Okt. 1998, BRB vom 2. März 1999 – AS **1999** 1239; BBl **1993** IV 554, **1994** III 1370, **1998** 4800, **1999** 2475 8768).

Figure 3: Example of the date stamping of textual units in contemporary legislative texts (excerpt from the Federal Constitution, SR 101)

more recently, or we can investigate how the continuous insertion of additional material has affected the overall structure of a text.

In addition to text segmentation (and the annotation of textual meta information), we have annotated the words in all three language versions of the SLC with their part of speech and their lemma. We used TreeTager (Schmid, 1994) for this task. Domain-specific words unknown to TreeTagger constituted the main problem. However, except for archaisms like *bejahendenfalls* 'in case of affirmation,' most unknown words turned out to be nouns (including proper names and abbreviations) or adjectives. In most cases, TreeTager was able to guess the part of speech of these words correctly; only their lemmas could not be inferred.[6] We are confident that this situation can be remedied by equipping TreeTagger with a hand-made list of domain-specific expressions and their lemmas.

3.4 Applications

The need for annotated corpora of legislative texts has grown with the recent advance of legal linguistics as a theoretical and applied academic discipline (see Grewendorf and Rathert, 2009). The SLC is meant to make a contribution to filling this gap.

We currently use the SLC to study the stylistic properties of legislative texts. We are interested in investigating to what extent present-day Swiss laws comply with established stylistic guidelines for legislative drafting. To this aim, we use the method of error modeling employed in automated language checkers for technical writing: We

[6] The evaluation refers to the German part of the corpus. We have manually assessed the tagging of 1,000 randomly selected tokens. 85 (8.5%) of these tokens were unknown to TreeTagger; in total, 399,872 (6.7%) of the 5,896,451 tokens contained in the corpus were unknown to TreeTagger. For 67 (79%) of the manually evaluated unknown tokens, TreeTagger was able to guess the part of speech correctly; only 18 (21%) were assigned a wrong part of speech.

	DS21	SLC
Source	Collection of Swiss Law Sources	Classified Collection of Swiss Federal Legislation
Temporal classification	Historical	Contemporary
Text types	Statutes, decrees, regulations, indentures, treaties, administrative documents, court transcripts, letters, and others	Federal acts, ordinances, federal and state constitutions, federal decrees, non-international treaties
Languages	German, French, Italian (historical regional variants)	German, French, Italian (parallel)
Units of alignment	N/A	Sentences, enumeration items
Number of texts	10,691 (total)	1915 (per language)
Time depth	1078 years (754–1832)	136 years (1875–2011)
Intra-textual time depth	Unknown	Up to 122 years
Annotated information	Meta information, structural units, persons and place names	Meta information, structural units, parts of speech; syntax in preparation
Format	XML: TEI P5	XML, IMS Corpus Workbench (CWB)
Current applications	Historical research	Stylistic analysis, definition extraction

Table 2: Key properties of the two corpora

first specify linguistic features and textual patterns that indicate the violation of a specific style guideline and then search the SLC for occurrences of these indicators.[7]

The oft-cited rule that, in a good legislative text, an article should not contain more than three paragraphs, a paragraph should only contain one sentence, and a sentence should not make more than one statement may serve as an example (Federal Office of Justice, 2007, p. 358). An evaluation of the first two parts of the rule can be done by accessing the annotated structural units: The rule is violated in articles with more than three paragraphs and in paragraphs with more than one sentence. Violations of the third part of the rule can be found by searching for specific keywords and syntactic structures. Höfler (2011), for instance, points out that, among other things, sentence coordination, relative clauses introduced by the adverb *wobei* 'whereby,' and certain prepositions (e.g., *vorbehältlich* 'subject to' or *mit Ausnahme von* 'with the exception of') indicate that a sentence makes more than one statement.

In a related strand of research, we use the SLC to extract legal definitions (Höfler et al., 2011). We exploit the fact that, by convention, legal definitions follow a relatively small inventory of sentence patterns. Searching the SLC for these patterns allows us to identify these definitions. An automatic extraction of the concepts and terms defined in the present legislation can be of value to legal practitioners, to scholars of law, and to professionals involved in the drafting and editing of new acts and ordinances.

[7] We also work on employing the same method to check draft laws for style guideline violations.

4 Summary

In this paper, we have described the construction of two corpora of Swiss legal texts: the DS21 corpus, a corpus of historical legal texts, and the SLC, a collection of contemporary laws. The two corpora are complementary: Together, they reflect the historical development of Swiss legal language almost in its entirety (except for the period from 1798 until the formation of the federal state in 1848).

We have illustrated that the availability of such corpora facilitates a plethora of humanities research, particularly in the fields of history, linguistics, and law. We have also shown that the peculiarities of the legal texts represented in the two corpora had a strong impact on the tasks that had to be solved in order to build them. The work presented in this paper emphasizes that the construction of domain-specific corpora also involves putting work and effort into developing domain-specific annotation tools.

References

Boschetti, F. (2007). Methods to extend Greek and Latin corpora with variants and conjectures: Mapping critical apparatuses onto reference text. In Davies, M., Rayson, P., Hunston, S., and Danielsson, P., editors, *Proceedings of the Corpus Linguistics Conference CL2007*. University of Birmingham.

Christ, O. (1994). A modular and flexible architecture for an integrated corpus query system. In *Proceedings of COMPLEX 1994, 3rd Conference on Computational Lexicography and Text Research*, pages 23–32.

Federal Office of Justice, editor (2007). *Gesetzgebungsleitfaden: Leitfaden für die Ausarbeitung von Erlassen des Bundes*. Berne, Switzerland.

Grewendorf, G. and Rathert, M., editors (2009). *Formal Linguistics and Law*, volume 12 of *Trends in Linguistics*. Mouton de Gruyter, Berlin, Germany.

Gschwend, L. (2008). Rechtshistorische Grundlagenforschung: Die Sammlung Schweizerischer Rechtsquellen. *Schweizerische Zeitschrift für Geschichte*, 58(1):4–19.

Höfler, S. (2011). "Ein Satz – eine Aussage." Multipropositionale Rechtssätze an der Sprache erkennen. *LeGes: Gesetzgebung und Evaluation*, 22(2):275–295.

Höfler, S., Bünzli, A., and Sugisaki, K. (2011). Detecting legal definitions for automated style checking in draft laws. Technical report, Department of Informatics, University of Zurich.

Lötscher, A. (2009). Multilingual law drafting in Switzerland. In Grewendorf, G. and Rathert, M., editors, *Formal Linguistics and Law*, volume 12 of *Trends in Linguistics*, pages 371–400. Mouton de Gruyter, Berlin, Germany.

Piotrowski, M. (2010a). Document conversion for cultural heritage texts: FrameMaker to HTML revisited. In Antonacopoulos, A., Gormish, M., and Ingold, R., editors, *DocEng 2010: Proceedings of the 10th ACM Symposium on Document Engineering*, pages 223–226, New York, NY, USA. ACM.

Piotrowski, M. (2010b). Leveraging back-of-the-book indices to enable spatial browsing of a historical document collection. In Purves, R., Clough, P., and Jones, C., editors, *Proceedings of the 6th Workshop on Geographic Information Retrieval (GIR'10)*, pages A17/1–2, New York, NY, USA. ACM.

Rechtsquellenstiftung des Schweizerischen Juristenverbandes, editor (2007). *Rechtsquellen der Stadt und Herrschaft Rapperswil*, volume SSRQ SG II/2/1 of *Sammlung Schweizerischer Rechtsquellen*. Schwabe, Basel, Switzerland. Prepared by Pascale Sutter.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49.

Dain Kaplan, Ryu Iida, Kikuko Nishina, Takenobu Tokunaga

Slate – A Tool for Creating and Maintaining Annotated Corpora

Recent research trends of the last five years show that richly annotated corpora inspire novel research. These richly annotated corpora are indispensable for progressing research, but also more difficult to manage and maintain due to increasing complexity – what is needed is a way to manage the annotation project in its entirety. However, annotation project management has received little attention, with tools predominately focusing on single document annotation. Therefore, we define a list of corpus creation and management needs for annotation systems, and then introduce our multi-purpose annotation and management system `Slate` to address these needs through use of a case study, showing how project management is essential to creating good corpora.

1 Introduction

A look at recent research of the last five years related to language resources reveals an increasingly growing wealth of literature utilizing corpora in many languages on an ever wider variety of topics, including spoken corpora, fluency, gender and age differences, semantic analysis, parallel corpora, work targeting poor-resource language development, and so on.These works have been made possible by the existence and development of corpora.We can see from research trends that richly annotated corpora inspire novel research. Yet, they are also more difficult to manage and maintain due to their increasing complexity. Our goal should then be to encourage the creation and extension of such richly annotated corpora by making it as easy as possible, regardless of discipline.It is important for the annotation tool to allow corpus builders to experiment with a variety of ways for expressing their tasks, i.e., to be flexible in how the annotation schema is expressed, since it dictates the style of work.

We aim to facilitate this kind of corpus creation by developing an annotation tool called `Slate`[1] Kaplan et al. (2010)which is not only capable of adapting to many kinds of annotation tasks, but differs from many other existing tools in that it also covers management of the annotation process, which is increasingly important as the sizes of corpora continue to grow Davies (2009). In the following sections we outline the needs for corpus creation and management (Section 2), briefly introduce our annotation and management tool `Slate` (Section 3), and then demonstrate its utility with a real world case study in Section 4. We briefly follow this up with more examples of use (Section 5), explain other annotation tools and types (Section 6), and then conclude (Section 7).

[1] Segment and Link-based Annotation Tool, *Enhanced*: `http://www.cl.cs.titech.ac.jp/slate`

2 Corpus Creation and Management Needs

As those that utilize language resources continue to diversify and grow in number, so naturally will the uses of the corpora. A more richly annotated corpus provides more value than a collection of raw text; not to underplay the importance of such corpora, but a corpus with various annotations opens new windows for new research and new ways of analyzing this new data. Therefore the barrier to entry for creating new corpora, or extending existing corpora should be as small as possible. Further, as these resources grow and expand, their evolving complexity also necessitates a means for *managing this complexity*, or they themselves will become unmanageable and the resources quickly unusable.

Thus we propose a set of needs that will be indispensable for this task of maintaining and managing, and also creating new annotated corpora. Traditional single document-style annotation tools may be increasingly difficult to use if you wish to coordinate annotation by multiple annotators, insure the proper versions are used and verify/check on progress. An annotation system that fulfills these needs will allow management of creation of the annotation resource as a whole, rather than piecemeal on a per-document-only basis. With simple corpora, managing the project may not be a concern, but in larger projects – and as corpus-based techniques continue to grow and advance so do the sizes of the corpora Davies (2009) – it becomes a serious issue. In order to truly facilitate corpus creation we must therefore also address the annotation project as a whole.

Let us then specify what is needed from an annotation system. Dipper et al. (2004) proposed seven categories for what is needed from an annotation tool (diversity of data, multi-level annotation, diversity of annotation, simplicity, customizability, quality assurance and convertibility).They can, however, be thought of as targeting the *document-level*. The following list serves as a compliment to this list, operating at a more macro, annotation *project-level*. As our list is complimentary, we do not wish to reiterate what was already well said, and therefore skip needs related specifically to the act of annotation (such as appearance).

(1) **User and role management**. For some annotation projects, a simple, single document-oriented annotation tool may be sufficient (such as if the data is small and there is only one annotator), but with larger, more complex corpora, it may be advantageous to have a system to manage the overall project, including roles for an *annotator* (who performs the annotation work), and for an *administrator* (who configures the annotation schema and oversees the project).

(2) **Delegation and monitoring of work**. The system should allow for an administrator to assign/reassign work to annotators, and to monitor their progress. In addition, each annotator should be isolated – or "sandboxed" – from others, so that they are not biased by others' work.

(3) **Adaptability to new annotation tasks**. The system must be flexible enough to accommodate a new annotation task. If administrators cannot easily create a

new project and define the annotation requirements (i.e. the annotation schema) then the system will not be useful to them.

(4) **Adaptability within the current annotation task**. During the lifespan of a project, it often meets with many changes, especially during its initial phases. It is crucial that the system allow for the adjustment of the annotation schema.

(5) **Diffing and merging**. Creating a corpus often entails the comparison of data from multiple annotators on a single resource and reconciling any differences (i.e. *diffing and merging*), to finalize the corpus (create a gold standard). However, if multiple annotators' data is desired, this should also be possible.

(6) **Versioning of corpora**. A corpus is a product, and like any other product, may go through lifecycles.There must be a way to package the corpus, and in the case of fixes or amendments, a way to add them and repackage it. Without management of versioning, the "current state" of the annotation project is all that is known; for large projects it is important to identify milestones or to mark[2] a given state.This way changes will not unknowingly be included into a release.

(7) **Extensibility in terms of layering**. As corpora are continuing to grow in size, it is no wonder that they are also becoming more complex. We have seen recently the layering of corpora upon one another. Examples of this include the Penn Discourse Treebank Miltsakaki et al. (2004) or VP Ellipsis Bos and Spenader (2011) on top of the Penn Treebank Marcus et al. (1993), or the NAIST Text Corpus atop the Kyoto Text Corpus Iida et al. (2007). The system should allow for adding new layers upon previous ones

(8) **Extensibility in terms of tools**. With larger, more complex corpora, it may be difficult for annotators to do all annotation by hand. Having a mechanism for calling plugins to process data allows for various kinds of semi-automatic tagging (e.g. *annotation projection*) while data is within the system. In addition, plugins could allow for changing the way tags are selected from tag-sets, or automatically tagging areas found by the plugin to be similar to areas already manually annotated.

(9) **Extensibility in terms of importing/exporting**. The system should be able to call plugins to convert the data during import and export. This allows the system to be agnostic to the data format and support a variety of standardized formats.Including generated comments into the format may facilitate in human verification of the exported resource.

(10) **Support for multiple languages**. Research today is carried out in a variety of languages, and the system should support them.

[2]Known in source-code versioning control as "tagging".

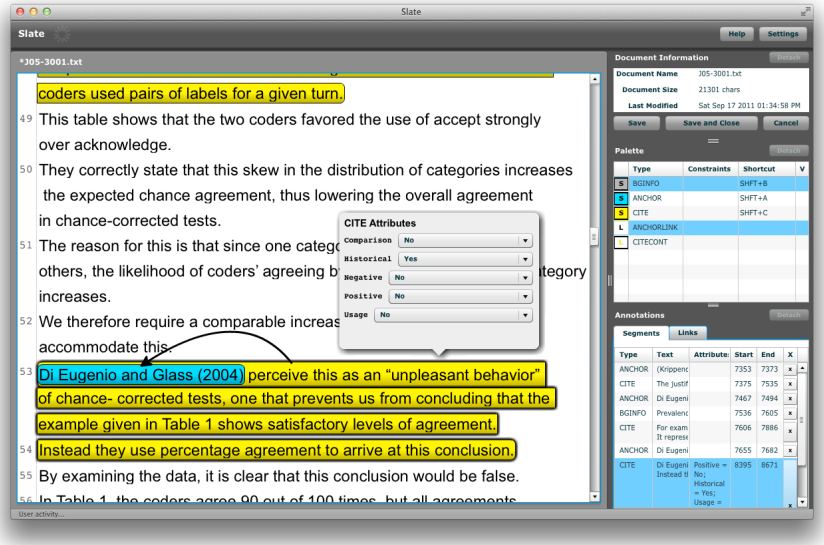


Figure 1: An example of the English Citation Corpus (see Section 5), showing an annotation of a citation's text pointing to an annotation of its citation anchor, and attribute panel with attributes related to the CITE tag.

3 `Slate` – A Web-based Annotation and Project Management Tool

We have developed `Slate` to try to meet the needs defined in Section 2.[3] One goal of the system is to be general enough to allow for a variety of textual annotation tasks. A conceptual framework is needed for representing annotations internally. ATLAS (Flexible and Extensible Architecture for Linguistic Annotation) Bird et al. (2000) is one possibility, very generalized and thus suitable for many tasks. Since our goal is textual annotation, we opted for using Segments and Links Takahashi and Inui (2006); Noguchi et al. (2008), a framework optimized specifically for this case. In this framework, annotations are stored in a stand-off[4] format.[5] Two concepts are used in annotation: *Segments* and *Links*. A *Segment* is defined as a span of text with a typed[6] label. Any textual annotation will be defined as a kind of Segment; using a stand-off

[3] At the time of writing, (5), (7), and (8) have not yet been fully realized.

[4] Meaning that the data is *not* directly embedded in the original text such as in XML, but separate to it; this supports partially overlapping annotations not possible in XML-like formats.

[5] `Slate` in fact uses a relational database for storing data while it is within the system.

[6] "Typed" as in "types" of Segments are defined, and those definitions are used to place those types as labels on spans of text.

format allows Segments to partially overlap. Similarly, a *Link* is a typed-relationship between two Segments. Any Segment or Link definition may further define attributes. For instance, a "Named Entity" Segment definition could have an attribute "Category" with possible values "Person", "Place", or "Thing".

Tag definitions are configured through a graphical user interface (GUI) within a Web-browser by the project administrator.Since `Slate` is both an annotation tool, and a tool for annotation project management, it allows project administrators to split up groups of documents to be annotated into smaller subsets, and to assign those subsets to annotators. `Slate` is Web-based, so annotators need not install any specialized software, only needing access to the internet. All of the annotators' changes are saved to the server, so their work can be monitored by the project administrator.

The current implementation is geared towards textual markup; but we have designed the system to be flexible, and it would be possible to replace the annotation canvas suited for textual markup with another suited for a different task, such as editing phrase-structure. (This is not impossible with the current annotation canvas, but definitely impractical.) `Slate` also supports an I/O plugin framework, so supporting an existing standard is only a matter of writing the interfacing code to read/write the format for importing/exporting.For more technical details, please refer to Kaplan et al. (2010).

The annotation screen is shown in Figures 1 and 2. For simplicity, we will use Figure 1 for explaining the annotator UI (as the example is in English). The left-side of the screen shows the main annotation panel; here is where the text to be annotated is shown. The annotator creates new annotations by selecting the type of Segment (tag) they wish to create from the middle panel on the right (where it says "Palette"), or by pressing that Segment's keyboard shortcut to select it, and then clicking and dragging the mouse over the text-span they wish to annotate, much like selecting text with a mouse in a text processor. To create a Link between two Segments, the annotator clicks and drags the mouse from the source to the destination Segment. The end result is an arrow (or undirected line if the Link definition is specified as such) between them.

The Palette in Figure 1 shows five definitions, three Segment types, and two Link types (the single letter in the leftmost column of the Palette panel shows the definition as a "S" for Segment, or "L" for Link). The Segment and Link definitions that appear here are determined by the schema that the administrator has defined for the project, as mentioned above. The main annotation panel shows a CITE Segment (yellow) with a Link (black arrow) to an ANCHOR Segment (teal). In the case that Segments overlap, Figure 1 shows that they become nested within one another.

If a Segment definition has attributes, the annotator may click on a selected Segment to open the attributes panel, where he/she may edit any values as specified by the Segment definition (compare Figure 1 with Figure 2). Links may also have attributes, and are editable in the same manner as for Segments.

The bottom right of the screen shows all created Segments and Links, letting the annotator easily navigate through the document. The top right of the screen has

buttons for saving or canceling changes, and basic information about the document, like when it was last saved.

4 A Case Study: Non-native Japanese Learner Composition Errors Corpus

The best way to see how `Slate` can facilitate creation of language resources without the need for developing a custom tool is to look at real-world examples. In this section we present a case study for an annotation task from Japanese Education in the Humanities, showing the problems encountered with their previous method of annotation, and how `Slate` helps to resolve many of these issues, increase the precision of the resulting data, and diversify the corpus as a whole. See Section 5 for a quick look at other examples.

4.1 Japanese Learner Composition Errors

During the process of learning a foreign language, one often makes mistakes along the way. A corpus of such common mistakes is indispensable to language instructors for snipping the bud of bad habits before they form; understanding learner tendencies for errors is crucial in designing lesson plans, general instruction, creating special learner dictionaries, devising tests, etc. As a compliment to modern day compositional support systems, such as `Natsume`[7], it is also possible to create compositional tools that automatically warn learners of potential mistakes as they are making them.

The underlying corpus texts are essays written by Japanese language learners on a series of topics. Currently there are over 261 essays (with 3,500 sentences) written by more than 164 learners.

4.2 Before `Slate`

At the time the corpus project began, there was no Japanese learner composition errors corpus in digital form, and therefore no specialized tool for its creation. The corpus creation started by using Microsoft Excel, which at the time, was an easily accessible tool at hand. However, using Excel proved to be a trying experience; some of the more major limitations encountered are outlined below.

Limitations on Data Integrity

I-1 Excel does not enforce formatting or style within cells; as a result an annotator's method for describing errors would very day-by-day based on his/her current disposition (e.g. using an arrow instead of parentheses, etc.).

I-2 Copy and paste made it difficult to keep consistent format, including the possibility for inserting accidental changes into the data.

I-3 The overall consistency of the data was poor due to I-1 and I-2, which made systematic analysis of the data difficult.

[7] `http://hinoki.ryu.titech.ac.jp/natsume/`

I-4 As a result of I-3, measuring inter-annotator agreement was also difficult.
I-5 The flow of text was difficult to grasp by looking at an Excel cell, often meaning the entire sentence could not be seen at once, increasingly likelihood for errors in annotation.
I-6 Many columns made the possibility for entering data in the wrong cell high.

Limitations on data diversity

D-1 Non-contiguous learner errors (such as a grammatical construction begun at the start of a sentence, but not finished until the end) where difficult to describe.
D-2 Annotating a sentence containing multiple leaner errors was not possible.
D-3 Relations between sentences, where a mistake is started or continued, were not possible.

4.3 With `Slate`

A screenshot of the annotation screen using `Slate` for this project is shown in Figure 2. This project has only one Segment definition, shown in purple, which marks composition errors. The attributes panel has a number of values related to the type of error, its correction, etc.

By converting the existing data, the project was able to migrate existing annotations into `Slate`. Once data is imported, the underlying text that is annotated is not changed, as a result this remedies miscopied data issues (I-2). The underlying text is never changed because all annotations are stored separately from the text; this allows multiple annotators to annotate the same document without being interfered or biased by one another. This makes inter-annotator agreement I-4 possible in the sense that now multiple annotators can work on the same document.

As explained briefly in Section 3, `Slate` does not predefine annotation tag-sets, but instead lets an administrator define them. Once the tag-set(s) are defined, all annotation is constrained to them. This means the user's input is limited, preventing issues where no system was constraining them from freely formatting text (I-1).

Because `Slate` is a visual tool, it eliminates many of the limitations prevalent in the previous annotation method. Non-contiguous errors (D-1) and errors that span multiple sentences (D-3) are possible to annotate with the ability to use *Links* to connect the *Segments*. Multiple annotations within the same sentence (D-2) is also now possible (see Figure 1). As a visual tool, it also means that the annotator is able to see the flow of text (I-5). Since users are directly interacting with what they are annotating, there is higher accuracy of annotation (I-6).

Further, now as the team of annotators grows to include members spread across the globe, there is no issue in sending and receiving work, as it is all managed online by `Slate`.

Most importantly, because `Slate` outputs a specified format with all the annotations, analysis (I-3) is now more straightforward.

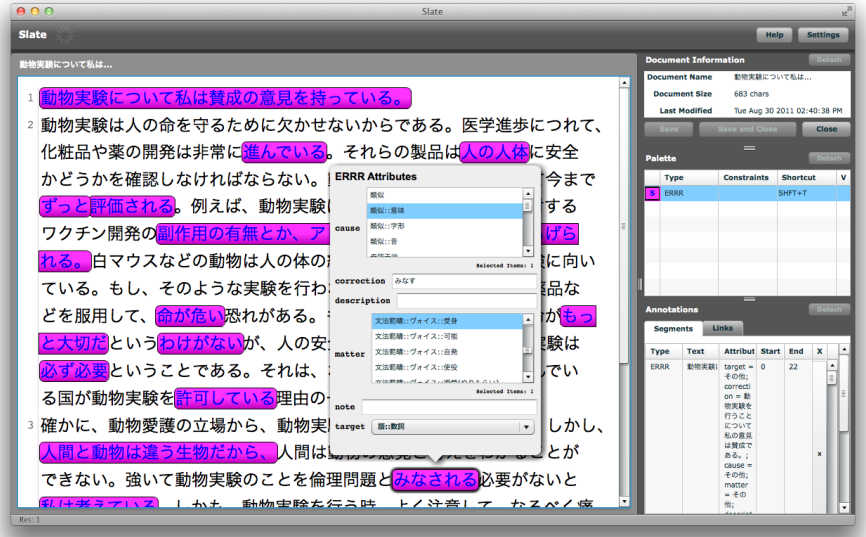


Figure 2: An example of composition errors annotated using Slate, with the panel showing editing attributes for a given Segment.

5 Other Corpus Creation Projects using `Slate`

Though space prevents a more detailed look, a few other projects utilizing `Slate` are briefly described below.

- **Japanese Blog Anaphora Corpus** – A corpus composed of blog texts, with annotations for the expressions that hint at the anaphoric relations (such as "this" or "that"), and the actual known referent of each, with explicit references between them. There are attributes on the annotations to classify them into various categories, such as between explicit and implicit anaphors. The anaphors and their referents are realized using Segments, and coreference relations using Links. `Slate` allows for free definition of attributes on annotations, so various kinds of simultaneous classifications are possible simply by defining multiple attributes for each annotation type.

- **Japanese Q&A Website Corpus** – A corpus composed of question threads[8] posted on Q&A Web sites, containing annotations marking the parts in the text

[8] I.e. the hierarchical flow of posts made by users asking a question and receiving answers, with possible additional posts by the initial asker with clarification about the problem along the way.

that indicate background about the problem setting, parts of the text that describe the problem itself, and parts that are directly answered by other users (the answers that directly correspond to parts of the question(s) asked). Two potential uses for such a corpus include improving the precision of Q&A site searching by knowing more semantically about the questions and their answers, and which parts of the text reflect each (e.g. limiting your search to within problem setting descriptions only). `Slate` allows for annotating these various parts of text, and for specifying relations between them quite freely.

- **English Citation Corpus** – A corpus of research papers with the citations in each annotated, including directly related background information that pertains to the citation, and various classifications for each citation, such as marking it as a comparison/contrast, refuting a claim, etc. In the case that the citations are non-contiguous Kaplan et al. (2009), they are connected to each other using *Links*. `Slate` allows for the various classifications using attributes on the annotations, and lets you link annotations together to make creation of this corpus possible. A sample of the annotation screen in `Slate` is shown in Figure 1.

6 Comparison with similar / related tools

In this paper we have shown how `Slate` can be a viable solution to many annotation tasks, especially for enabling those otherwise unable, to create richer, more precise language resources. It has been designed with corpus management functionality, an area overlooked by most other tools of this kind. There are, however, many other tools available that satisfy a wide range of different tasks to varying degrees. Major differences are outlined below.

In essence three kinds of annotation tools exist, ranging from specialized to general: project-specific tools, task-oriented tools, and generalized or multi-purpose tools. In addition to these three kinds of tools, there are tools that run as desktop applications, and those that are Web-based; further, there are tools that are single document-centric, or that are concerned only with the current document being opened by the user, and those that attempt to help the user manage the annotation task as a whole (by grouping files, providing comprehensive search, etc.). In general, non-Web-based tools have several drawbacks related to data management and data consistency, such as maintaining consistent versions of the files on multiple computers, transferring the files to the annotators' computers (including potential licensing issues involved as well as file sizes), redistributing the files to different annotators in the case of reassignment, and remembering which files have been distributed to which annotators.

Project-specific tools are very specialized. Usually they support a single corpus format, so they can only work with a single project. Such tools are good in one sense because they meet the specific needs of an annotation project, but require resources for their development/maintenance, and may also not be compatible with other corpora, and thus not support layering (preventing the development of richer,

more complex resources). Some more recent examples of project-specific tools include a Korean treebank Park et al. (2006), lexical chains Stührenberg et al. (2007), image annotation Russell et al. (2005), and sentiment tagging Francisco et al. (2011). Of these, some are Web-based Stührenberg et al. (2007); Russell et al. (2005); Francisco et al. (2011). **Task-specific tools** are more generalized than project-specific tools, but still often exist because of a certain project and its descendants, such as a treebank annotator reading and writing files in Penn Treebank format. Some recent examples include: text alignment to speech Draxler (2005), word-alignment Madnani and Hwa (2004), syntactic annotation Noguchi et al. (2006), and frame-set annotation Choi et al. (2010b,a). Some are desktop applications Madnani and Hwa (2004); Choi et al. (2010b,a), and some Web-based Draxler (2005); Noguchi et al. (2006).

Multi-purpose tools are capable of adapting to a variety of tasks. The most famous of these may be MMAX2 Müller and Strube (2006), and also the NITE XML Toolkit Carletta et al. (2005). Glozz Widlöcher and Mathet (2009) and Word Freak Morton and LaCivita (2003) have plugin architectures and nice interfaces, comparable in many ways to `Slate`, but lack project management and as easy to use schema definition UI. For multimedia annotation ELAN is a popular choice Auer et al. (2010). There are other tools in various states of activity/dormancy Asan and Orăsan (2003); Cunningham et al. (2002); Dennis et al. (2003); Mueller and Strube (2001); Callisto (2002).

The tool most similar to `Slate` is probably Djangology Apostolova et al. (2010), which is a Web-based tool supporting project management. Its UI however, does not allow for visual creation of links/relations between segments like `Slate` does. The other multi-purpose tools are not Web-based, nor do they support management of the annotation project.

7 Conclusion

There is of course no silver bullet to annotating corpora; their creation comes at the cost of substantial time and labor. However, because of this there is all the more reason to make sure that the time and effort spent are only spent where they are necessary, and are kept to a minimum everywhere else. To this end, we have tried to make sure `Slate`'s ability to manage corpus creation and to create a flexible annotation schema allows for different styles of working. We have also designed the system with extensibility in mind, so that future developers may more easily adapt it to new tasks and standards.As the various project examples have shown, by allowing freedom in how the schema is defined `Slate` can adapt to a number of textual annotation tasks. Further, it allows annotators to get started quickly since no setup is required, and provides project administrators instant access to the annotators' work.

`Slate` is still under active development as we work towards realizing all the needs outlined in Section 2. The project website is: `http://www.cl.cs.titech.ac.jp/slate/`.

Acknowledgements

This work has been supported in part by the Grant-in-Aid for Scientific Research Priority Area Program "Japanese Corpus" (2006 – 2010), sponsored by MEXT (Ministry of Education, Culture, Sports, Science and Technology – Japan).

References

Apostolova, E., Neilan, S., An, G., Tomuro, N., and Lytinen, S. (2010). Djangology: A light-weight Web-based tool for distributed collaborative text annotation. In *Proc. of the 7th conference on International Language Resources and Evaluation (LREC'10)*, pages 3499–3505.

Asan, C. O. and Orăsan, C. (2003). Palinka: A highly customisable tool for discourse annotation. In *Proc. of the 4 th SIGdial Workshop on Discourse and Dialogue, ACL '03*, pages 39–43.

Auer, E., Russel, A., Sloetjes, H., Wittenburg, P., Schreer, O., Masnieri, S., Schneider, D., and Tschöpel, S. (2010). ELAN as flexible annotation framework for sound and image processing detectors. In *Proc. of the 7th conference on International Language Resources and Evaluation (LREC'10)*, pages 890–893.

Bird, S., Day, D., Garofolo, J. S., Henderson, J., Laprun, C., and Liberman, M. (2000). ATLAS: A flexible and extensible architecture for linguistic annotation. *CoRR*, cs.CL/0007022.

Bos, J. and Spenader, J. (2011). An annotated corpus for the analysis of VP ellipsis. *Language Resources and Evaluation*, 45(2):1–32.

Callisto (2002). `http://callisto.mitre.org`.

Carletta, J., Evert, S., Heid, U., and Kilgour, J. (2005). The NITE XML Toolkit: Data model and query language. *Language Resources and Evaluation*, 39(4):313–334.

Choi, J. D., Bonial, C., and Palmer, M. (2010a). PropBank FrameSet annotation guidelines using a dedicated editor, Cornerstone. In *Proc. of the 7th conference on International Language Resources and Evaluation (LREC'10)*, pages 3650–3653.

Choi, J. D., Bonial, C., and Palmer, M. (2010b). PropBank instance annotation guidelines using a dedicated editor, Jubilee. In *Proc. of the 7th conference on International Language Resources and Evaluation (LREC'10)*, pages 1871–1875.

Cunningham, H., Maynard, D., Bontcheva, K., and Tablan, V. (2002). GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proc. of the 40th Annual Meeting of the ACL*, pages 168–175.

Davies, M. (2009). Contemporary American English (1990-2008+). *International Journal of Corpus Linguistics*, 14(2):159–190.

Dennis, G., Sherman, B., Hosack, D., Yang, J., Gao, W., Lane, H. C., and Lempicki, R. (2003). DAVID: Database for annotation, visualization, and integrated discovery. *Genome Biology*, 4(5).

Dipper, S., Gotze, M., and Stede, M. (2004). Simple annotation tools for complex annotation tasks: An evaluation. In *Proc. of the LREC Workshop on XML-based Richly Annotated Corpora*, pages 54–62.

Draxler, C. (2005). Webtranscribe – an extensible Web-based speech annotation framework. In *Proc. of the 8th Text, Speech and Dialog conference (TSD'05)*, pages 61–68.

Francisco, V., Hervás, R., Peinado, F., and Gervás, P. (2011). EmoTales: Creating a corpus of folk tales with emotional annotations. *Language Resources and Evaluation*, 45.

Iida, R., Komachi, M., Inui, K., and Matsumoto, Y. (2007). Annotating a Japanese text corpus with predicate-argument and coreference relations. In *Proc. of the Linguistic Annotation Workshop*, pages 132–139.

Kaplan, D., Iida, R., and Tokunaga, T. (2009). Automatic extraction of citation contexts for research paper summarization: A coreference-chain based approach. In *Proc. of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*, NLPIR4DL '09, pages 88–95.

Kaplan, D., Iida, R., and Tokunaga, T. (2010). Annotation process management revisited. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, pages 3654–l3661.

Madnani, N. and Hwa, R. (2004). The UMIACS word alignment interface. `http://www.umiacs.umd.edu/~nmadnani/alignment`.

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of english: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Miltsakaki, E., Prasad, R., Joshi, A., and Webber, B. (2004). The Penn Discourse Treebank. In *Proc. of the 4th conference on International Language Resources and Evaluation (LREC'04)*.

Morton, T. and LaCivita, J. (2003). WordFreak: an open tool for linguistic annotation. In *Proc. of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Demonstrations - Volume 4*, NAACL-Demonstrations '03, pages 17–18.

Mueller, C. and Strube, M. (2001). MMAX: A tool for the annotation of multi-modal corpora. In *Proc. of the 2nd IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, pages 45–50.

Müller, C. and Strube, M. (2006). Multi-level annotation of linguistic data with MMAX2. In *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, pages 197–214.

Noguchi, M., Ichikawa, H., Hashimoto, T., and Tokunaga, T. (2006). A new approach to syntactic annotation. In *Proc. of the 5th conference on International Language Resources and Evaluation (LREC'06)*, pages 1412–1417.

Noguchi, M., Miyoshi, K., Tokunaga, T., Iida, R., Komachi, M., and Inui, K. (2008). Multiple purpose annotation using SLAT – segment and link-based annotation tool. *Proc. of 2nd Linguistic Annotation Workshop*, pages 61–64.

Park, S.-Y., Song, Y.-I., and Rim, H.-C. (2006). A segment-based annotation tool for Korean treebanks with minimal human intervention. *Language Resources and Evaluation*, 40(3-4):281–289.

Russell, B. C., Torralba, A., Murphy, K. P., and Freeman, W. T. (2005). Labelme: A database and web-based tool for image annotation. Technical report, Massachusetts Institute of Technology.

Stührenberg, M., Goecke, D., Diewald, N., Mehler, A., and Cramer, I. (2007). Web-based annotation of anaphoric relations and lexical chains. In *Proc. of the Linguistic Annotation Workshop*, pages 140–147.

Takahashi, T. and Inui, K. (2006). A multi-purpose corpus annotation tool: Tagrin. *Proc. of the 12th Annual Conference on Natural Language Processing*, pages 228–231. (in Japanese).

Widlöcher, A. and Mathet, Y. (2009). La plate-forme Glozz: environnement d'annotation et d'exploration de corpus. In *Actes de TALN 2009*, Senlis, France. (in French).

Timo Korkiakangas, Marco Passarotti

Challenges in Annotating Medieval Latin Charters

No annotation guidelines concerning substandard Latin are presently available. This paper describes an annotation style of substandard Latin that supplements the method designed for standard Latin by the Perseus Latin Dependency Treebank and the *Index Thomisticus* Treebank. Each word of the corpus can be assigned only one morphological analysis. In our system, the analysis can be either functional or formal. Functional analysis is applied when a form is language-evolutionarily deducible from the corresponding standard Latin form used in the same (semantico-)syntactic function (e.g. *solidus* pro *solidos* 'gold coins' as a direct object: analysis "accusative"). Formal analysis applies when no connection to the functionally required classical form exists (e.g. *heredibus* pro *heredes* 'heirs' as a subject: analysis "ablative" or "dative"). When running queries on the corpus, the formally analysed forms can be isolated, and percentages of standard and substandard forms can be counted. In addition, further principles concerning syntax and specific morphological issues are introduced.

1 Introduction

The present paper is related to a PhD project on the Latin case system in a corpus of ca. 500 Tuscan private charters (ca. 200,000 words) from the 8th and 9th centuries. So far, 1,452 sentences (28,488 words) have been annotated. Special attention is given to the core arguments (subjects and objects) and to prepositional phrases. The charters, published in three copyright-free diplomatic editions, have been digitised, proof-read and converted into XML.[1]

Research on the morphosyntax of the charters is performed by annotating the charters with the Latin Dependency Treebank (LDT) online tools provided by the Perseus Digital Library Project. The Latin and Ancient Greek Dependency Treebanks environment is suitable for our purpose, as it enables syntactic annotation, is user-friendly and publicly available.[2] Our annotation style is based on the *Guidelines for the Syntactic Annotation of Latin Treebanks* (BAMMAN et al. 2007[2]), which were launched to reconcile the practices of the annotators of LDT and the *Index Thomisticus* Treebank (IT-TB)[3] and to provide a general framework for all prospective treebanking projects in Latin. These guidelines and the related programs supporting annotation are designed for standard[4] Latin. The early medieval charters, however, differ from the standard in many respects (concerning orthography, morphology and syntax).

In this paper, we present a solution to the above-mentioned problem by introducing the concepts of formal and functional analysis plus further principles to supplement the existing guidelines. Even with these supplements, practical annotation requires highly subjective judgements on problematic cases, which is inevitable when dealing with charter texts and their language variety.

2 Standard and Substandard Latin

The *Guidelines for the Syntactic Annotation of Latin Treebanks* of LDT and IT-TB are designed according to the framework of dependency grammar as used on the analytical layer of annotation in the Prague Dependency Treebank (PDT) (HAJIČ et al. 1999) and adapted to Latin with the help of the *Latin Syntax and Semantics* of HARM PINKSTER (1990). Dependency grammar is an appropriate scheme of representation for highly inflected languages with a relatively free word order, such as Latin (BAMMAN et al. 2007[2], 3).

In LDT, both morphological and syntactic annotation is performed through a semi-automatic procedure provided by an online user interface. The morphological tagset reports information on the following: part of speech proper, person, number, tense, mood, voice, gender, case and degree. The syntactic annotation comprises syntactic tags (e.g. PRED, SBJ, OBJ, ATR, ADV) and head-dependent relations (BAMMAN et al. 2007[2], 4).[5]

If a word form already occurs in the treebank, the system provides its morphological analysis. If not, which is often the case when early medieval charters are concerned, the analysis must be typed manually in the table editor. If more than one analysis is provided by the system, annotators must choose the correct one from a drop-down menu. When combined, morphological and syntactic annotations allow performing advanced queries with *ad hoc* search engines, such as Annis used by LDT or Netgraph used by IT-TB.[6]

The Latin of the Italian charters from the 8th and 9th centuries is a technical, non-literary language variety resembling the style of the Lombard Laws. This variety seems to form a separate genre which is deliberately closer to the developments of spoken language than the literary texts of the same period, although it does not reflect spoken language directly nor is an attempt to act as a new "vulgar" language, distinct from Latin.[7]

The main issue concerning the Latin of early medieval charters is orthographic variation, which often concerns inflectional endings. These variations make it difficult to understand the syntactic structure of the texts. The existing annotation guidelines, designed for standard Latin, are not always able to manage substandard forms or standard forms used in a substandard way. Thus, new methods are needed with our corpus of medieval charters.[8] In standard Latin, each syntactic function is usually encoded by a relevant case form, which makes annotation process straightforward. However, with the Latin used in medieval charters, the equivalence between form and function is often not transparent.

3 The Solution: Functional and Formal Analyses

Each word in the corpus receives only one morphological tagging. In principle, we want to label all the forms functionally, i.e. according to their (semantico-)syntactic function in standard Latin. However, this is not always possible and several specifications are needed, mainly for nouns and other nominals.

If a word appears in its correct standard form, morphological tagging has no relevance since form and function are matching. If, however, a form is substandard, it is provided with a functionally based morphological analysis on condition that the form is language-evolutionarily deducible from the corresponding standard Latin form used in the same function. If no connection with the functionally required standard form exists, the substandard word is assigned a formal instead of a functional analysis.

For instance, we functionally annotate as accusatives the following substandard forms occurring as direct objects (although their form is not accusative) because they are meant to stand for the standard accusative forms: *solido* (standard: *solidum* 'gold coin'), *terra* (standard: *terram* 'land'), *testis* (standard: *testes* 'witnesses'), *solidus* (standard: *solidos* 'gold coins'). In *CDL* 23: *in tua cui supra emturi sit potestatem* (standard: *in tua cuius supra emptoris potestate* 'in the possession of you, the above purchaser'), the two words of the noun phrase *tua potestatem* ('your possession') are labelled functionally as singular ablatives dependent on the preposition *in*, although *potestatem* is formally an accusative singular in standard Latin. Finally, in *CDL* 45: *auris soledus trentas* (standard: *auri solidos triginta* '30 gold coins'), the standard ablative/dative plural form *auris* ('of gold') is functionally labelled as a genitive singular showing an additional *-s*.

Clear linguistic errors represent a class of their own and are always tagged according to their formal appearance. For instance, if a standard ablative/dative plural form, such as *heredibus* ('heirs'), functions as a subject (but does not occur in an ablative absolute construction), the form cannot be tagged functionally as a nominative because it is not possible to interpret it as a descendant of the nominative form. Thus, we label the *heredibus* according to its form, i.e. as ablative/dative plural. The form is an error probably due to the contamination between two or more formulae, a phenomenon common in medieval charters, or to the wrong interpretation of the abbreviation *hhd* (for *heredes*).

Sorting out such anomalous usages is relevant, as they are indirect (or "negative") clues of the corresponding, functionally correct form. This "negativity principle" represents, along with the functionality-formality approach, another pillar of our method. When running queries on the corpus, the distinction between formal and functional labelling allows us to isolate the formally analysed forms and to count the percentages of standard and substandard forms.

Both formal and functional labellings are based on standard Latin grammar. Although the language of these medieval charters may be quite different from standard Latin, analysing the charter texts in the framework of the traditional case system is justified because the charter texts try to resemble the standard language and, in spite of several disturbing factors, they reflect a multi-case system essentially similar to that of standard Latin. Adhering to standard Latin is also due to practical reasons: first, both LDT and IT-TB are based on standard Latin grammar; second, the language of the utilitarian texts of the 8th and 9th centuries, such as charters and laws, was never described in terms of prescriptive grammar similar to that of Classical Latin.

Distinguishing between functional and formal analyses is not the only possible method for annotating substandard Latin. In principle, one could also provide both types of annotations side by side, but this sort of multilevel annotation would be often redundant, as it would reduplicate the same information in most cases.

Another possible solution would be to provide functional analysis only, thus refining the query results according to the endings (for instance, by selecting all the subjects ending in *-ibus*). However, this solution would result in clearly erroneous analyses: for instance, the form *heredibus* would be tagged as "nominative". Our purpose is to provide morphological analyses that reflect the real language-evolutionary origin of the forms, in order to make

possible both to exploit the morphological tagging and to detect the 'anomalous' cases, i.e. those whose morphological tagging is incompatible with their syntactic function (reported by the dependency relation tag).

4 Additions to the LDT/IT-TB Guidelines

The principles described in the previous section are the backbone of our annotation style. This section introduces further specifications and individual rules designed in order to treat recurrent problematic structures consistently. This is of special relevance to morphology because it differs extensively from standard Latin.

4.1 Lemmatisation

Reducing lemmas. Almost all the words in the charters have two or more graphical variants. Likewise, one single morph may have several realisations. Therefore, particular attention must be paid to lemmatising all its graphical variations under one common lemma in order to avoid proliferation of lemmas in the Perseus Dynamic Lexicon database (BAMMAN – CRANE 2008, 11–13). For instance, nouns facing gender change, such as the masculine nominative plural *saeculi* ('centuries'), as well as adjectives facing declension change, such as the second declension nominative singular *inanus* ('void'), are lemmatised under the standard lemmas: *saeculum* (neuter) and *inanis* (third declension), respectively. The aim is to respect the choices taken by the scribe as far as they are traceable. This is also the motive for formally labelling those functionally impossible case forms, such as *heredibus*, in order to show their anomalous status.

Proper names. Several Germanic and Latin proper names exhibit much variation. For example, the form *Delmati* is lemmatised under *Dalmatius* and the forms *Guntifrido* and *Cuntefrid* under *Guntifridus*. However, it is sometimes difficult to establish the correct lemma, as no variant seems to be more justified (or more frequent) than the others. In the charters, there are also several unidentified place names. Unknown second declension toponyms, such as *Brancalo*, are lemmatised as neuters ending in *-um*. Although in some cases the lemma can be reconstructed on the grounds of the modern name of the place in question[9], those names that are completely opaque must be labelled as "unknown".

4.2 Syntax

Omitted elements. As our research focuses primarily on the syntactic constructions concerning the core arguments (subjects and direct/indirect objects) and prepositional phrases, we leave unannotated all the non-nominal adverbials, except negation particles, and the punctuation marks, except those commas which have a role in coordinated or appositive tree structures (cf. the lacking "," and "*et*" in Figure 1). Terminal punctuation marks are always tagged with the technical label AuxK (BAMMAN et al. 2007[2], 33–34).

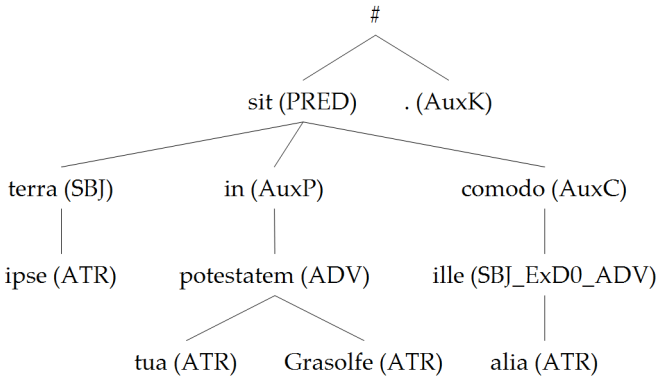


Figure 1: Dependency tree of *CDL* 220: *ipse terra in tua Grasolfe sit potestatem, comodo et ille alia*.

Ellipsis and fragmentary parts. In annotating ellipsis, we follow the LDT style and reconstruct the omitted nodes. In fact, the formulaicity of the charters very often allows deducing even the exact wordings of the missing parts with a high degree of reliability. For example, in *CDL* 220: *ipse terra in tua Grasolfe sit potestatem, comodo et ille alia* ('this plot be in your possession, Grasolfus, just as that other one'), the omitted verb of the subordinate clause *comodo* (*sit*) *et ille alia* ('just as (be) that other one') is reconstructed through the complex tag SBJ_ExD0_ADV. This means that *ille* ('that') is the subject (SBJ) of the omitted verb (ExD0, "externally dependent", is the technical label for missing items) that in the tree would be the head of an adverbial (ADV) subordinate clause (see Figure 1). This is the only aspect where the annotation style of IT-TB differs from the LDT one. As a matter of fact, IT-TB (as PDT) does not resolve the ellipsis and, thus, would assign to *ille* the simple tag ExD. In those cases where an elliptic structure is ambiguous or where words are missing because the original source is damaged, we follow the IT-TB style and link the orphan nodes directly to their assumed parents via ExD (BAMMAN et al. 2007[2], 36–37; BAMMAN et al. 2007[1], 4).

Indirect objects. We introduce a specific tag (c="1") to annotate indirect objects while LDT and IT-TB use the same label OBJ for both direct and indirect objects. Even though the latter solution is suitable for standard Latin, where indirect objects always occur in dative or as prepositional phrases, it cannot be applied to our texts, which feature a high degree of formal variation. In *CDL* 125: *in terra*, *que offerui sancti Petri cum ipsa fossa* ('in the plot, which we donated to St. Peter, with the ditch'), the direct object is *que* (standard: *quam*) and the indirect object is *sancti Petri* (standard: *sancto Petro*). Although they are both labelled with OBJ (see Figure 2), *sancti Petri* is assigned the additional tag c="1" in order to make clear its status as an indirect object.[10] In this case, the morphological formal analysis of *sancti Petri* (genitive singular) also helps to detect the anomaly.

Vocatives. Although the Guidelines demand to link the vocatives to their verbal heads with the label ExD (BAMMAN et al. 2007[2], 41), we link them to their nominal heads via ATR since, in our charters, the vocatives mainly represent the function of nominal attributives.

See, for example, the words *Uuarniperte* and *Lamprande* in *CDL* 269: *uouis Uuarniperte et Lamprande presbiteri* ('to you, priests Warnipertus and Lamprandus'), and *Grasolfe* in Figure 1: *in tua Grasolfe sit potestatem* ('in your possession, Grasolfus').

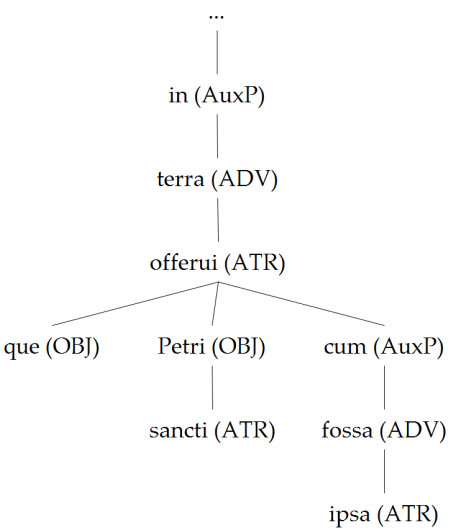


Figure 2: Dependency tree of *CDL* 125: *in terra, que offerui sancti Petri cum ipsa fossa*.

4.3 Morphology

Subjects. The following annotation style only applies to subjects of clauses whose verb occurs in finite form. The standard case of a subject headed by a finite verb is nominative. The subjects of *accusativus cum infinitivo* constructions and ablative absolutes are not discussed here. In standard Latin, they are encoded with accusative and ablative, respectively.

The second and fourth declension masculine singular subjects ending in <*-o -u -um*>, such as the second declension form *Deo* (standard: *Deum* 'God'), are tagged formally as accusatives because, according to the bicasual hypothesis, they cannot be deduced from the standard nominative form. The neuter subjects ending in <*-o -u -um*>, such as *pretio* (standard: *pretium* 'price'), are tagged functionally as nominatives. In principle, the neuter subjects could equally well be tagged as accusatives because the standard nominative and accusative forms of the second declension neuters are identical.

The formal tagging also applies to those third declension singular subjects ending in <*-e -em*> whose stem has an additional syllable in all cases except nominative, such as nominative *potes-tas* vs. accusative *potes-ta-tem* ('possession'). Instead, those third declension singular subjects ending in <*-e -i -em*> whose stem has the same number of syllables in all cases, such as nominative *tes-tis* vs. accusative *tes-tem* ('witness'), and all the first declension singular subjects ending in <*-a -am*> are tagged functionally as

nominatives because the word-final /*m*/ was no more pronounced in Late Latin. For example, *potestate* and *potestatem* as subjects are tagged as accusatives, while *teste* and *testem* as subjects are tagged as nominatives.

These principles are based on the following reason: When annotating third declension subjects with an equal number of syllables in all cases, no distinction can be made between the language-evolutionary outcomes of the standard nominative and accusative forms. However, in the second and fourth declension subjects as well as in the third declension subjects with an additional stem syllable, the nominative and the accusative forms still differed from each other because the second and fourth declension final /*s*/ in the nominative and the third declension stem extension mark the nominative as distinct. Indeed, the opposition between the second declension nominative singular and accusative forms seems to have been partially neutralised in the Latin of Tuscany in the 8th and 9th centuries, but the bicasual assumption is a good working hypothesis (cf. ZAMBONI 2000, 233–235, 243–244). Cases such as the second declension subject *Deo* (tagged as an accusative) illustrate the role of subjective decisions in the annotation process: determining the status of certain forms is not uncontroversial, but the decisions can be systematic if they are based on a well-grounded theory. Indeed, the annotation style depends considerably on the chosen theoretical framework, and the choice of the annotation framework is dictated by the purpose of the corpus. As our corpus is mainly designed for studying case marking, the bicasual assumption seems to be a valid background assumption for our annotation style.

In the fifth declension, the confusion appears to be so massive that the analyses must be especially delicate. In the plural forms of all the declensions, the deviations from the standard forms are fewer.

Genitives and the oblique case. In late substandard Latin, the genitive was often replaced by an oblique case, most likely derived from the standard accusative. The development, which started in spoken language, led to a situation where the standard case system was probably reduced to a bicasual system (nominative vs. accusative). The accusative gradually absorbed the functions of all the other cases except the nominative and finally even that of the nominative (VÄÄNÄNEN 1981, 116–117; cf. ZAMBONI 2000, 248). The oblique forms are tagged formally as accusatives. For instance, in the subscription *CDL* 261: *signum manus Alprand filio quondam Teudualdi testis* (‘mark of the hand of Alprandus, son of late Teudualdus, witness’), the word *filio* (‘son’) is labelled as an accusative and linked to its head *Alprand* via ATR.

Prepositions. As a general principle, we label as accusatives the complements of prepositions governing accusative in standard Latin and as ablatives the complements of prepositions governing ablative, if the case-endings can be claimed to represent the original accusative and ablative forms, respectively. This requires looking at the meanings of the prepositional phrases, as some prepositions govern different cases according to what they mean. For instance, the prepositions *in* and *super* govern accusative when expressing motion and ablative when expressing state. In *CDL* 23: *sup die quartum* (‘on the fourth day’), we label both *die* (‘day’) and *quartum* (‘fourth’) as singular ablatives since *sub* governs ablative if it means state, and accusative if it means motion.

Nominal attributives. Nominal attributives occur mainly in the titles of commissioners and addressees of legal transactions, for example in *CDL* 266: *ego Autulu uir religiosus clirico filio quondam Bonuald de uico Turrite* ('I Autulus, *vir religiosus*, clerk and son of late Bonualdus from the village of Turrite'). Several problems arise when the head-dependent relations in such noun phrases are labelled. As a rule, we choose as the head of the noun phrase the member with the highest ranking in the following hierarchy of animacy: personal pronouns > proper names > other nouns referring to humans. Thus, the head of the above noun phrase is *ego*, under which *Autulu* is attached as an attributive; *uir religiosus*, *clirico* and *filio* are then linked to *Autulu* as attributives.

Absolute constructions. Some substandard absolute constructions, such as the accusative absolute and the *post* construction, had been quite firmly established even in the late literary language (HELTTULA 1987, 6–7, 91–92). As far as morphological annotation is concerned, we do not force the absolute structures into the form of standard Latin. In the medieval charters, almost all case forms can occur in absolute constructions, and we do not want to reduce such formal variety to any expected pattern, such as accusative absolute, because we take into account the scribes' freedom in choosing the case form in absolute constructions.

This applies, for instance, when a case form might be interpreted as a descendant of the standard ablative. For example, in *CDL* App. *postea*, *inimicum eum suadente* (standard: *inimico eum suadente*), *inuolauit mihi ipsam cartulam* ('later on, he stole me the charter, incited by the Devil'), the noun *inimicum* ('Devil') can be interpreted as an ablative, but the structure rather seems to be an accusative absolute.

Post constructions are treated as if they were normal prepositional constructions. Examples of these are *MED* 424: *post fructum de ipsa res recollecto* ('having collected the yield on this property') and *CDL* 260: *spondeo … conponere tibi post hanc cartulam ostensam … quae tibi subtraxerimus* ('I promise … to compensate you for what we may have seized from you, if this charter is brought in evidence') (see Figure 3).

Vocatives. The label "vocative" is assigned only to the forms showing a clear vocative ending, such as *Uuarniperte* and *Lamprande* in the above-mentioned *CDL* 269: *uouis Uuarniperte et Lamprande presbiteri*; the form *presbiteri* is tagged as a nominative plural.

Gender change. Gender change from neuter singular to masculine singular and from neuter plural to feminine singular is a relevant example of the changes occurring in Latin declension. In our annotation style, the neuters occurring in masculine or feminine forms are lemmatised under their standard lemmas and still labelled as neuters. For instance, see *pretius* ('price', masculine) in *CDL* 66: *suscipemus … pretius* (standard: *suscepimus … pretium* 'we received … the price', neuter), or *adiacentia* ('neighbourhood', feminine) in *CDL* 266: *cum omnem adiacentia sua ... pertenente* (standard: *cum omnibus adiacentibus suis … pertinentibus* 'with all its neighbourhood … that belongs to…', neuter). Thus, the annotation does not reveal gender change. This is only revealed when the words labelled as neuters are sorted by their endings or when they are read in their context, as in *cum omnem adiacentia sua ... pertenente*.

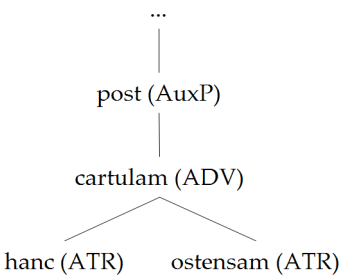


Figure 3: Dependency tree of *CDL* 260: *post hanc cartulam ostensam*.

Number and person in verbs. If it is not possible to determine the number of a verb, it is tagged according to its formal appearance. This phenomenon mainly occurs with the third person of verbs expressing actions performed by the addressee of a charter, as it is often unclear whether the addressee is acting alone or with his/her heirs. However, in *CDL* 23: *petras que iniui esse uiditor* ('the stones that are (lit. is) seen there') the singular verb *uiditor* may be due to impersonalisation of the passive structure (GIANOLLO 2005, 100). The relative pronouns (*que* 'that') were already on their way to becoming indeclinable.

The person of the verb is usually tagged functionally because the person is normally easier to recognise than the number. The context may be helpful: for instance, in *CDL* 28: *abbas … habeas* ('the abbot … may have'), the form *habeas* 'you may have' is analysed as a third person singular (standard: *habeat*). In more complex cases, such as *donatores … habeas* (standard: *habeant* 'the donators … may have'), the annotator has to make delicate decisions which depend on the amount of graphical variation observed in the charter.

5 Two Case Studies

In order to demonstrate how helpful the distinction between formal and functional analysis is for organising and retrieving data, we briefly report two case studies concerning two simple constructions which occur in our corpus.

The first construction concerns those prepositional phrases that are headed by the preposition *ad* ('to'). In Latin, the preposition *ad* governs nouns and pronouns inflected in the accusative case. In our corpus, however, several exceptions to this rule occur: these exceptions can be retrieved by exploiting the annotation. Table 1 reports the results concerning this construction.

	Non-accusative case	Accusative case
Formal = Functional	---	81
Functional	---	119
Formal	19	---
TOTAL by case	19	200

TOTAL	219

Table 1. The results concerning prepositional phrases headed by *ad*.

In the part of our corpus annotated so far, there are 219 lexical items governed by the preposition *ad*. Among them, only 19 are tagged formally as clearly non-accusative forms showing no connection with the functionally required standard form. We report two examples of such items, one presenting a direct governance and the other showing a coordinated governance: *ad heredibus uestris* ('to your heirs') and *ad Laurentio et Ualentini* ('to Laurentius and Valentinus').

The remaining 200 items are tagged as accusatives. Among them, 81 are standard accusative forms (*ad ecclesiam*: 'to the church') and, thus, their formal and functional tagging are matching. The remaining 119 items are substandard accusative forms (language-evolutionarily deducible from the corresponding standard Latin forms): hence, they are tagged functionally as accusatives (*ad ecclesia*).

The second construction is ablative absolute. This construction consists of one participle (in the ablative case) and one subject (also in the ablative case). Table 2 presents the results concerning this construction.

	Subject in non-ablative case	Subject in ablative case
Formal = Functional	---	47
Functional	---	71
Formal	18	---
TOTAL by case	18	118
TOTAL	136	

Table 2. The results concerning ablative absolute.

On a total of 136 items occurring as subjects of ablative absolute constructions, only 18 are annotated formally as clearly non-ablative forms. One example is *Dominus interueniente* ('with the intervention of God'), where *Dominus* is a nominative form. Most of the ablative absolute constructions present a subject tagged as ablative (118 cases). Among these, 47 are tagged as standard ablative forms (*regnante Liutprando*: 'under the reign of Liutprand'); 71 are substandard ablative forms, which are tagged functionally (*regnante Liutprando*).

6 Conclusions

In order to overcome the incompatibility between the annotation of Latin in the medieval charters and the annotation style provided by the LDT/IT-TB guidelines, two distinct forms of analysis (formal and functional) and a number of additional principles were introduced.

Four issues can be distinguished: (a) functional analysis is applied when a form is deducible from the corresponding standard Latin form used in the same function; (b) formal analysis is applied when a form is not deducible from the standard Latin form used in the same function; (c) the linguistically impossible forms can be isolated when querying the

data; (d) the query results of the data can be further processed by classifying the results according to endings; the percentages of standard, early medieval and linguistically impossible forms can be counted.

Building and querying an annotated corpus of substandard language that shows much variation is a challenging task. An inherent disadvantage of introducing new rules in annotation is that the corpus becomes more difficult to use. The user must consider several different parameters that were applied when building the annotated corpus. This, along with separating formal and functional labellings, implies that the pure quantitative results from the queries on our corpus cannot be compared with those acquired from corpora in standard Latin. However, following the same general principles of syntactic annotation (in terms of theoretical framework, syntactic labels and head-dependent attachment) allows us to compare the syntactic constructions occurring in our corpus with those of LDT and IT-TB.

Acknowledgements

We would like to thank the anonymous reviewers for their valuable comments and suggestions to improve the paper, Hilla Halla-aho, PhD, who commented an earlier version of the paper, and Leena Enqvist, MA, who proof-read the final version.

---

[1] The three editions are *Codice diplomatico longobardo* (*CDL*) 1–2 (LUIGI SCHIAPARELLI, 1929–1933); *Codice diplomatico toscano*, part 2, vol. 1 (FILIPPO BRUNETTI, 1833) and *Memorie e documenti per servire all'istoria del Ducato di Lucca* (*MED*), part 5, vol. 2 (DOMENICO BARSOCCHINI, 1837). *CDL* is digitised and proof-read by the Institut für Mittelalterforschung of the Austrian Academy of Sciences while the other two are digitised by Google and proof-read by us. Almost all the charters were also published recently in the *Chartae Latinae Antiquiores* (2nd series).

[2] The Perseus Latin and Ancient Greek Dependency Treebanks are projects aimed at treebanking texts in Classical Latin and Greek; they are both hosted at Tufts University in Boston, USA (*http://nlp.perseus.tufts.edu/syntax/treebank/index.html*). Another project in the field is the Laboratoire d'Analyse Statistique des Langues Anciennes in Liège, Belgium (LASLA, *http://www.cipl.ulg.ac.be/Lasla/*). The annotation style of syntax by LASLA concerns subordination patterns only.

[3] The *Index Thomisticus* Treebank is an ongoing project aimed at the syntactic annotation of the *Index Thomisticus*, a morphologically annotated corpus of the texts of St. Thomas Aquinas. The project is hosted at the Catholic University of the Sacred Heart in Milan, Italy (*http://itreebank.marginalia.it)*.

[4] By "standard" Latin we mean the variant of Latin mostly used by the Classical authors and reported in the pedagogical grammatical tradition.

[5] For the morphological tagset, see the README file for the Latin Dependency Treebank at *http://nlp.perseus.tufts.edu/syntax/treebank/ldt/1.5/docs/README.txt*.

[6] The Annis search engine is not yet publicly available. The *Index Thomisticus* Treebank can be browsed through Netgraph at *http://circse.marginalia.it/~passarotti/netgraph/client/applet/NGClientAen.html*.

[7] See BARTOLI LANGELI 2006, 24–28 about the status of the Latin of the Lombard charters and laws.
[8] See PHILIPPART DE FOY [forthcoming] about changes in the LASLA annotation procedures to face similar problems in a medieval hagiographic corpus.
[9] The *Chartae Latinae Antiquiores* editions usually report the modern equivalents of the place names occurring in the charters.
[10] The verb heading a relative clause is linked to its antecedent as an attributive (ATR) (BAMMAN et al. 2007[2], 37–38).

References

BAMMAN, D. – CRANE, G. (2008). „Building a Dynamic Lexicon from a Digital Library". In: Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2008, Pittsburgh). New York: ACM, 11–20.

BAMMAN, D. – PASSAROTTI, M. – CRANE, G. – RAYNAUD, S. (2007[1]). „A Collaborative Model of Treebank Development". In: Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories (TLT2007, Bergen), 1–6.

BAMMAN, D. – PASSAROTTI, M. – CRANE, G. – RAYNAUD, S. (2007[2]). Guidelines for the Syntactic Annotation of Latin Treebanks (v. 1.3). *http://nlp.perseus.tufts.edu/syntax/treebank/ldt/1.5/docs/guidelines.pdf*

BARTOLI LANGELI, A. (2006). Notai. Scrivere documenti nell'Italia medievale. Roma: Viella.

CDL = Codice diplomatico longobardo 1–2. A cura di LUIGI SCHIAPARELLI. (1929–1933). Roma: Tipografia del Senato.

GIANOLLO, C. (2005). „Middle Voice in Latin and the Phenomenon of Split Intransitivity". In: Calboli, G. (ed.) (2005). Latina lingua! Proceedings of the Twelfth International Colloquium on Latin Linguistics (ICLL 2003, Bologna). Roma: Herder, 1: 97–110.

HAJIČ, J. – PANEVOVÁ, J. – BURÁŇOVÁ, E. – URešOVÁ, Z. – BÉMOVÁ, A. (1999). Annotations at Analytical Level. Instructions for annotators. Institute of Formal and Applied Linguistics, Prague. *http://ufal.mff.cuni.cz/pdt/Corpora/PDT_1.0/References/aman_en.pdf*

HELTTULA, A. (1987). Studies on the Latin Accusative Absolute. Tammisaari: Societas Scientiarum Fennica.

MED = Memorie e documenti per servire all'istoria del Ducato di Lucca 5:2. A cura di DOMENICO BARSOCCHINI. (1837). Lucca: Francesco Bertini.

PHILIPPART DE FOY, C. [forthcoming] „Lemmatiser un corpus de textes hagiographiques: enjeux et modalités pratiques". In: Biville, F. (ed.) Latin vulgaire – latin tardif IX. Actes du IX[e] colloque international sur le latin vulgaire et tardif (LVLT 2009, Lyon).

PINKSTER, H. (1990). Latin Syntax and Semantics. London: Routledge.

VÄÄNÄNEN, V. (1981). Introduction au latin vulgaire. Paris: Éditions Klincksieck: (Bibliothèque française et romane, A:6.

ZAMBONI, A. (2000). „L'emergere dell'italiano: per un bilancio aggiornato". In: Herman, J. – Marinetti, A. (eds.) (2000). La preistoria dell'italiano. Atti della Tavola Rotonda di Linguistica Storica. Università Ca' Foscari di Venezia, 1999. Tübingen: Niemeyer, 231–260.

Cerstin Mahlow, Britta Juska-Bacher

Exploring New High German texts for evidence of phrasemes

Most dictionaries containing phraseological information are restricted to a synchronic perspective. Diachronic information on structural, semantic, and pragmatic change over time has to be reconstructed by a time-consuming consultation of various dictionaries providing only punctual insights. In the OLdPhras, project we construct an online dictionary for diachronic phraseology in German from ca. 1650 to the present by combining dictionary exploration with corpus-based methods. This paper highlights some challenges we have met: How to select the "interesting" phrasemes, i.e., those that underwent some change? How to deal with historical corpora? How to include different kinds of phraseme variation? We present a semi-automatic corpus-based approach for the investigation of phraseme development. We argue for a combination of dictionary exploration and corpus-based methods to provide reliable and extensive information about the diachronic development of German phrasemes.

1 Introduction

Phraseology as a subfield of linguistics investigates form, meaning, use, and change of phrasemes (also referred to as phraseological units, idioms, or set phrases). Phrasemes are defined by polylexicality, relative stability, and idiomaticity (Burger, 2010, 36ff). Dictionaries—whether printed or electronic ones—usually describe the characteristics of phrasemes at a certain point in time, i.e., they are restricted to a synchronic perspective. For example, in a contemporary dictionary of German phraseology, one finds information on the current meaning of phrasemes such as *gegen den Strom schwimmen* ("to swim against the current", i.e., 'to oppose the opinion or the habits of the majority') and an example. General-purpose dictionaries give fewer explanations on a phraseme's development; etymological dictionaries like Kluge (2002) only provide information on the development of single words, not of multi-word units.

Metalexicographers have repeatedly criticized the neglect or the unsystematic presentation and placement of phrasemes in general and phraseological dictionaries (Kühn, 2003; Stantcheva, 2003; Burger, 2010). Often users cannot determine whether an example given represents established usage and can be found in real-world texts or whether it was made up by the author of the dictionary. In current phraseography research, empirical methods, i.e., analyzing large corpora, are emphasized to overcome some of these problem (see for example Mellado Blanco, 2009).

The project "German Proverbs and idioms in language change. Online-dictionary for diachronic phraseology (OLdPhras)" (started August 2010) funded by the Swiss National Science Foundation aims to provide information on the development—i.e.,

structural, semantic, and pragmatic change—of German phrasemes from ca. 1650 to the present. The resulting electronic dictionary is intended to serve researchers as a resource for further investigations as well as interested laypersons for information purposes. Alongside common lexicographic information on lexical units and their grammatical structure, we focus on evidence concerning lexical and semantic variants of phrasemes, and on changes in meaning and use by offering authentic examples from (New) High German texts. Comments on diachronic development will be based primarily on evidence from corpora considering observable characteristics with respect to lexical units, syntactic and morphosyntactic properties, semantic concepts, or pragmatic aspects. Additionally, we consider synchronic information from existing dictionaries covering that period and we describe usage and meaning at the time of a specific evidence, see Juska-Bacher and Mahlow (2012) for an example of results to be expected by using the example of *gegen den Strom schwimmen*.

In this paper we first look at the state of the art with respect to handling German phrasemes and refer to related work. Then we comment on the resources used in the OLdPhras project and outline specific challenges. We present our approach, with particular attention on how to overcome some of the obstacles due to the diachronic perspective, and report on first findings; as this is still ongoing work, we cannot provide extensive evaluations.

2 State of the art and related work

Defining phrasemes as *non-Fregian collocations*, we can search texts for potential collocations, which then have to be classified by phraseologists with respect to idiomaticity, resulting in a semi-automatic process. In our project we face two of the challenges pointed out by Rothkegel (2007, 1027): exploring which phrasemes are used in which forms and variants, and automatically identifying phrasemes in texts.

Fritzinger et al. (2009) propose the extraction of potential collocations for German to be based on fully syntactically parsed text. However, the implementation seems to be prototypical only. Seretan and Wehrli (2010) report on the extraction of collocations as preprocessing step to make the work of lexicographers easier. Rothkegel (2007) and Heid (2007) present various attempts to extract collocations from texts in different languages, Evert (2005) compares various approaches. A common feature is the attempt to reach high recall—to not miss a potential phraseme—resulting in rather low precision, requiring manual efforts to identify phrasemes.

All studies try to identify whether there are *any* collocations in a given text; there is no previous assumption on what to expect. Applying these approaches would be a completely corpus-driven method. In contrast, in the OLdPhras project, we follow a rather corpus-based approach—given a set of "interesting" phrasemes (see section 5.1) we explore corpora for evidence.

However, the approaches and methods developed so far are applied to *modern* texts. Annotated databases like Kuiper et al. (2003) describe phrasemes at a certain point in time, which then can be used in NLP tools. For our project with a strong diachronic

focus, two major issues arise: First, methods and resources developed for or trained on modern texts cannot be applied to texts from older language stages—there are differences with respect to orthography, lexicon, morphology, as well as syntax. Second, existing electronic resources like dictionaries or lexical databases (e.g., Kuiper et al. (2003)) reflect the *current* state of a language. We can distinguish compositional and non-compositional phrasemes in today's language use, but our interest is the point in time when arbitrary multi-word units started to be used in an idiomatic way rather than literally, or when certain kinds of variation of a phraseme were not used or not even allowed anymore. Synchronic resources (e.g., collections, tools) may thus only serve as a starting point. The lexicon developed by Keil (1997) for NLP purposes and used in experiments by Fischer and Keil (1996) seems to be no longer available; however, the proposed structure is of interest for our purposes.

The electronic resource most closely related to the aims of our project is the *Idiomdatenbank*, developed between 2003 and 2006 at the Berlin-Brandenburgische Akademie der Wissenschaften, see Fellbaum (2007); Fellbaum and Geyken (2005). However, *OLdPhras* is not simply an extension, but poses specific challenges related to a very simple fact: The time period to be investigated is longer (350 years vs. 100 years). During 350 years, more variation and change can be expected than during 100 years—more variation is assumed the further back in time we go (Burger and Linke, 1998). Change might occur on all levels relevant to a phraseme: (a) spelling, (b) lexical components, (c) syntactical structure of the multi-word unit, (d) semantics of single units, (e) semantics of the multi-word unit, (f) pragmatics. Taking as many levels of change as possible into account, we have to consider a wide range of possible instantiations of one phraseme—preferably automatically formulated as search string—to be looked up in a corpus.

3 Resources

3.1 Corpora

In recent years there have been several attempts to create diachronic corpora for German for various research perspectives. Given our interest in the time from 1650 until today, two corpora are specifically relevant: *Deutsches Textarchiv* (= DTA)[1] with 532 texts from 1650 to 1900 and *GerManC* with texts from 1650 to 1800 (Bennett et al., 2010). DTA aims to make available the most relevant cross-disciplinary German-language books. GerManC aims to provide "a basis for comparative studies of the development of the grammar and vocabulary of [...] German and the way in which they were standardized."[2] The corpus consists of representative 2000-word samples from nine genres from various regions.

While DTA makes available whole texts under a Creative Commons License, GerManC provides snippets only. However, both projects aimed to digitize the most authentic

[1] http://www.deutschestextarchiv.de
[2] http://www.llc.manchester.ac.uk/research/projects/germanc

versions of the texts, i.e., first editions. There are other freely available collections of texts from the relevant time, like the online library provided by TextGrid—also under a Creative Commons License—with texts from the beginning of publishing until the beginning of the 20th century[3]. This collection is volume 125 from the "Digitale Bibliothek" (DB125), consisting of roughly 2700 fictional texts with about 87 million running word forms. These texts are usually later editions.

We will also be able to explore two special-purpose collections: *Text+Berg digital* (Volk et al., 2010) consisting of the yearbooks of the Swiss Alpine Club from 1864 to 2009 with 36 million running word forms, and a subset of the *Collection of Swiss Law Sources* (Gschwend, 2007), containing about 4 million running word forms in texts from ca. 1000 to 1798 (Piotrowski, 2010).

3.2 Dictionaries and collections

Synchronic phraseological information at various points in time can be found in general-purpose dictionaries like Adelung (1801) (= Adelung), Campe (1812), or Sanders (1865), or in special-purpose dictionaries like Wander (1880) (= DSL), Friedrich (1976), or Dudenredaktion (2008) (= Duden11). These dictionaries list phrasemes known and used at a specific point in time; some, like (Grimm and Grimm, 1971) (= DWB) or Borchardt (1888) integrate some diachronic and etymological information to different extents. Etymological information provided by the folklorist Röhrich (2002) often tends to be quite vague, using expressions like "originally" or "formerly". Looking at one dictionary or collection at a time, we get mainly synchronic impressions; moreover, the sources are rarely, if ever, identified, and of course every dictionary claims to be the most authoritative one.

If a phraseme in Duden11 is not listed in DSL, like *einen Quantensprung machen* ("to make a quantum leap", 'to make huge progress'), this might be evidence that this specific phraseme was not known 130 years ago, possibly (as in this case) because one of the lexical units or the concept was not known then. Vice versa, if a phraseme is listed in an older dictionary, but not in a modern one, this might be evidence that this phraseme is not used any more, like *einen Krebs im Beutel haben* ("to have a crab in the bag", 'to be short of money').

To get a first impression of diachronic change, we explored existing dictionaries and collections from different dates. Some, like DSL were already available in electronic format, others, like Duden11 were digitized by us for internal use.[4] Comparing listed phrasemes allows phraseologists to decide which phrasemes to inspect further because of (potential) changes in use, meaning, and/or pragmatics. After analyzing the dictionary data and selecting phrasemes, we will search for evidence in the corpora described in section 3.1 and annotate the results to serve as source for describing diachronic changes.

Fischer and Keil (1996) distinguish non-compositional and compositional idioms, referring to syntactic and semantic flexibility, the latter allowing to vary parts of

[3] `http://www.textgrid.de/digitale-bibliothek.html`

[4] In the meantime, Duden11 is available online.

the phraseme by adjectival modifications, quantification, or by using demonstrative determiners. We do not follow this differentiation, and allow for variation in all phrasemes of interest. There is also no consensus among phraseologists on how to define and to distinguish *variants* and *synonyms*. We therefore provide information about the characteristics of similarity of two examples concerning form and meaning. This allows for searching and displaying phrasemes (or their instantiations in our corpus) by formal aspects—e.g., sharing similar syntactic structure or lexical units—or by meaning—e.g., expressing similar semantic concepts.

4 Challenges

Given our aim and considering the limited resources with respect to manpower and time, we face several challenges.

First, it is impossible to investigate *all* phrasemes listed in one or more of the available collections: we were able to extract 33,200 potential phrasemes from Küpper (1997) (=WdDU), 45,729 potential phrasemes from DSL, 11,500 from "Redensartenindex" (RA-I)[5], 13,300 from Duden11, etc. Due to different conventions for formulating the citation form used by the authors or editors of these collections and dictionaries, it is impossible to compare the entries as such to detect phrasemes listed in more than one collection. However, even given the smallest number of extracted phrasemes—3'834 manually annotated phrasemes from Adelung—it was too much to explore the diachronic evolution of all of them. We therefore had to select phrasemes meeting several constraints: they should be of interest for the intended audience (i.e., researchers as well as laypersons), the phrasemes should have undergone some change over time, the sample should not be restricted to the most common or most unknown phrasemes used today, and there should be a certain frequency of occurences of the phrasemes in corpora. In section 5.1 we report on this aspect.

Second, the corpora of interest for us do not come in comparable formats. All corpora mentioned in section 3.1 are annotated according to the TEI guidelines. TEI P5 (Wittern et al., 2009) allows projects to use various subsets of TEI, so that actual annotation may differ between corpora; however, most of the corpora are not deeply annotated, we can reduce all annotation to the most shallow one, allowing for an easy mapping between annotations to provide a common ground. What is more important, although for example DTA put a lot of effort in normalizing and lemmatizing the texts (with very good results, see for example Jurish (2010)), users can download the non-lemmatized texts only. The Collection of Swiss Law Sources also provides no normalization or lemmatization. The same is true for the TextGrid library. Only the small corpus of alpine texts provides lemmatized texts. As we cannot investigate normalization or lemmatization of old German language variants during project time and as today there are no such tools available providing reasonable quality to be applied without further

[5] `http://www.redensartenindex.de`

effort, we cannot use standard corpus-linguistic tools as in the Idiomdatenbank project. In section 5.2 we report on first attempts to overcome this issue.

Third, searching for phrasemes in corpora means looking for evidence of a sequence of words allowing for inclusion of particles or adjectives as well as for morphological and syntactical variation. Whether a found sequence is indeed a phraseme or whether the words are used literally, can only be decided by looking at the context; in most cases this decision has to be made by the phraseologist, who generally cannot rely on intuition if it comes to older texts. It is hardly possible to reduce this manual effort (see also Rothkegel (2007)).

5 Approach

From a diachronic point of view, polylexicality involves language changes on various levels—structure and meaning of the lexical components as well as structure and meaning of the whole phraseme. When deciding on which phrasemes to investigate we have to allow for changes on all levels to find evidence for these phrasemes in our corpora. As mentioned in section 4, we have to define a sample of phrasemes used for investigation. The OLdPhras dictionary will contain entries with detailed descriptions of their development, while for others we will provide selected information only. We will first report on how to select this sample and we then develop searching strategies for our corpora.

5.1 Choosing the sample of phrasemes for investigation

From Adelung and DSL (both in digitized versions) we extracted phrasemes representing German in the late 18th and in the 19th century—the *historical phrasemes* (HP). For DSL we could make use of typographical structuring of the entries and extract all potential phrasemes automatically. As there was no such typographical structure used in Adelung, the text was annotated manually using the author's markers as starting point[6]—which had the advantage that information concerning meaning, use, and variation could be annotated as belonging to a specific phraseme at the same time. We used *Stripey Zebra*, the current version of the German Malaga Morphology (Lorenz, 1996) to identify nouns used in the extracted phrasemes.[7] The continuously most frequent nouns indicate a constant productivity of components: for somatisms, i.e., words for body parts, like hand, head, eye, ear, or nose, we find a great number of phrasemes.[8]

[6] Adelung used markers like "Sprichwort", "Redensart" or "RA", but not consistently.

[7] *Stripey Zebra* is a rule-based morphological analyzer, providing detailed, hierarchically structured results; using pruning and weighting *Stripey Zebra* can provide "the best" analysis according to the morphological principles of derivation, compounding, and inflection. For unknown words a hypothesis is generated. See Mahlow and Piotrowski (2009) for a detailed description and performance data.

[8] Using a modern morphological analyzer like *Stripey Zebra* poses some bias, as older spelling variants might result in wrong results or no results at all. However, manual inspection showed

Extracting potential phrasemes from contemporary sources—Duden11 and RA-I, the *contemporary phrasemes* (CP)—by making use of typographical information and using lemmatization again, we could compare rankings of nouns. We were interested in nouns being part of a large variety of phrasemes today *and* in former times—suggesting ongoing productivity (we did not compare the individual phrasemes in which they occur, but only the number of phrasemes). Focusing on changes, we were also interested in nouns showing higher productivity in older collections than in newer ones and vice versa—indicating loss of phrasemes and emergence of new ones.

We set a threshold of 2%, meaning that a noun was considered *frequent*, if it belongs to the top 2% in the frequency list of all nouns of a collection. A noun was considered *infrequent*, if it was found in one or two phrasemes of a collection only[9]. Based on this we assembled (a) historical frequent nouns, (b) contemporary frequent noun, (c) historical infrequent nouns, and (d) contemporary infrequent nouns.
Based on that we could identify:

- Nouns frequently used in HP *and* in CP, like *Hand* ("hand"), *Teufel* ("devil"), or *Kopf* ("head")
- Nouns frequently used in HP, but not in CP, especially animals like *Affe* ("monkey"), *Laus* ("louse"), or *Kuh* ("cow"), as well as *Narr* ("fool"), *Schnee* ("snow"), or *Feder* ("feather")
- Nouns frequently used in CP, but not in HP, *Nerv* ("nerve"), *Fall* ("case"), or *Punkt* ("point")
- Nouns infrequently used in HP *and* in CP, like *Affenschande* ("apish shame"), *Friedenspfeife* ("calumet"), or *Gnadenbrot* ("charity")
- Nouns infrequently used in CP, but more frequently in HP, like *Krebs* ("crab"), *Käse* ("cheese"), or *Weib* ("woman")
- Nouns infrequently used in CP, but not used at all in HP, like *Fleischwolf* ("meat grinder"), *Brechstange* ("crow bar"), *Sprungbrett* ("diving board"), or *Abstellgleis* ("holding track")

Keeping interested laypersons in mind, we did not look for infrequently used nouns in HP with no evidence in CP or more frequently used in CP.

Having identified diachronically "interesting" nouns, we explored the phrasemes in which these nouns occur. For each resource we independently identified and allocated variants and synonyms of phrasemes by assigning specific *phraseme types*. A phraseme type represents a specific semantic concept. Instantiations include all lexical and

that results on the nouns occuring in our extracted phrasemes, are quite acceptable, there is more spelling variation in verbs.

[9] We decided to use two phrasemes as the lower bound instead of one, as manual inspection had shown that for nouns with two associated phrasemes, the phrasemes typically tend to be variants of each other.

structural variants. For example, in Adelung we find the phraseme *jemanden Staub in die Augen streuen* ("to throw dust into someone's eyes", 'to pull the wool over someone's eyes'), whereas in Duden11 we have *jmdm. Sand in die Augen streuen* ("to throw sand into someone's eyes"), both of them belong to the same phraseme type and express the same meaning.

Based on phraseme types, we can then flip the matrix and see for each phraseme type which nouns in which phrasemes are associated to this particular phraseme type in which collection. We thus get an impression of the variants already reported in various collections and can thus decide which phraseme types to investigate further. By annotating other resources like Borchardt (1888) or WdDU with phraseme types as well, we create a rich resource that allows us to get a first diachronic impression. Note that up to this moment, we have made use of already existing lexical information, which we have rearranged and recombined. We still lack empirical evidence but rely on statements of other phraseologists only. In the next section we look at the empirical part, which is work in progress.

5.2 Searching corpora

With respect to corpora, we have to face a different notion of *frequent* and *infrequent*: an infrequent noun like *Friedenspfeife* with only one associated phraseme type (*die Friedenspfeife mit jemandem rauchen* "to smoke the calumet together", 'to reconcile') might be found quite often in texts and thus be relatively frequent. Additionally, we can calculate the frequency of the lexical units of a phraseme as occurring in the text regardless of whether it is used in a phraseme or in its literal meaning in other contexts. However, one fundamental problem searching corpora for phrasemes is their generally low frequency compared to other multi-word units. (Colson, 2007) Due to variation of the phrasemes and to the decreasing size of corpora, phrasemes get more and more difficult to find the further back in time we search (see also Claridge, 2008).

Our first intuition—based on our definition of variants and synonyms of phrasemes and taking into account the state of the art—was a lemma-based search for phrasemes and their variants, allowing for syntactical, lexical, and morphological variation. However, as most of the relevant corpora do not provide lemmata and we will not be able to lemmatize them automatically, we have to come up with other strategies, taking into account spelling variants, too.

Using vector-based approaches from the field of information retrieval (IR) (Salton et al., 1975) like computing co-occurence vectors for the phrasemes in question, we can use the already identified instantiations of a phraseme type to find them in the corpora by matching the query-vector to the corpus allowing for variation—the phraseologists will then have to decide if a match is indeed idiomatic. However, we also have to take into account that the texts in the corpora are written in several variants of German.

Spelling variation and different inflectional paradigms[10] might influence recall and precision of vector-based approaches.

For queries considering spelling variation, we will make use of data provided by the project "Freiburger Anthologie", a collection of the 1000 most important German poems.[11] We have enriched this data with observations in the texts of DB125. We found further spelling variants, out-dated vocabulary, and different inflectional paradigms especially for verbs. Note that using vectors we can look for surface-similarity only, not for semantic similarity.

For finding variants including semantically similarity we will use GERMANET (Hamp and Feldweg, 1997) to identify synonyms, hyperonyms, and hyponyms for the lexical units used in a phraseme. For example from *der Apfel fällt nicht weit vom Stamm* ("the apple does not fall far from the stem", 'like father like son') we can create the forms *der Apfel fällt nicht weit vom Baum* ("the apple does not fall far from the tree"), *die Birne fällt nicht weit vom Stamm* ("the pear does not fall far from the stem"), *die Birne fällt nicht weit vom Baum* ("the pear does not fall far from the tree"). Considering spelling variation and allowing for changes in word order we are then able to find *die birn nit wey vom baum falt* (Rechtsquellenstiftung des Schweizerischen Juristenverbandes, 2009, 121)

Automatically creating search queries including semantic variation for single lexical units will allow us to automatically create variants of the phrasemes we are investigating. Using vector-based IR algorithms we will then look for evidence in the corpora. The results will contain context to enable phraseologists to decide whether a particular match is a phraseme or a non-idiomatic co-occurrence only. If a match is not a phraseme, but the words are used literally, the match will not be rejected but marked as non-idiomatic. Idiomatic evidence will be annotated with all information needed to serve as source for a general comment on diachronic change of the phraseme under investigation as well as information to be directly displayed to the user of the resulting dictionary.

Based on the results of all these procedures described above, we will be able to enrich the lexicographic information for a particular phraseme type. We will also be able to provide statistical information showing some trends concerning increase, decrease, or stability of use of a phraseme type (or a specific variant) over time.

6 Conclusion

We presented our semi-automatic approach for investigating phrasemes in German from a diachronic perspective. Due to the diachronic aspect, several issues arise which can be solved by using manual effort in combination with automatic processing steps. Searching for variants of phrasemes (or multi-word units in general) in historical texts

[10] A word might have belonged to a different inflectional paradigm a few hundred years ago than it does today, an example would be the verb *backen* ("to bake") with weak inflection today (*backte* and strong inflection formerly *buk*. However, the strong inflection is still used in Swiss Standard German, but not in Germany or Austria.

[11] `http://freiburger-anthologie.ub.uni-freiburg.de/fa/fa.pl?cmd=gedichte\&sub=analog\&add=`

emphasizes the need to solve problems of normalization and lemmatization—higher-level applications as the OLdPhras project rely on those annotations to allow the use of state-of-the-art NLP, information retrieval, or text mining tools. In particular, if lemmatization has already been performed, freely available corpora should be distributed including this annotation.

However, including the human in the loop at various steps at the process, we developed a semi-automatic approach that is transferable to other situations—other languages or texts from other periods. We will thus be able to provide some information on the evolution of form, meaning, and use of German phrasemes that goes beyond example-based explorations. We will also be able to annotate respective information in the corpora, which might later be used by other researchers investigating other questions.

7 Acknowledgments

We thank our colleagues from the OLdPhras project for collaboration on concepts and for the thorough manual annotation of the various extracts described. We also thank the anonymous reviewers for helpful comments on an earlier version of this paper.

References

Adelung, J. C. (1793–1801). *Grammatisch-kritisches Wörterbuch der Hochdeutschen Mundart*. Breitkopf & Sohn, Leipzig. (= Adelung).

Bennett, P., Durrell, M., Scheible, S., and Whitt, R. J. (2010). Annotating a historical corpus of German: A case study. In *Proceedings of the LREC 2010 Workshop on Language Resource and Language Technology: Standards - state of the art, emerging needs, and future developments*, pages 64–68, Paris. ELRA.

Borchardt, W. (1888). *Die Sprichwörtlichen Redensarten im deutschen Volksmund nach Sinn und Ursprung erläutert*. Brockhaus, Leipzig.

Burger, H. (2010). *Phraseologie*. Erich Schmidt, Berlin.

Burger, H. and Linke, A. (1998). Historische Phraseologie. In Besch, W., Betten, A., Reichmann, O., and Sonderegger, S., editors, *Sprachgeschichte: Ein Handbuch zur Geschichte der deutschen Sprache und ihrer Erforschung*, pages 743–755. Walter de Gruyter, Berlin/New York.

Campe, J. H. (1807–1812). *Wörterbuch der deutschen Sprache*. Schulbuchverlag, Braunschweig.

Claridge, C. (2008). Historical corpora. In Lüdeling, A. and Kytö, M., editors, *Corpus Linguistics*, pages 242–259. Walter de Gruyter, Berlin/New York.

Colson, J.-P. (2007). The World Wide Web as a corpus for set phrases. In Burger, H., Dobrovol'skij, D., Kühn, P., and Norrick, N. R., editors, *Phraseology*, Handbooks of Linguistics and Communication Science, pages 1071–1077. Walter de Gruyter, Berlin/New York.

Dudenredaktion (2008). *Redewendungen: Wörterbuch der deutschen Idiomatik*. Dudenverlag, Mannheim. (= Duden11).

Evert, S. (2005). *The statistics of word cooccurrences: Word pairs and collocations*. PhD thesis, Universität Stuttgart, Institut für Maschinelle Sprachverarbeitung.

Fellbaum, C., editor (2007). *Idioms and Collocations: Corpus-based Linguistic and Lexicographic Studies*. Research in Corpus And Discourse. Continuum, London/New York.

Fellbaum, C. and Geyken, A. (2005). Transforming a corpus into a lexical resource the Berlin Idiom Project. *Revue française de linguistique appliquée*, X(2):49–62.

Fischer, I. and Keil, M. (1996). Parsing decomposable idioms. In *Proceedings of the 16th conference on Computational linguistics - Volume 1*, COLING '96, pages 388–393, Stroudsburg, PA, USA. Association for Computational Linguistics.

Friedrich, W. (1976). *Moderne deutsche Idiomatik. Systematisches Wörterbuch mit Definitionen und Beispielen*. Huber, München.

Fritzinger, F., Kisselew, M., Heid, U., Madsack, A., and Schmid, H. (2009). Werkzeuge zur Extraktion von signifikanten Wortpaaren als Webservice. In Hoeppner, W., editor, *GSCL-Symposium Sprachtechnologie und eHumanities*, pages 32–43.

Grimm, J. and Grimm, W. (1852–1971). *Das deutsche Wörterbuch*. Hirzel, Leipzig. (= DWB).

Gschwend, L. (2007). Die Sammlung Schweizerischer Rechtsquellen, herausgegeben von der Rechtsquellenstiftung des Schweizerischen Juristenvereins: Ein Monumentalwerk rechtshistorischer Grundlagenforschung. *Zeitschrift für Schweizerisches Recht*, 126(I):435–457.

Hamp, B. and Feldweg, H. (1997). GermaNet – a lexical-semantic net for German. In *Proceedings of the ACL workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15, Somerset, NJ, USA. Association for Computational Linguistics.

Heid, U. (2007). Computational linguistic aspects of phraseology II. In Burger, H., Dobrovol'skij, D., Kühn, P., and Norrick, N. R., editors, *Phraseology*, Handbooks of Linguistics and Communication Science, pages 1036–1044. Walter de Gruyter, Berlin/New York.

Jurish, B. (2010). More than words: Using token context to improve canonicalization of historical German. *Journal for Language Technology and Computational Linguistics*, 25(1):23–39.

Juska-Bacher, B. and Mahlow, C. (2012). Phraseological change – a book with seven seals? tracing back diachronic development of German proverbs and idioms. In Durell, M., Scheible, S., and Whitt, R. J., editors, *TBA*, volume 3 of *Corpus linguistics and Interdisciplinary perspectives on language*. Gunter Narr, Tübingen, Germany.

Keil, M. (1997). *Wort für Wort – Repräsentation und Verarbeitung verbaler Phraseologismen (Phraseo-Lex)*, volume 35. Max Niemeyer Verlag, Tübingen.

Kluge, F. (2002). *Etymologisches Wörterbuch der deutschen Sprache*. Walter de Gruyter, Berlin, New York, 24th revised and expanded ed edition.

Kühn, P. (2003). Phraseme im Lexikographie-Check: Erfassung und Beschreibung von Phrasemen im einsprachigen Lernerwörterbuch. *Lexicographica*, 19:97–118.

Kuiper, K., McCann, H., Quinn, H., Aitchison, T., and van der Veer, K. (2003). *SAID: A syntactically annotated idiom database*. Linguistic Data Consortium, Philadelphia.

Küpper, H. (1997). *PONS Wörterbuch der Deutschen Umgangssprache*. Klett Verlag, Stuttgart. (= WdDU).

Lorenz, O. (1996). Automatische Wortformerkennung für das Deutsche im Rahmen von MALAGA. Master's thesis, Friedrich-Alexander-Universität Erlangen-Nürnberg.

Mahlow, C. and Piotrowski, M. (2009). A target-driven evaluation of morphological components for German. In Clematide, S., Klenner, M., and Volk, M., editors, *Searching Answers – Festschrift in Honour of Michael Hess on the Occasion of his 60th Birthday*, pages 85–99. MV-Verlag, Münster.

Mellado Blanco, C. (2009). Einführung. Idiomatische Wörterbücher und Metafraseografie: zwei Realitäten, eine Herausforderung. In Mellado Blanco, C., editor, *Theorie und Praxis der idiomatischen Wörterbücher*, pages 1–20. Max Niemeyer, Tübingen.

Piotrowski, M. (2010). From Law Sources to Language Resources. In Sporleder, C. and Zervanou, K., editors, *Proceedings of the ECAI 2010 workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2010)*, pages 67–71.

Rechtsquellenstiftung des Schweizerischen Juristenverbandes, editor (2009). *Appenzeller Landbücher*, volume SSRQ AR/AI 1 of *Sammlung Schweizerischer Rechtsquellen*. Schwabe, Basel, Switzerland.

Röhrich, L. (2002). *Das große Lexikon der sprichwörtlichen Redensarten*. WBG, Darmstadt.

Rothkegel, A. (2007). Computerlinguistische Aspekte der Phraseme I. In Burger, H., Dobrovol'skij, D., Kühn, P., and Norrick, N. R., editors, *Phraseology*, Handbooks of Linguistics and Communication Science, pages 1027–1035. Walter de Gruyter, Berlin/New York.

Salton, G., Wong, A., and Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Sanders, D. (1859–1865). *Wörterbuch der deutschen Sprache*. Wigand, Leipzig.

Seretan, V. and Wehrli, E. (2010). Tools for syntactic concordancing. In *Proceedings of the International Multiconference on Computer Science and Information Technology*, pages 493–500. IEEE.

Stantcheva, D. (2003). *Phraseologismen in deutschen Wörterbüchern: Ein Beitrag zur Geschichte der lexikographischen Behandlung von Phraseologismen im allgemeinen einsprachigen Wörterbuch von Adelung bis zur Gegenwart*. Dr. Kovač, Hamburg, Germany.

Volk, M., Bubenhofer, N., Althaus, A., Bangerter, M., Furrer, L., and Ruef, B. (2010). Challenges in building a multilingual Alpine heritage corpus. In *Seventh International Conference on Language Resources and Evaluation (LREC)*, pages 1653–1659, Paris. European Language Resources Association (ELRA).

Wander, K. F. W. (1867–1880). *Deutsches Sprichwörter-Lexikon*. Brockhaus, Leipzig. (= DSL).

Wittern, C., Ciula, A., and Tuohy, C. (2009). The making of TEI P5. *Literary and Linguistic Computing*, 24(3):281–296.

Massimo Manca, Linda Spinazzè, Paolo Mastandrea, Luigi Tessarolo, Federico Boschetti

Musique Deoque: Text Retrieval on Critical Editions

This paper aims at illustrating the main features of the *Musisque Deoque Project*, which provides a fully freely searchable archive of Latin poetry equipped with critical apparatus. The first part explains how variants are mapped on the reference edition and the second part illustrates the web interface to retrieve sequences of words taking into account possible variants.

1 Introduction

The *Musisque Deoque Project* (MQDQ) aims at creating a digital archive of Latin poetry, from its origins to the late Italian Renaissance, equipped with critical apparatus and various exegetical and linguistic information. This project is focused on the study of synchronical and diachronical intertextuality as illustrated, e.g., in Cicu (2005). For this reason, we give strong attention to formal and material aspects of the text that actually played a relevant role in the poetical tradition. The fixed text of printed critical editions, aimed at the reconstruction as close as possible to the lost originals, provides just a snapshot of the tradition, which is intrisically dynamic, and gives to the modern reader a distorted image of what an ancient text was in fact.

Fully searchable digital collections currently available are based on traditional critical editions, which are, as we just said, authoritarian texts; this authoritarianism is emphasized by the conversion from printed text to database, because usually the critical apparatus is cut away and there is no way for the reader to check a variant different from the one the editor put in the main text, often *dubitanter*, simply because he *had* to choose a variant. Limiting lexical searches to editor's choices drives unavoidably both to false positives and false negatives, which need to be verified back on printed critical editions. False positives are due to possibly wrong emendations made by modern and contemporary scholars, provided by the text retrieval systems among the genuine occurrences, whereas false negatives are the likely variants excluded by editors biased by prejudices against specific linguistic and stylistic phenomena (such as the short-term repetiton, systematically emended by philologists of the last centuries).

The purpose of Musisque Deoque is to overcome these limitations, retrieving not only the word keys quoted in the reference edition, but also the variants lying in the critical apparatus. In this way, further knowledge on the accomplished itinerary – from ancient operas during the subsequent ages until the Humanism and the Renaissance – can emerge.

2 Background

Musisque Deoque is the result of the continuous evolution of projects focused on the digitization of the Latin poetry: *Almae Latinitatis Bibliotheca* (Classical Latin Poetry), *Poetria Nova* (Medieval Latin Poetry) and *Poeti d'Italia in Lingua Latina* (Humanist and Renaissance Italian Poetry in Latin). Along the decades, additional information has been encoded to the text-only documents related to metrical genres, to biographical data of the authors and other information and, consequently, features have been added to the search engines available on CD or online (in particular, *Poeti d'Italia*: http://www.poetiditalia.it and *Musisque Deoque*: http://www.mqdq.it).

Important projects that deal with digital variants have been developed in the last decades: see, in particular, Calabretto and Bozzi (1998) and Calabretto et al. (2005). These projects are focused on the collation of manuscripts and are aimed to provide tools that help the philologist to check variants on the images of the manuscripts or to produce an automated collation of digital diplomatic editions. On digital philology and Medieval texts, see Stella and Ciula (2007).

The Perseus Project stressed out the importance of a cyberinfrastructure for the classical studies able to deal also with variants. See, for instance, Crane (2009) and Crane (2010).

Peter Robinson, the promoter of *Interedition* and *Virtual Manuscript Room*, considers the process of editing digital editions as a collaborative enterprise: see Robinson (2010) and Babeu (2011); see also Price (2009) and McGann et al. about digital scholarship. In this perspective, the main purpose of *Interedition* is offering a sort of public, social and sharing context in order to improve, compare and discuss first of all the tools for digital scholarly edition publishing. Very similar ideas about the future development of digital scholarly editions are asserted by Gabler (2010).

MQDQ does not aim to the *constitutio textus* nor offers new protocols for publishing digital editions; its goal is rather to offer a tool to study the literary influences among the tradition. The ideal end-user of MQDQ is a scholar interested in analyzing the *Fortleben* and the mutual relationship of texts at a more deeper level than the one allowed by the common authoritative databases.

Even if MQDQ takes into account the theoretical models and the practices to represent variants, as expressed in recent contributions, such as Boschetti (2007), McGann (2010), Gabler (2010), Marotti (2010), May (2010), Mandell (2010), the main goal of MQDQ project is to achieve a very extensive database, which includes almost all the Latin poetry with a wide range of variants.

MQDQ is a work in progress and its features and improvements have been illustrated in several conferences and articles, such as Zurli and Mastandrea (2009), Manca (2009) and Mastandrea (2011).

3 Encoding Text and Critical Apparatus

Musisque Deoque is based on dynamic repertories of texts and critical apparatus. On one hand, the text of a classical work is a faithful transcription of the text established by the editor of the most authoritative printed critical edition currently available and only in few cases it is digitized from the text established by scholars on articles, Ph.D. theses, etc. Pages are scanned, OCR is performed, and skilled operators select only text boxes excluding the critical apparatus. Manual corrections made by the operators are reviewed by the project managers and their collaborators.

On the other hand, many critical editions of the same work are examined by skilled operators. They prepare the digital *conspectus codicum* and they insert the variants that they consider the most relevant for the study of the textual tradition.

The concept of "significant variant" isn't so subjective as it sounds. Manca (2009) defines "significant variant" a "*lectio* we can credit to the author himself, or to an editor, but more often introduced by readers or copists still in the ancient phase of the tradition, and which may bring to new perspectives in intertextual researches". A variant that trace the path of the textual tradition must be considered significant, even when it is an error from a metrical, syntactic, pragmatic or encyclopedic point of view. For example, a variant such as *Gallia omnis est divisa in partes quattuor* would be surely rejected in any traditional critical apparatus; but if this mistake would have been elaborated by the literary tradition, one should accept it in a *corpus-oriented* apparatus. For a more realistic case, see the success, in the scriptural tradition, of the 'wrong' expression *Nabuzardan princeps coquorum* over the 'correct' *Nabuzardan princeps militiae* in Manca (1999). In the MQDQ enviroment it is significant also a variant *deterior* or *facilior*, completely useless for the *constitutio textus*, if this different reading somehow spreaded itself in literature. Usually corrupted variants, cases of *scriptio continua* or wrong division are not significant into our archive.

In order to enrich very quickly the existent database of latin texts with more variants for every single works, the first technical effort was to build a user-friendly tool that permits to scholars unskilled in IT but very competent philologists to become MQDQ-operators. The MQDQ operator works with a cross-platform sofware written in Java called MQDQ2. The philologist that creates new digital editions with MQDQ may decide to download and modify the text present in the pre-existent database or to replace it with another plain text (i.e. without tags). The user have to initialize the text for adding apparatus information through a sequence of dialogue boxes (Fig. 1): in this phase the operator writes the header of xml file for the apparatus and decides how to create the *conspectus codicum*.

The table of manuscripts and bibliography can be encapsulated in the same XML file of the apparatus, or can be saved as an indipendent XML file: it is up to the operator in the first phase to choose the preferred method. The wizard that guides the operator that starts working with a new text offers the author a choice to build a new *conspectus codicum* or to share an existent one, usually taken from a different section of the same

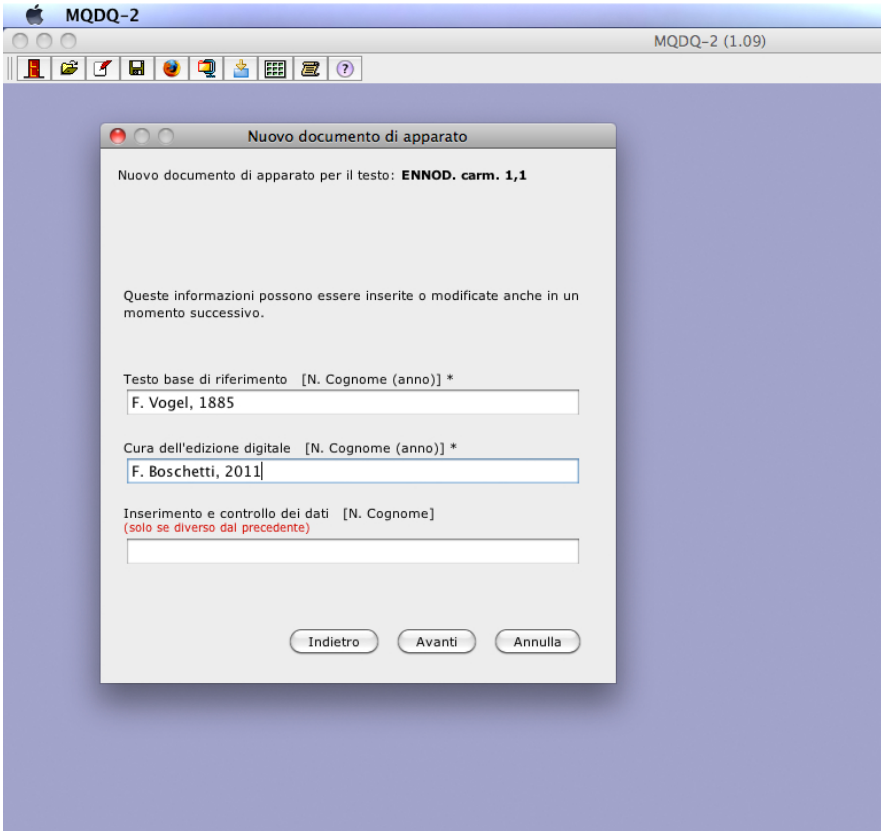


Figure 1: Initializing the interface 1/1

author. Very often, indeed, it is useful to choose the same table of *codices* among several sections of the same work, or different works of same author(s) (Fig. 2).

After the preliminary operations, the user accesses the main page of the application, where he can build a *conspectus codicum et uirorum doctissimorum* (Fig. 3) and add the variants or other kind of paratextual notes (Fig. 4). The system is enough flexible to allow the operator to cope with the almost endless problems of information representation in a text.

So, variants and conjectures registered in different critical editions are inserted in the new digital critical apparatus, and each textual variant is mapped on the reference edition.

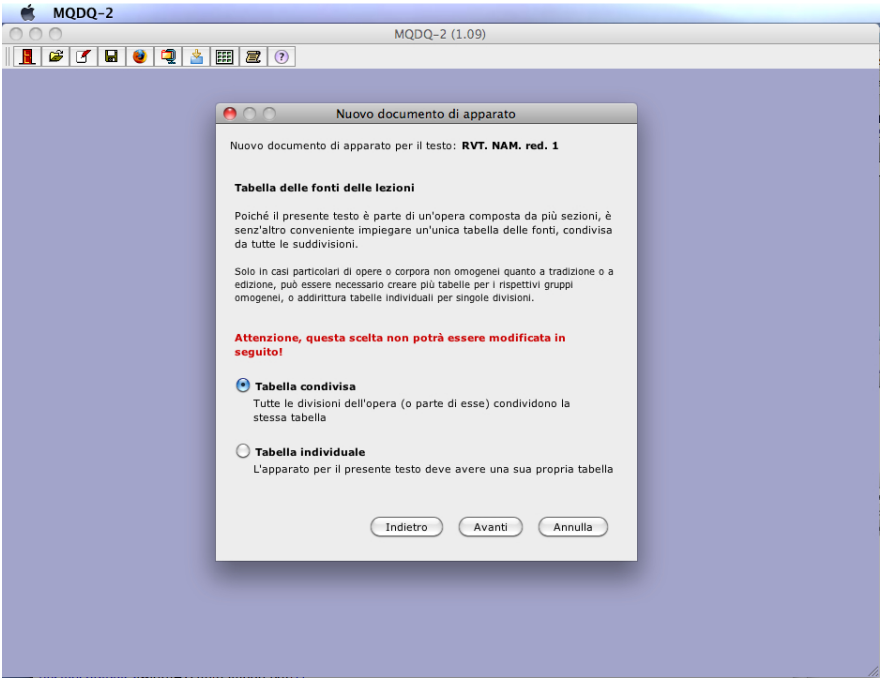


Figure 2: Initializing the interface 1/2

The structure of the back-end is transparent to the operators, which are not supposed to be skilled in XML annotation, but the hidden structure is worthy of mention.

In order to decouple text and apparatus, the text is fully segmented at the word level. Each word or punctuation mark receives a unique identifier, which is used to map the variant on the correct position in the context of the verse, as illustrated below:

```
<line id="l38" name="36" type="verse">
        <word id="w239">patrem</word>
        <word id="w240">probaui</word>
<punctuation id="w241" space="post">,</punctuation>
        <word id="w242">gloriae</word>
<word id="w243">feci</word>
<word id="w244">locum</word>
<punctuation id="w245" space="post">.</punctuation>
</line>
```

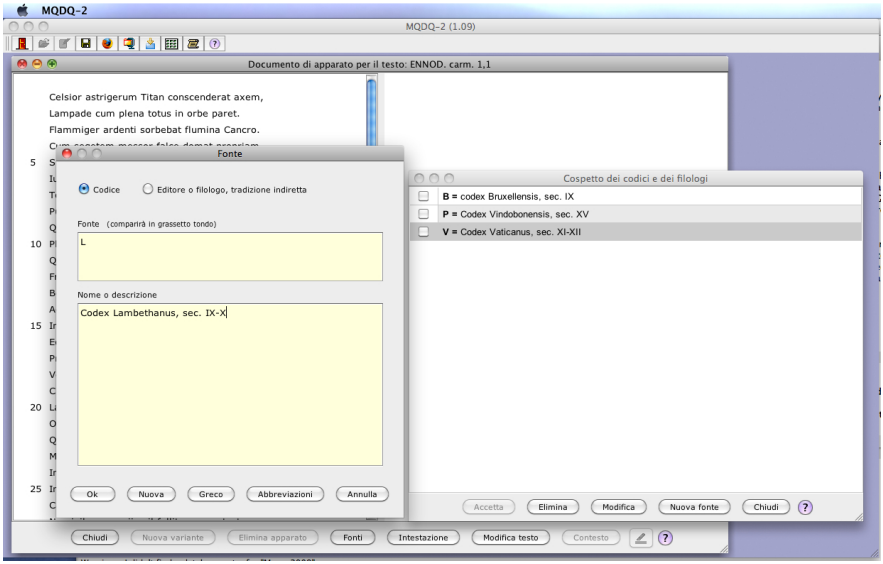


Figure 3: Creating conspectus codicum

```
<line id="139" name="37" type="verse">
<word id="w246">qua</word>
<word id="w247">Sol</word>
<word id="w248">reducens</word>
<word id="w249">quaque</word>
<word id="w250">deponens</word>
<word id="w251">diem</word>
</line>
```

The *conspectus codicum et virorum doctissimorum* is built in a separate file (or encapsulated in the same app.xml file, see above), encoding the names of the editions from which the information is extracted, *sigla*, description of manuscripts, and the name possibly with reached bibliographical information about scholars that proposed conjectures. As seen above, this *conspectus* collects information from many critical editions of the same work. The editor of the printed edition and the editor of the electronic edition are mentioned within the `head` tag. Each source has an identifier, which is unique for the metadata related to a specific work (e.g. *s1* for the code **R** that contains Seneca's tragedies). But this identifier can be equivalent to the identifiers related to other works (e.g. when a miscellaneous manuscript is mentioned in different ways by the scholars).

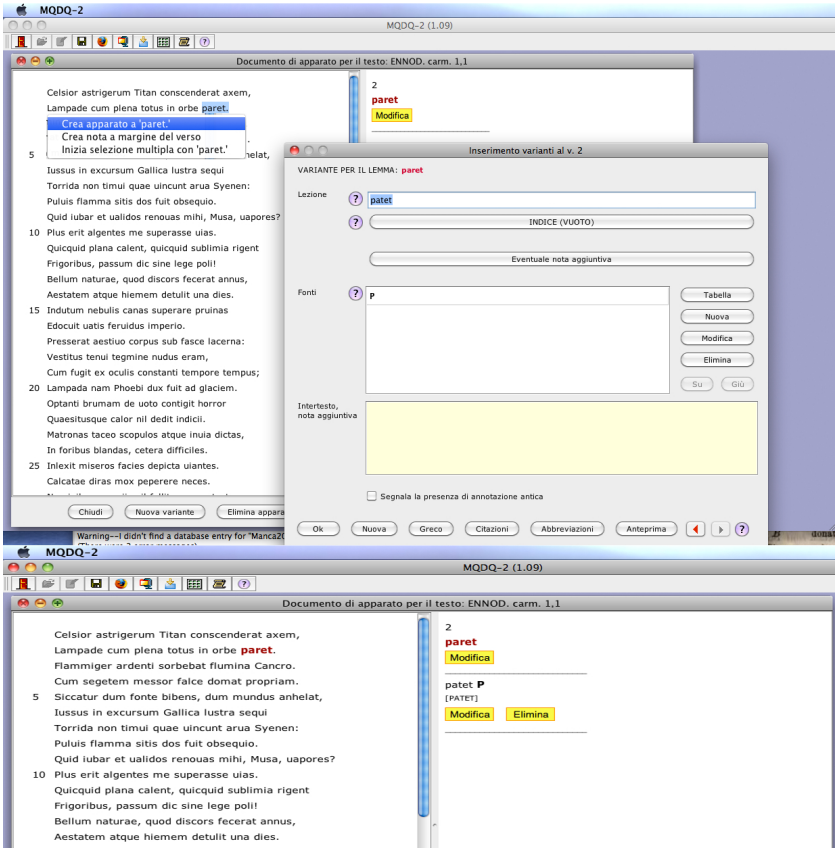


Figure 4: Creating variant readings

A conversion table allows the suitable correspondence. Each source has also a type, which can be `cod` (for a manuscript) or `auth` (for a scholar that suggested a conjecture), as shown below:

```
[...]
    <head>
        <editor>O. Zwierlein (1986)</editor>
        <e_editor>G. C. Musa (2008)</e_editor>
    </head>
    <links>
        <text>xml-app/SEN-hefu-001-txt.xml</text>
    </links>
```

```
    <conspectus>
        <source id="s1" type="cod">
            <name>R</name>
            <explication>Ambrosianus G 82 sup., saec. V in., 5 foll. rescripta</explication>
        </source>
        <source id="s2" type="cod">
            <name>E</name>
            <explication>Laurentianus Plut. 37. 13
                (&amp;quot;Etruscus&amp;quot;), saec. XI ex., foll. 165</explication>
        </source>
        [...]
        <source id="s67" type="auth">
            <name>Commelinus</name>
            <explication>H. Commelinus apud Scriuerium</explication>
        </source>
        [...]
```

After the preamble that contains the *conspectus*, chunks affected by multiple readings are registered. Each chunk is uniquely identified and it contains a reference to the textual positions where insertions, deletions, substitutions or translations are required by alternative readings.

The reading accepted by the editor of the reference edition is marked as `pos`, whereas the other variants are marked as `var`. The text of the reading is inserted and words are indexed with the positions in the verse. In case of addition of text, positions can be identified by decimal numbers. In case of deletion of text, the `operation` tag is used. `note` is used to insert unstructered information from the printed critical apparatus, such as evaluations of scholars (e.g. *dubitanter...*); the encoding of a chunk of information is shown below:

```
[...]
<chunk id="i12" nameVerse="37" idRef="w246w247w248w249w250">
        <reading type="pos">
                <reading>qua Sol reducens quaque deponens</reading>
                <source idSources="s2">
                    <operation></operation>
                    <note></note>
                </source>
        </reading>
        <reading type="var">
                <reading>aperitque thetis qua ferens titan</reading>
                <index idRef="w246">APERITQVE</index>
                <index idRef="w247">THETIS</index>
                <index idRef="w248">QVA</index>
                <index idRef="w249">FERENS</index>
                <index idRef="w250">TITAN</index>
                <source idSources="s16">
                    <operation></operation>
                    <note></note>
                </source>
        </reading>
</chunk>
[...]
```

As said above, the resulting XML document is not visible to the operators, mostly graduate or Ph.D. students of Latin literature, which insert data via a front-end interface

that allows the selection of text on the reference edition and the completion of related forms with the information about variants. This back-end system produces a coherent XML code among all the operators and, as no XML training is necessary, a scholar with no previous knowledge of tagging may be operative with only a short two-hours briefing.

4 Querying

The simplest function of the web interface is the retrieval of word sequences. MQDQ inherits the metrical metadata encoded in the previous projects directed by P. Mastandrea and L. Tessarolo, such as Poeti d'Italia in Lingua Latina. It is possible to find words in special positions of the verse (in particular the beginning and the end), to filter specific metres (e.g. dactylic metres) and also to search words inside the extra-text, the sender of a letter, the speaker of dialogues etc. (see Fig. 5).

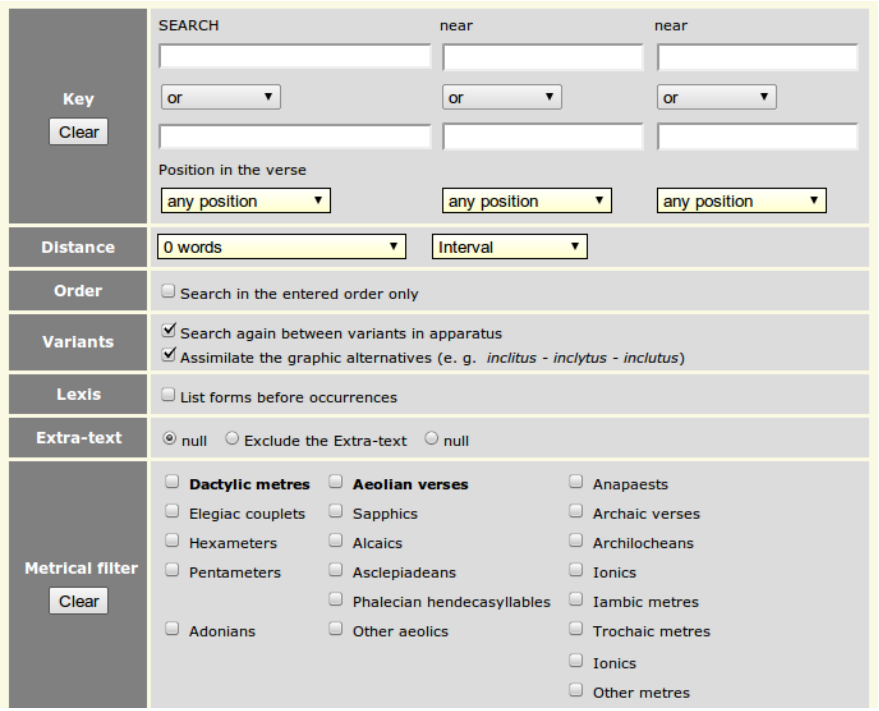


Figure 5: Querying Interface

The query mask allows to search word sequences (and graphic alternatives, such as words written with `-dc-` or with `-cc-`) not only in the reference editions, but also in the collection of variants, collocated in the correct context of the verse. That means that a sequence constituted by a word of the reference edition followed by a word registered in the critical apparatus can be recognized as adjacent and retrieved.

Figure 6: Variants

Searching for *amore leget*, it is possibile to retrieve also the variants *amore legat* or *amor releget*. Searching for *Captus amor*, it is possible to find the occurrence even if one word was originally recorded in the reference edition and the other was originally encoded only in the critical apparatus (see Fig. 6).

5 Conclusion

In conclusion, Musisque Deoque provides tools both to build digital critical editions and to query a large database of variants, which are mapped on reference editions. MQDQ is focused on the study of the intertextuality, and for this reason is based neither on digital diplomatic editions of manuscripts nor on the mere text of traditional critical editions, but on a selection of relevant variants recorded in printed critical editions. With increasing digital material such as manuscripts and ancient edition on the Web, the team of MQDQ are now ready to work in the direction of interoperability, expanding the traditional intertextuality among ancient texts to the new intertextuality that the power of the Internet nowadays offers, according to the developments illustrated in Spinazzè (2011).

References

Babeu, A. (2011). *Rome Wasn't Digitized in a Day: Building a Cyberinfrastructure for Digital Classicists*. CLIR Reports.

Boschetti, F. (2007). Methods to extend greek and latin corpora with variants and conjectures: mapping critical apparatuses onto reference text. In *Proceedings of the Corpus Linguistics Conference*.

Calabretto, S. and Bozzi, A. (1998). The philological workstation bambi (better access to manuscripts and browsing of images). *J. Digit. Inf.*, 1(3).

Calabretto, S., Bozzi, A., Corradini, M. S., and Tellez, B. (2005). The eumme project: towards a new philological workstation. In *ELPUB*.

Cicu, L. (2005). *Le api il miele la poesia. Dialettica intertestuale e sistema letterario greco-latino*. Roma.

Crane, G. (2009). Cyberinfracstructure for classical philology. *DHQ*, 3(1).

Crane, G. (2010). Give us editors! re-inventing the edition and re-thinking the humanities. In *Online Humanities scolarship: the Shape of things to Come*, <http://cnx.org/content/m34316/latest>. Mellon Foundation, Jerome McGann, ed.

Gabler, H. W. (2010). Theorizing the digital scholarly edition. *Literature Compass*, 7(2):43–56.

Manca, M. (1999). Nabuzardan princeps coquorum. una lezione vulgata oltre la vulgata. *Quaderni del Dipartimento di Filologia, Linguistica e Tradizione Classica (Università degli Studi di Torino*, (13):491–498.

Manca, M. (2009). Database and corpora of ancient texts towards the "second dimension": theory and practice of musisque deoque project. In P. Anreiter, M. K., editor, *Computational Linguistics and Latin Philology. 15th Colloquium on Latin Linguistics.*, pages 697–702.

Mandell, L. (2010). Special issue: 'scholarly editing in the twenty-first century'– a conclusion. *Literature Compass*, 7(2):120–133.

Marotti, A. F. (2010). Editing manuscripts in print and digital forms. *Literature Compass*, 7(2):89–94.

Mastandrea, P., editor (2011). *Nuovi archivi e mezzi d'analisi per i testi poetici. I lavori del progetto Musisque Deoque, Venezia 21-23 giugno 2010*. Hakkert.

May, S. W. (2010). All of the above: The importance of multiple editions of renaissance manuscripts. *Literature Compass*, 7(2):95–101.

McGann, J. (2010). Electronic archives and critical editing. *Literature Compass*, 7(2):37–42.

McGann, J., Gabler, H. W., Wolfson, S. J., Pratt, L., Curran, S., Marotti, A. F., May, S. W., Ezell, M. J. M., O'Donnell, D. P., and Mandell, L. *Literature Compass*, (2):134–144.

Price, K. (2009). Edition, project, database, archive, thematic research collection: What's in a name? *Digital Humanities Quaterly*, 3(3).

Robinson, P. (2010). Editing without walls. *Literature Compass*, 7(2):57–61.

Spinazzè, L. (2011). Risalire alle fonti: dall'edizione mqdq ai testimoni manoscritti. In Mastandrea, P., editor, *Nuovi archivi e mezzi d'analisi per i testi poetici*, pages 59–71. Hakkert.

Stella, F. and Ciula, A., editors (2007). *Digital philology and medieval texts. Atti del seminario, Arezzo 2006*. Pacini.

Zurli, L. and Mastandrea, P., editors (2009). *Poesia latina, nuova E-filologia. Opportunità per l'editore e per l'interprete. Atti del convegno Internazionale. Perugia 13-15 settembre 2007*. Hedder Editrice e Libreria.

Eiríkur Rögnvaldsson, Anton Karl Ingason, Einar Freyr Sigurðsson, Joel Wallenberg

Creating a dual-purpose treebank

We describe the background for and building of IcePaHC, a one million word parsed historical corpus of Icelandic which has just been finished. This corpus which is completely free and open contains fragments of 60 texts ranging from the late 12th century to the present. We describe the text selection and text collecting process and discuss the quality of the texts and their conversion to modern Icelandic spelling. We explain why we choose to use a phrase structure Penn style annotation scheme and briefly describe the syntactic annotation process. Furthermore, we advocate the importance of an open source policy as regards language resources.

1 Introduction

The parsed corpus, or treebank, reported on in this paper, Icelandic Parsed Historical Corpus or IcePaHC (WALLENBERG et al. 2011) is the product of three different projects which originally had different aims. The earliest and largest of these projects was a subpart of a large language technology project which had the aim of developing three different basic language resources for Icelandic. The aim of this subproject was to build a treebank of Modern Icelandic for use in language technology and to develop efficient parsing methods and tools for less resourced languages. Since some of the participants had been involved in historical syntax research, we also wanted to include a few texts from older stages of the language. However, the main emphasis was on language technology use – we intended to use the texts to train a statistical parser for Modern Icelandic.

At the same time, two other projects with the aim of developing resources for studying diachronic Icelandic syntax were in preparation. After some discussion, the participants in these three projects decided to join forces and make a combined effort to build a large parsed corpus covering the history of Icelandic syntax from the earliest sources up to the present. This corpus thus serves the dual purpose of being one of the cornerstones of Icelandic language technology and being an invaluable tool in Icelandic diachronic syntax research. The corpus is now finished and has been made available through free download (*http://linguist.is/icelandic_treebank/Download*) – in fact, we released preliminary versions every three months through the whole project period.

We believe the corpus is unusual in many ways. First, it is designed from the beginning to serve both as a language technology tool and a syntactic research tool, and developed by people with research experience in both diachronic syntax and computational linguistics. Most parsed corpora are developed either for language technology use (such as the Penn Treebank, *http://www.cis.upenn.edu/~treebank/*) or for syntactic research (such as the Penn Parsed Corpora of Historical English, *http://www.ling.upenn.edu/hist-corpora/*).

Secondly, the corpus spans almost ten centuries – the oldest texts are written in the final decades of the 12th century and the youngest are from the first decade of the 21st century. As far as we know, no other single parsed corpus comes close to that. Most other languages

have changed so much in the course of the last thousand years that it would be impractical to have text from such a long period in a single treebank.

Third, our corpus contains over one million words and is thus among the largest parsed corpora that have been published for any language. As far as we know, only English and Czech have larger hand-checked treebanks.

Fourth, the corpus is completely free and open without any registration or paperwork, and the same goes for all the software that has been used to build it and the software that was developed within the project. Both the software and the corpus itself are distributed under the LGPL license.

This paper describes the background of the treebank. In the next section, we explain how it is possible and why it is feasible to build a diachronic treebank spanning almost ten centuries in the history of Icelandic. After that, we discuss several aspects of the material in the treebank – the selection of the texts, their quality, and their conversion to modern Icelandic spelling. We then go on to explain why we choose to build a Penn style treebank instead of a dependency treebank, which might perhaps seem a more obvious choice. Following a brief description of the annotation process, we finally present our open source policy and set forth "10 basic types of user freedom" for language resources.

2 The diachronic dimension

Icelandic is a language with a rich literary heritage ranging from the 12th century to the present. The oldest preserved texts are mainly religious ones, such as for instance the Old Icelandic Homily Book (*Íslensk hómilíubók*, Stock. Perg. 4to no. 15), a large manuscript from around 1200, and a few translations from Latin.

In the 13th century, Icelanders started writing narrative texts, many of which are considered great literature and have been much celebrated. The most important of these texts are the Family Sagas (*Íslendingasögur*), stories about people living 300 years earlier, in the age of the settlement; *Heimskringla* (Sagas of the Kings of Norway) by the famous author Snorri Sturluson; *Sturlunga Saga*, a collection of stories about people and events in 13th century Iceland; and sagas of bishops.

The writing of these texts continued into the 14th century, but in the late 14th and 15th centuries, legendary sagas (*fornaldarsögur*) and romances (*riddarasögur*) become dominant, most of them translations from continental or English sources. However, Icelanders continued writing saga-style narratives on a small scale up to the 19th century. After the reformation in 1550, religious texts in the vernacular become more prominent. Some of the most important texts from the 17th and 18th centuries are biographies and travelogues. The first Icelandic novels were written in the first half of the 19th century.

It is a commonly accepted fact that Icelandic morphosyntax has changed much less during the last thousand years than most other European languages. This has often been attributed to the strong literary tradition and the isolation of the country. However, it must be emphasized that some features of the language have in fact changed considerably since Old Icelandic. Thus, the phonological system has undergone dramatic changes, especially the vowel system. The phonetic quality of many of the vowels has changed, and the quantity system has changed such that vowel length is now context-dependent instead of being fixed.

On the other hand, the inflectional system and the morphology has in all relevant respects remained unchanged from the earliest texts up to the present, although a number of nouns have shifted inflectional class, a few strong verbs have become weak, one inflectional class of nouns has been lost, and the dual in personal and possessive pronoun has disappeared. The syntax is also basically the same, although a number of changes have occurred. The changes mainly involve word order, especially within the verb phrase, and the development of new modal constructions (cf. for instance RÖGNVALDSSON and HELGADÓTTIR 2011).

Thus, present-day Icelanders can read many texts from the 13th century without special training, although that doesn't necessarily mean that they can read the texts directly from the manuscripts. There was no accepted spelling standard until the 20th century, and the same sounds, sound combinations and words can be written in many ways. However, since the morphology is the same, it is usually relatively straightforward to convert older spelling to the modern standard and get legible text.

These two features – the stability of the morphology and the changes in the syntax – are the reasons why it is both possible and feasible to build one treebank with texts spanning a period of ten centuries. If the morphological system had changed dramatically, it would have been difficult and pointless to apply the same annotation scheme to old and modern texts. On the other hand, the known syntactic changes and variation do not greatly complicate the annotation scheme, making it feasible to build a tool that enables us to study these changes and variation in a systematic way. The parsed historical corpus is such a tool.

3 Text selection

Selecting texts to parse for the corpus was a challenging task. We wanted to have the corpus both representative of different text genres and comparable through the centuries. This meant that we excluded some genres which have emerged only recently, such as newspaper texts. We decided in the beginning on a goal of parsing one million words – approximately 100,000 from each century of Icelandic literary tradition.

Our original plan was to have samples from five different genres of text for each century – preferably 20,000 words from each text. The genres we had in mind were narrative texts, religious texts, biographies, law, and science. We knew from the beginning that it would be impossible to reach this goal, simply because texts belonging to some of the genres do not exist from all 10 centuries. We started with narrative texts and religious texts, since texts from these two genres were easiest to get hold of.

When we were well into the project, we decided to abandon the original plan and concentrate on these two genres. Narrative texts are the overwhelming majority of preserved medieval texts, and those which have been most studied and are easiest to get. It is also relatively easy to find religious texts from most centuries, but biographies, laws, and scientific texts are much fewer and harder to find in edited editions. Thus, we decided to stick to the original plan of having around 100,000 words from each century, but instead of dividing this evenly among five genres, we aimed at having 80,000 words of narrative texts and 20,000 words of religious prose. This also increases the data set for the two genres, allowing for more reliable studies of style-shifting phenomena.

By and large, this plan could be upheld. However, we didn't manage to find any religious text that could be attributed to the 15th century, and it proved to be difficult to find enough narrative texts from the 16th through 18th centuries. Instead, we included more of religious texts from the 16th century and some biographies from the 18th and 19th centuries. The distribution of the texts across genres and centuries is shown in table 1.

	nar	rel	bio	sci	law	Total
12th	0	40871	0	4439	0	45310
13th	93463	21196	0	0	6183	120842
14th	77370	21315	0	0	0	98685
15th	111560	0	0	0	0	111560
16th	35733	60464	0	0	0	96197
17th	46281	28134	52997	0	0	127412
18th	63322	22963	22099	0	0	108384
19th	100362	20370	0	3268	0	124000
20th	103921	21234	0	0	0	125155
21st	43102					43102
Total	675114	236547	75096	7707	6183	1000647

Table 1: Text types

The corpus contains (samples of) 60 different texts which came from various sources. Approximately 20 texts were taken from text repositories on the Internet, especially the Icelandic Netútgáfan (*http://snerpa.is/net*) but a few came from the Project Gutenberg website (*http://www.gutenberg.org*), the Internet Archive (*http://www.archive.org/*) and the Medieval Nordic Text Archive (*http://www.menota.org/*). Around 10 texts came from the Árni Magnússon Institute text archive (*http://www.lexis.hi.is/corpus/*). We received around 10 texts directly from scholars who have been editing them or publishing companies that had published them. The remaining texts, around 20, were keyed in for us by students working on the project. Four texts from the 20th and 21st centuries are still under copyright, but we contacted the authors who gave us permission to use them.

4 Text quality

The quality of the texts varies a lot. Very few Old Icelandic texts are preserved in the original, and exact dating of the texts is often very difficult. Usually, the preserved manuscripts are assumed to be several decades and even centuries younger than the original text. We know that the scribes did not copy the manuscripts letter for letter – often they just used their own spelling instead of retaining the spelling of the original. This makes it very difficult to use the text to study phonology and morphology (cf. for instance BERNHARÐSSON 1999).

For those who use the text to study syntax and syntactic change, however, this is not a serious drawback, although in exceptional cases case distinctions in endings may be lost due to phonological changes and/or changes in spelling. On the other hand, it is usually assumed that scribes more or less retained the word order and other syntactic features of the manuscript they were copying, although there are a number of exceptions to this.

Most of the medieval texts that we used are taken from editions with a detailed introduction where the editor, usually a trained philologist, speculates about the dating of both the preserved text and the original. We have in most cases chosen to use the assumed dating of the original. If the scribes changed the syntax when copying older manuscripts – which they no doubt did occasionally – some syntactic features in some of the texts are actually younger and may thus lead us to believe that certain syntactic changes in fact occurred earlier than they actually did.

Another option would have been to use the dating of the preserved manuscript. That would have been misleading if it is assumed – as we do – that scribes usually didn't change the syntax when copying older manuscripts. Using the date of the manuscript, then, would lead us to believe that certain syntactic changes in fact occurred later than they actually did.

The third option would have been to use only those texts which can be dated fairly accurately and which exist in the original or in a manuscript close to the original in time. Unfortunately, this would have left us with only a couple of texts. None of the Sagas, for instance, is preserved in the original, and some of them only in manuscripts that are one, two, or even three centuries younger than the assumed writing of the text. Thus, we decided to choose quantity over quality in some cases and use texts which cannot be dated exactly and/or which only exist in manuscripts considerably younger than the original text. However, we always gave preference to the most reliably dated texts for a given time period when we had an option.

This may of course give rise to wrong or misleading results when the treebank is used to trace the origin or development of certain syntactic feature. However, the treebank is accompanied by detailed "info" files which users can consult and make their own decisions on using or disregarding data from certain texts. Of course, the treebank can also be used to check the dating of the texts. If, for instance, we are studying a certain syntactic phenomenon which increases or declines regularly through the centuries, but one text stands out as an exception, this might be an indication that this text has not been correctly dated.

These problems are by no means confined to our project – the developers of all historical treebanks are faced with similar problems. Note, however, that they have nothing to do with the construction and quality of the treebank per se. They only become problems when we want to interpret the results we get from searching the treebank for certain constructions and try to trace the development of a certain syntactic feature through the centuries. We are restricted by the available texts and have to interpret them somehow – we cannot ask for the native judgments of living speakers. This is of course one of the major problems that all diachronic syntacticians have to deal with.

5 Text conversion

We decided to convert all our texts to modern Icelandic spelling. There were two reasons for this. One was that this makes it possible to search for individual words without having to capture all possible spelling variants using fuzzy search, regular expressions and the like. The main reason was, however, that we wanted to use the open-source IceNLP package for preprocessing. This package (available at *http://icenlp.sourceforge.net*) contains a tokenizer, a PoS tagger, a lemmatizer, and a shallow parser (LOFTSSON 2008; LOFTSSON and RÖGNVALDSSON 2007; INGASON et al. 2008). It was written for Modern Icelandic texts and its dictionary assumes that words have Modern Icelandic spelling. If we had given the package input in the original spelling of each text, the result of the preprocessing would have been much poorer.

All major texts from the medieval period have been published, although the editions are not always as good as one would wish. Many texts from the 16th up to the 19th century, however, have never been published. We decided in the beginning that we would only use texts from printed sources – it would have been prohibitively time-consuming and expensive to digitize texts from manuscripts.

Editions of medieval Icelandic texts have one of three formats: 1) Diplomatic editions, where the text is printed exactly as in the manuscripts. 2) Standardized Old Norse spelling, which is a standard developed in the 19th century and is meant to mirror the sound system of 13th century Icelandic. 3) Modern Icelandic spelling. For most of the 20th century, editions of medieval texts intended for the public were usually in the standardized Old Norse spelling. Since the 1970s, however, it has become customary to use modern Icelandic spelling in new editions of medieval texts. Editions mainly aimed at scholars, however, usually try to mirror the spelling of the manuscript as closely as possible. Texts from the 19th century onwards usually only have minor deviations from the modern spelling.

A number of texts were in modern Icelandic spelling and could be used as they were. However, the majority of them were either in standardized Old Norse spelling or diplomatic, and thus had to be changed. For the texts in the standardized Old Norse spelling, the task was rather easy, and a few simple scripts could be used to make most of the changes. The diplomatic editions were much harder. Some scripts and simple search-and-replace could help, but since the spelling in these texts is often highly irregular, we had to go over them word by word and correct them, which was rather tedious and time-consuming.

6 Annotation scheme

One of the main questions which had to be answered before the annotation started was which annotation scheme to use. Most of the treebanks that have been built for the Scandinavian languages use some version of dependency parsing (e.g. KROMANN 2003; BICK 2003; NIVRE, NILSSON and HALL 2006; EYTHÓRSSON and KARLSSON 2011), so in some sense it would have been most natural for us to follow them. However, we had close contacts with the treebank team at the University of Pennsylvania from the early stages of the project, so it was a natural choice for us to use the phrase structure annotation scheme that they have developed for their parsed historical corpora (KROCH, SANTORINI and DELFS 2004; KROCH

and TAYLOR 2000; SANTORINI 2010). Thus, IcePaHC uses the same general type of labeled bracketing as the Penn Treebank (with dash-separated lemmata added) as shown below:

```
(1)( (IP-MAT   (NP-SBJ (PRO-N Hann-hann))
               (VBDI spurði-spyrja)
               (CP-QUE (WADVP-1 (WADV hvernig-hvernig))
                      (C 0)
                      (IP-SUB (ADVP *T*-1)
                             (NP-SBJ (NPR-D Grími-grímur))
                             (VBDS liði-líða))))
       (ID 1888.GRIMUR.NAR-FIC,.301))
```

This proved to be a very fortunate decision. The Penn annotation scheme has already been adapted for Old English (TAYLOR et al. 2003), which is rather similar to Icelandic in many respects, both as regards the syntax and the morphological system. Thus, the scheme could be applied to Icelandic with only slight modifications. Furthermore, the Penn team has written extensive annotation guidelines which were of tremendous help in our work (SANTORINI 2010). We were careful to write our own guidelines and document all deviations from and additions to the Penn guidelines.

The decision to model our annotation on the Penn annotation system also meant that we could use the software that has been written especially to facilitate the annotation (CorpusDraw) and to search the corpus (CorpusSearch; RANDALL 2005). An extra bonus is that it is now very easy to compare Icelandic and older stages of English. We can write search queries for English in CorpusSearch and by and large use the same queries for Icelandic, although minor modifications are sometimes necessary. Furthermore, Penn-style treebanks have been built or are currently being built for a number of other languages, such as French (MARTINEAU et al. 2010), Portuguese (GALVES and BRITTO 2002), Early High German (LIGHT 2010), Classical Greek (BECK 2011), Yiddish (SANTORINI 1997/2008), and more. This development means that cross-linguistic, comparative diachronic studies can be carried out in a controlled and reproducible way with the same search queries across these datasets.

Yet another reason for choosing the Penn annotation system was that it is relatively rich, compared to most dependency-based schemes. Thus, it should – in principle, at least – be possible to convert our treebank to a dependency treebank, although some information will be lost in the conversion. Going the other way, that is, converting a dependency-based treebank to a Penn-style phrase structure treebank, would, on the other hand, be impossible without adding information.

Even though we followed the Penn scheme in most cases, we found it necessary to make some slight modifications, as mentioned above. The most important of these are that we annotate the words for lemma and nominals also for case, neither of which is done in the English historical corpora (excepting case in Old English).

7 The annotation process

After the texts had been converted to modern Icelandic spelling, they were handed over to student assistants who had the task of dividing them into clauses. Some periods do not signal the end of a tree and not all trees end in periods. Sentence boundary detection for English has been shown to classify periods such that 98.5% of sentences boundaries are correctly identified, a considerable improvement over the 90% precision of a baseline classifier which assumes every period to be a boundary (PALMER and HEARST 1994). While this may seem encouraging we have reasons to prefer a manual approach. Failure to identify clause boundaries interacts with determining phrase structure as demonstrated by (2) below.

(2) a. We saw the problem that affected [NP the man and the woman] # and [NP the problem] made Jupiter look small.
b. We saw the problem that affected [NP the man] # and [NP the woman and the problem] made Jupiter look small.

A reviewer points out that it would be useful to quantify this problem but unfortunately we do not currently have reliable estimates. Nevertheless, we have two good reasons for our manual approach to clause boundary detection. First, while rules for classifying periods as boundary marking or not are fairly simple, the rules for inserting clause boundaries (usually between well-formed matrix clauses but sometimes sentence fragments without enough material to reconstruct clausal structure reliably and consistently) are more complicated and require interpretation of gapping structures where the nature and amount of omitted material affects the boundary status of conjunctive elements. Such problems can in principle be addressed using computational methods but the required tools are not currently available for Icelandic and their development was beyond the scope of our project. Second, while clause boundary detection is not a trivial computational task it is fairly simple for a human and this part of the annotation could be carried out fast and reliably by research assistants which did not have to be trained in the complexities of full phrase structure annotation.

After running IceNLP we ran a few programs developed within the project to prepare the text for manual annotation. The PoS tagset was converted to a format nearly identical to the Penn Parsed Corpora of Historical English, the format of the labeled bracketing was converted to the Penn treebank format for compatibility with existing software and various structures were partially annotated using CorpusSearch revision queries (RANDALL 2005). Such partial annotation includes building the left edge of subordinate clauses whose right edge is subsequently determined by a human annotator.

The manual annotation phase comprised the bulk of the work. In the beginning, we used the CorpusDraw software to correct the parse, but we soon realized that it would be possible to speed up the annotation if we had software that was better suited for the task. Therefore, we wrote the annotation tool Annotald which made it possible to speed up the annotation considerably. Annotald is a browser based cross-platform visual tree editor which combines keyboard and mouse shortcuts such that the annotator can always keep the left hand on the keyboard and the right hand on the mouse. This avoids moving the right hand back and forth between mouse and keyboard which leads to improved speed and accuracy over CorpusDraw (see Figure 1 for the overall impact of improved methods and training on tree pro-

duction). Annotald is released under the LGPL license and continues to be developed by a growing team of programmers at the University of Pennsylvania (the latest version is BECK et al. 2011).

Three annotators worked on the project. In the beginning, they reviewed each other's work and spent a lot of time consulting the annotation manual for the Penn Historical Corpora (SANTORINI 2010), which we succeeded in adapting to Icelandic. When the annotators had become well acquainted with the annotation scheme and developed special annotation rules for most of the cases where Icelandic deviates from Old(er) English, they stopped reviewing each other's annotations and placed the emphasis on speeding up the annotation as much as possible. Figure 1 shows the annotation progress for the whole project period.

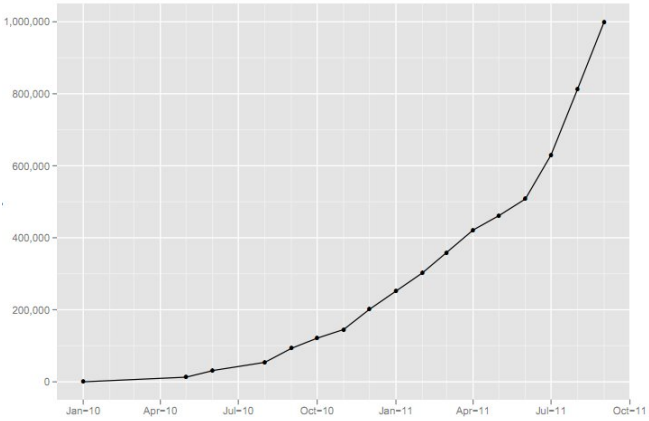


Figure 1: The annotation process

We are in no doubt that the speed of the annotation process, and the fact that a large part of the annotation has not been reviewed, has resulted in a considerable number of annotation errors and discrepancies. The errors are nevertheless a small minority of potential errors. Our current approach to correction is to systematically enforce more constraints on well-formed structures. For example, the latest release of the corpus (WALLENBERG et al. 2011) incorrectly contains 51 clauses with two phrases labeled as direct objects (NP-OB1). This is about 1% of the 4727 double object clauses in the corpus and in most of the cases one of the two objects should be labeled as an indirect object (NP-OB2).

These errors and many more have been corrected for the next release of the corpus. However, we want to emphasize that the corpus is meant to be used for quantitative research, not qualitative. It is not possible to take the parsing of any single sentence from the corpus and rely on it without reflection. The reasons are both that the text may be of disputable age and the parse may contain an error. However, we believe that the quantity of the text and its overall quality make the corpus safe to use in quantitative studies.

8 Maximizing distribution and user freedom

We believe strongly in the sharing of resources. True to that spirit, we decided at the beginning of the project that we wanted to make our work as open and widely distributed as possible. To emphasize that, we defined the following "10 basic types of user freedom":

1. Raw data available can be downloaded for local use (corpus not hidden behind a search interface)
2. Comprehensive documentation freely available online
3. Available without registration, user identification of some sort, or signing of contracts
4. Development process of corpus relies only on free/open source software tools (for transparent replication of annotation process)
5. Open development (annotation is carried out in an open online version control repository for transparency regarding the actual steps taken in the development and immediate access to work-in-progress)
6. Regular scheduled releases of numbered versions during development as well as for more permanent milestone versions so that researchers can always produce replicable results on a recent version of the corpus
7. Users can improve the corpus and release modified versions without special permission
8. Free of cost to academia
9. Free of cost to commercial users
10. Corpus released under a standard free license of some sort for straightforward compatibility with other projects (GPL, LGPL, CC, etc.)

We decided not to wait until the treebank was finished to release it. Instead, we released a new version every three months, in the hope of incrementally building up a user base and getting feedback from users which we could use to improve the treebank. This worked very well – for instance, Version 0.4, released in April 2011 and containing around 440,000 words, was downloaded (in different formats) more than 450 times from the project website ((*http://linguist.is/icelandic_treebank/Download*). Furthermore, the treebank had already been used in a number of studies before the current version was released in August 2011 (for instance SAPP 2011; INGASON, SIGURÐSSON and WALLENBERG 2011; LIGHT and WALLENBERG 2011).

Even though the treebank is practically finished, the current version is numbered 0.9 because some minor corrections remain to be made. The treebank is released in three versions; a zip-file containing the raw data of the of the corpus in labeled bracketing format; an easy-to-use setup executable for Windows that installs the corpus and a graphical user interface; and a zip-file containing the corpus and a platform independent user interface in Java.

9 Conclusion

In this paper, we have described the parsed historical corpus of Icelandic, IcePaHC, and its motivations. As pointed out in the introduction, the corpus was built in order to serve two purposes: first, to be used within language technology to train parsers etc., and secondly, to be used as a tool for diachronic syntactic research.

Its usefulness for the latter purpose has already been demonstrated. For instance, four papers that were presented at the 13th Diachronic Generative Syntax Conference (DiGS) in

June 2011 made use of the corpus (see *http://www.ling.upenn.edu/Events/DIGS13/*). As for the first purpose, the corpus has not yet been put to use but there is no reason to doubt that it can serve that purpose too. The corpus contains around 300,000 words which can safely be considered Modern Icelandic – texts from the 19th, 20th and 21st centuries. That is more than enough material to train a statistical parser.

As mentioned above, we believe that our corpus is unusual for a number of reasons. The most important one is that we have brought together a group of researchers who come from different fields and have different motives, but saw the benefits of joining their forces in building an important language resource which serves a dual purpose. The interdisciplinarity of the team should ensure that both humanist researchers and language technologists feel at ease in using the corpus in their work.

Finally, we emphasize the importance of distributing language resources under an open source license. This is especially important when working on less-resourced languages where duplication of work must be avoided. We hope that other researchers will follow in our steps and make their resources and tools publicly available for the benefit of all.

Litterature

BECK, J.E. (2011). Penn Parsed Corpora of Historical Greek (PPCHiG). (*http://www.ling.upenn.edu/~janabeck/greek-corpora.html*).

BECK, J.E., A. ECAY and A.K. INGASON (2011). Annotald, version 11.11. [Software for treebank annotation.] (*http://github.com/janabeck/Annotald*).

BERNHARÐSSON, H. (1999). Málblöndun í sautjándu aldar uppskriftum íslenskra miðaldahandrita. Reykjavík: Instite of Linguistics, University of Iceland.

BICK, E. (2003). Arboretum, a Hybrid Treebank for Danish. In: Nivre, J., and E. Hinrichs (eds.) (2003). TLT 2003. Proceedings of the Second Workshop on Treebanks and Linguistic Theories, 14-15 November 2003, Växjö, Sweden. Växjö: Växjö University Press, 9-20.

EYTHÓRSSON, T., and B.M. KARLSSON (2011). Greinir skáldskapar: an Annotated Corpus of Old Icelandic Poetry. Paper presented „Bragarmál“, an international conference on Germanic and Icelandic metrics, University of Iceland, Reykjavík, September 24th, 2011.

GALVES, C., and H. BRITTO (2002). The Tycho Brahe Corpus of Historical Portuguese. Department of Linguistics, University of Campinas. Online publication, first edition. (*http://www.tycho.iel.unicamp.br/~tycho/*).

INGASON, A.K., S. HELGADÓTTIR, H. LOFTSSON and E. RÖGNVALDSSON (2008). A Mixed Method Lemmatization Algorithm Using a Hierarchy of Linguistic Identities (HOLI). In: Raante, A., and B. Nordström (eds.) (2008). Advances in Natural Language Processing. (Lecture Notes in Computer Science, Vol. 5221.) Berlin: Springer, 205-216.

INGASON, A.K., E.F. SIGURÐSSON and J. WALLENBERG (2011). Distinguishing Change and Stability: a Quantitative Study of Icelandic Oblique Subjects. Paper presented at DiGS 13, University of Pennsylvania, Philadelphia, June 3rd, 2011.

KROCH, A., B. SANTORINI and L. DELFS (2004). The Penn-Helsinki Parsed Corpus of Early Modern English (PPCEME). Department of Linguistics, University of Pennsylvania. CD-ROM, first edition, (*http://www.ling.upenn.edu/hist-corpora/*).

KROCH, A., and A. TAYLOR (2000). The Penn-Helsinki Parsed Corpus of Middle English (PPCME2). Department of Linguistics, University of Pennsylvania. CD-ROM, second edition, (*http://www.ling.upenn.edu/hist-corpora/*).

KROMANN, M.T. (2003). The Danish Dependency Treebank and the DTAG Treebank Tool. In: Nivre, J., and E. Hinrichs (eds.) (2003). TLT 2003. Proceedings of the Second Workshop on Treebanks and Linguistic Theories, 14-15 November 2003, Växjö, Sweden. Växjö: Växjö University Press, 217-220.

LIGHT, C. (2010). Parsed Corpus of Early New High German. (*http://enhgcorpus.wikispaces.com/home*).

LIGHT, C., and J. WALLENBERG (2011). On the Use of Passives across Germanic. Paper presented at DiGS 13, University of Pennsylvania, Philadelphia, June 4th, 2011.

LOFTSSON, H. (2008). Tagging Icelandic text: A linguistic rule-based approach. Nordic Journal of Linguistics 31(1):47-72.

LOFTSSON, H., and E. RÖGNVALDSSON (2007). IceParser: An Incremental Finite-State Parser for Icelandic. In: Nivre, J., H.-J. Kaalep, K. Muischnek and M. Koit (eds.) (2007). NODALIDA 2007 Conference Proceedings. Tartu: University of Tartu, 128-135.

MARTINEAU, F., P. HIRSCHBÜHLER, A. KROCH and Y.C. MORIN (2010). Corpus MCVF (parsed corpus), Modéliser le changement : les voies du français, Département de français, University of Ottawa. CD-ROM, first edition (*http://www.arts.uottawa.ca/voies/voies_fr.html*).

NIVRE, J., J. NILSSON and J. HALL (2006). Talbanken05: A Swedish Treebank with Phrase Structure and Dependency Annotation. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC). Genoa: 1392-1395.

PALMER, D.D., and M.A. HEARST (1994). Adaptive sentence boundary disambiguation. In: Proceedings of the fourth conference on Applied natural language processing. Stroudsburg, PA: Association for Computational Linguistics, 78-83.

RANDALL, B. (2005). CorpusSearch 2 Users Guide. University of Pennsylvania, Philadelphia. (*http://corpussearch.sourceforge.net/CS-manual/Contents.html*).

RÖGNVALDSSON, E., and S. HELGADÓTTIR (2011). Morphosyntactic Tagging of Old Icelandic Texts and Its Use in Studying Syntactic Variation and Change. In: Sporleder, C., A.P.J van den Bosch and K.A. Zervanou (eds.) (2011). Language Technology for Cultural Heritage: Selected Papers from the LaTeCH Workshop series. Berlin: Springer, 63-67.

SANTORINI, B. (1997/2008). The Penn Yiddish Corpus. University of Pennsylvania. For details, contact: beatrice@babel.ling.upenn.edu.

SANTORINI, B. (2010). Annotation manual for the Penn historical corpora and the PCEEC. University of Pennsylvania, Philadelphia. (*http://www.ling.upenn.edu/hist-corpora/annotation/index.html*).

SAPP, C. (2011). A Relative Pronoun in Old Norse? Paper presented at DiGS 13, University of Pennsylvania, Philadelphia, June 5th, 2011.

TAYLOR, A., A. WARNER, S. PINTZUK, AND F. BETHS (2003). The York-Toronto-Helsinki Parsed Corpus of Old English Prose. University of York. (*http://www.users.york.ac.uk/~lang22/YcoeHome1.htm*)

WALLENBERG, J., A.K. INGASON, E.F. SIGURÐSSON and E. RÖGNVALDSSON (2011). Icelandic Parsed Historical Corpus (IcePaHC). Version 0.9. (*http://www.linguist.is/icelandic_treebank*)

Arne Skjærholt

More, faster: Accelerated corpus annotation with statistical taggers

We present our experiments with annotating a Latin corpus using an assisted annotation procedure where the corpus to be annotated is pre-annotated by a statistical tagger. This assisted procedure gives a notable reduction in annotator error compared to the unassisted annotation of previous annotation efforts, even with a huge tagset (1 000 tags) and modest tagger accuracy due to limited training data and domain effects.

1 Introduction

When creating corpora of richly inflecting languages like Latin, the most time-consuming (and boring) task is morphological annotation. Qualified labour for a classical language is also hard to come by, adding unneeded strain to already limited budgets and slowing down the pace of corpus development. Thus, we would very much like to speed up this process.

We present here a simple, effective and cheap way of achieving this. Using almost no custom components, relying instead on off-the-shelf software, we leverage an existing Latin corpus to accelerate annotation of a new text with the help of an HMM tagger whose output is corrected by annotators. Even though the tagger is far from the 95% accuracy of the state-of-the-art in tagging in general and the tagset is extremely large (more than 20 times larger than the Penn Treebank tagset), we improve both the speed and error rate of manual annotation considerably.

After a quick review of related research, both into statistical taggers for Latin and annotation assisted by taggers, we present the corpus used for our experiments and its annotation procedure, as well as a brief outline of the particularities of Latin as a language. Then we present the details of our experiments: the accuracy of the tagger itself and how it compares to previous taggers for Latin, the effect of assisted annotation on annotation speed, and its effect on annotator error. Finally, we sum up our conclusions from these experiments and mention some possible avenues for future research.

1.1 Previous work

The automatic analysis of Latin morphology has been the subject of a few previous studies. Poudat and Longrée (2009) used the LASLA corpus[1] to explore the automatic analysis of Latin morphology with HMMs, and the influence of factors such as author,

[1] http://www.cipl.ulg.ac.be/Lasla/

genre and time period on tagging performance. Skjærholt (2011) used the PROIEL corpus (more on this in the next section) to compare the viability of HMMs versus the more sophisticated CRF models and studied the possibility of using constrained decoding to increase tagger performance on out-of-domain data. Bamman and Crane (2008) and Passarotti (2010) also explore statistical tagging of Latin in the greater context of developing lexical resources; they both achieve comparable results to the more in-depth studies, Bamman and Crane (2008) using TreeTagger (Schmid, 1994) and Passarotti (2010) using HunPos (Halácsy et al., 2007).

The basic idea of using a computer to generate output to be post-processed is essentially the same as Bar-Hillel's (1960) suggestion that machine translation output be corrected by a human translator and the translator's amanuensis proposed in Kay (1997). The concrete idea of letting human annotators correct tagger output rather than starting from scratch is a common one, and has been explored in depth by several studies. When creating the Penn Treebank, Marcus et al. (1993) found that manual tagging took twice as long and gave double the inter-annotator disagreement compared to correcting tagger output, a position largely corroborated by Fort and Sagot's (2010) more in-depth study of the influence of tagger accuracy. However, their experiments used the Penn Treebank, and we need to verify that their results still hold with a tagset as large as ours. Furthermore, Fort and Sagot used in-domain data to train their tagger, whereas we use out-of-domain data, which means that we require more data to get similar tagger performance. Dandapat et al. (2009) are guardedly optimistic, but they too show that correcting tagger output yields better data than annotating from scratch.

2 Language & corpus

2.1 A crash course in Latin

Latin is an Indo-European language with a long and interesting history, ancestor of the Romance languages of today. The very first traces of Latin language date back to eight century BCE, and the oldest literature to survive until our day, the comedies of Plautus, date to 200 BCE or thereabouts. The language is typical of classical Indo-European languages of roughly the same age, such as Ancient Greek and Sanskrit; it is a richly inflecting language with synthetic morphology and a large array of forms and morphemes, even though the Latin system represents a radical departure and restructuring of the ancestral system better preserved in Greek and Sanskrit.

The history of Latin is usually divided into several periods, the most important being Classical Latin, which dates from around the 1st century BCE to the first century CE; most of the Latin authors commonly known today, like Caesar, Cicero, Virgil and Horace, belong to this era. Anything earlier than Classical Latin is counted as Old Latin. After the Classical era, the language starts to split into two languages: Vulgar Latin, the language of the people, already quite different from the literary language in the Classical era, becomes more and more a separate language, eventually becoming

the Romance languages. The literary language on the other hand remains relatively unchanged for several centuries.

Roughly speaking, the morphology of Latin (which is the interesting part, vis-a-vis the present work) can be divided into two largely independent parts: nominal inflection and verbal inflection. The verbal system governs all finite forms of the verb, while the nominal system covers inflection of the remaining infinite forms of the verb, nouns, and adjectives. A few words fall outside of these two groups, most notably the pronouns. Both the nominal and verbal systems are further subdivided into five declinations and four conjugations, respectively; these subdivisions are again more or less independent, forming the different forms with different morphemes, especially in the case of the nominal system. Finally, it is quite common for several forms of a word to be identical, especially in the nominal system.

2.2 Corpus

For our experiments, we used the Pragmatic Resources of Old Indo-European Languages[2] (PROIEL) corpus to train the tagger whose output the annotators correct and to gather data from unassisted annotation to compare our assisted approach with. The PROIEL project aims to study the pragmatics of several classical IE languages (Ancient Greek, Old Church Slavic, Classical Armenian, Gothic, and Latin) by creating a large parallel corpus of several such languages, to allow for large-scale contrastive analysis. The main part of the corpus is the translation[3] of the New Testament in the respective language, but some other texts from the various languages are included as well.

The corpus is morphologically annotated with two tagsets: a part-of-speech (PoS) tagset for features belonging to the lemma, and a morpho-syntactic descriptor (MSD) tagset for features that vary according to the form of the word. The PoS tagset is relatively coarse with only 23 tags, corresponding to the ten parts of speech of traditional grammar, augmented with finer subdivisions for some parts of speech (most notably nine kinds of pronoun) and a foreign word class; the full list is given in table 1. The MSD tagset is a fixed-width format ten characters wide, where each position corresponds to a particular morphological feature such as case, mood or tense; a list of tags relevant to Latin is presented in table 2. The PoS tag is attached to the lemma of a word, and MSD tags to each token in the corpus, and in total there are 962 distinct PoS-MSD pairs in the Latin part of the corpus. The syntactic annotation is in the style of dependency grammar, with the addition of secondary dependencies to fill the external roles of open functions, similar to structure sharing in LFG and HPSG (Haug, 2010, 1).

The PROIEL annotation procedure is somewhat idiosyncratic; instead of each sentence being annotated by two independent annotators and then resolving any disagreements, the PROIEL annotation procedure is in two steps: First an annotator (graduate students, for the most part) analyses the morphology and syntax of the sentence. The

[2] `http://foni.uio.no:3000/`
[3] Or original, in the case of ancient Greek

Tag	Meaning	Tag	Meaning
A–	adjective	Pc	reciprocal pronoun
C–	conjunction	Pd	demonstrative pronoun
Df	adverb	Pi	interrogative pronoun
Dq	relative adverb	Pk	personal reflexive pronoun
Du	interrogative adverb	Pp	personal pronoun
F–	foreign word	Pr	relative pronoun
G–	subjunction	Ps	possessive pronoun
I–	interjection	Pt	possessive reflexive pronoun
Ma	cardinal numeral	Px	indefinite pronoun
Mo	ordinal numeral	R–	preposition
Nb	common noun	V–	verb
Ne	proper noun		

Table 1: PoS tagset

Position	Feature	Values
1	Person	1st (1), 2nd (2), 3rd (3)
2	Number	singular (s), plural (p)
3	Tense	present (p), imperfect (i), future (f), perfect (r), pluperfect (l), future perfect (t)
4	Mood	indicative (i), subjunctive (s), imperative (m), infinitive (n), participle (p), gerund (d), gerundive (g), supine (s)
5	Voice	active (a), passive (p)
6	Gender	masculine (m), feminine (f), neuter (n), m/n (o), m/f (p), m/f/n (q), f/n (r)
7	Case	nominative (n), vocative (v), accusative (a), genitive (g), dative (d), ablative (b)
8	Degree	positive (p), comparative (c), superlative (s)
9	Unused[a]	—
10	Inflection	inflecting (i), non-inflecting (n)

[a] Used for strong/weak inflection in Gothic and Old Church Slavonic

Table 2: MSD tagset

Corpus	Sentences	Tokens	Avg. tok/sen
BG	1 417	26 663	18.8
Vulgata	12 459	112 135	9.0
Peregrinatio	921	17 553	19.1

Table 3: Annotated corpus sizes at the time of writing

sentence is then reviewed by a more senior annotator to ensure that the analysis is correct (Haug and Jøhndal, 2008, 27–28).

The Latin part of the corpus is comprised of three texts: the *Vulgata* translation of the Bible, Caesar's *Bellum Gallicum* (*BG*) and *Peregrinatio Aetheriae*, a 5th century Vulgar Latin account of a pilgrimage to the Holy Land. Of the three, the *Vulgata* corpus is by far the largest at more than 100 000 annotated tokens, with the *BG* corpus at 25 000 tokens. Detailed statistics are given in table 3. We omit the *Peregrinato* corpus from the experiments in the present work, since the Vulgar Latin of this text is simply too different from the Classical Latin of Caesar and the literary style of Jerome's *Vulgata*. In particular, the restructuring of the rich morphological system of Latin into the more modest Romance system is well under way, which means that many inflections are used in ways that are flat out wrong in the more classically informed Latin.

Many students' first encounter with Latin is Caesar's *BG*, and its opening sentences will serve us well as an example:

Gallia est omnis divisa in partes tres, quarum unam incolunt Belgae, aliam Aquitani, tertiam qui ipsorum lingua Celtae, nostra Galli appellantur. Hi omnes lingua, institutis, legibus inter se differunt. Gallos ab Aquitanis Garumna flumen, a Belgis Matrona et Sequana dividit.

In all, Gaul is divided in three. Of these, the Belgians inhabit one, the Aquitans another, and those who are called Celts in their own language, or Gauls in our own, inhabit the third. All of them differ between each other in language, traditions and laws. The river Garonne separates Gauls from Aquitans, and the Seine and Marne from the Belgians.

The morphological annotation of the opening of the first Latin sentence, corresponding to the first sentence of the translation, is shown in figure 1; the first two characters correspond to the PoS tags of table 1 and the remaining ten to the MSD tagset of table 2. Thus, *in* is a preposition (R–) which is indeclinable (––––––––n) and *divisa* the feminine nominative singular of the perfect passive participle (–srppfn––i) of a verb (V–).

Finally, a brief word on how PROIEL compares to the LASLA corpus. First of all, the latter project started in 1961, which means that the size of the corpus is quite

Gallia	est	omnis	divisa
Ne-s---fn--i	V-3spia----i	Px-s---pn--i	V--srppfn--i
in	partes	tres	[quarum ...]
R----------n	Nb-p---fa--i	Ma-p---pa--i	

Figure 1: Morphological annotation of *BG* 1.1.1

significant: 1.6 million words[4], an impressive figure compared to the 140 000 words in the PROIEL corpus. The LASLA corpus is primarily morphologically annotated however, only a limited amount of information related to verbs is annotated. Second, the LASLA tagset is quite a bit larger than the PROIEL tagset, due to its encoding of inflectional classes in the PoS part of the tagset; Poudat and Longrée (2009) report a total of 3 732 distinct tags, more than three times the 960 tags in the PROIEL corpus. Unfortunately, the raw data of the LASLA corpus aren't publicly available, so we couldn't use it to train our tagger.

Index Thomisticus and the Perseus Latin Dependency Treebank are two other publicly available corpora of Latin. In general, more training data for a statistical model is a good thing, but we decided not to use these sources. Index Thomisticus is a treebank of the works of Thomas Aquinas, and is medieval Latin and excluded on the same basis as *Peregrinatio*. The Perseus LDT is a 50 000 token treebank, made up of selections of various important Latin author's work. Linguistically, these texts are a good fit with our training corpus, but unfortunately the Perseus and PROIEL tagsets are not the same, and converting from the Perseus tagset to PROIEL is a non-trivial task which would most likely introduce quite a bit of noise to our data.

3 Assisted annotation

In order to investigate the properties of assisted annotation and its efficacy for the annotation of Latin, we selected a new text for annotation. The text to be annotated is Cicero's *Epistulae ad Atticum* (*Att*), a collection of letters to his friend Titus Pomponius Atticus. This is a fairly large corpus, composed of 4 561 sentences and 61 193 tokens (giving an average of 13.4 tokens per sentence). Linguistically and stylistically, this text is most closely aligned with Caesar's *BG* rather than the later (and simpler) *Vulgata* text.

At the outset, the primary objective of this new assisted annotation procedure is to provide a faster annotation rate compared to the unassisted annotation. It would also be nice if the assisted annotation results in better annotation (that is, fewer errors) compared to the unassisted procedure. We will quantify both of these dimensions. Inter-annotator agreement is another standard measure of annotation quality; we will

[4] `http://www.cipl.ulg.ac.be/Lasla/tlatins.html`, retrieved 5/9/2011

not quantify this, for the simple reason that it is not possible to do so with our present dataset, since each sentence is only annotated by a single annotator.

3.1 The tagger

The tagger we used for our experiments is Thorsten Brants' Trigrams'n'Tags (TnT), described in Brants (2000). TnT is a fairly straightforward trigram HMM tagger, but with one important addition: the unknown word model. Instead of estimating emission probabilities of words not seen in training by some kind of discounting strategy, the suffixes of words seen in training are matched against the suffixes of the unknown word, and the emission probability of the longest matching suffix is used as the word's emission probability. This strategy works very well for Latin, since it's exclusively suffix-inflecting.

The model used to pre-process the *Att.* corpus was trained on the concatenation of the *Vulgata* and *BG* corpora, using TnT's default options. The tagger output was then combined with a partial finite-state morphology that was made available to us, such that if TnT's analysis was one of those licenced by the morphology, the lemma of the finite-state analysis was added as well. If the analyses did not match, the lemma was set to "FIXME".

A brief interview with the annotators who have annotated the new corpus makes it clear that the preprocessed corpus at least makes the annotation work more bearable. However, we would like to quantify the effects of preprocessing the corpus with the statistical taggers as well. An important first datum is simply the raw performance of the tagger, and how this compares to previous results. Both Poudat and Longrée (2009) and Skjærholt (2011) evaluate tagger performance on in-domain and out-of-domain (OOD) data, and despite the important differences between the two corpora, obtain remarkably similar results, summarised in table 4. In particular, their results result in an overall accuracy of 84.3%, even though Poudat and Longrée's training corpus is larger than Skjærholt's. Given this, it seems reasonable to believe that our results using the PROIEL data can be meaningfully compared to Poudat and Longrée's other results. The accuracy of 76.9% is encouraging and somewhat unexpected, given previous tagging results; before evaluating the tagger we expected a result closer to the 63% of row *e*. The *c* result of 77.2% in table 4 is very close to our final accuracy, but the training corpus in that experiment was quite a bit larger: 352 820 tokens compared to our 139 620.

Instead of using a statistical tagger, another option would be to use a rule-based tagger, such as Words by William Whitaker[5], Morpheus (Crane, 1991), originally developed for classical Greek and later adapted to Latin, or Lemlat (Passarotti, 2000). However, such a tagger outputs all possible analyses for an ambiguous token, which doesn't fit very well with the DB schema of the annotation tool, and it is preferred that the annotators correct a single analysis rather than choose from a potentially long list of options.

[5] `http://ablemedia.com/ctcweb/showcase/whitakerwords.html`

Experiment	TA	OOV	IV
Poudat and Longrée (2009)[a]	84.3	?	?
Poudat and Longrée (2009)[b]	63.7	?	?
Poudat and Longrée (2009)[c]	77.2	?	?
Skjærholt (2011)[d]	84.3	60.7	88.9
Skjærholt (2011)[e]	62.8	33.3	85.0
Vulgata & *BG* on *Att*	76.9	50.0	85.7

[a] LASLA, *BG* books 1–2,4–7 on book 3
[b] LASLA, *BG* and *Bellum Civile* on *1st Catilinarian*
[c] LASLA, historical texts on *1st Catilinarian*
[d] PROIEL, *BG* 10-fold cross-validation
[e] PROIEL, trained on *BG*, tested on *Vulgata*

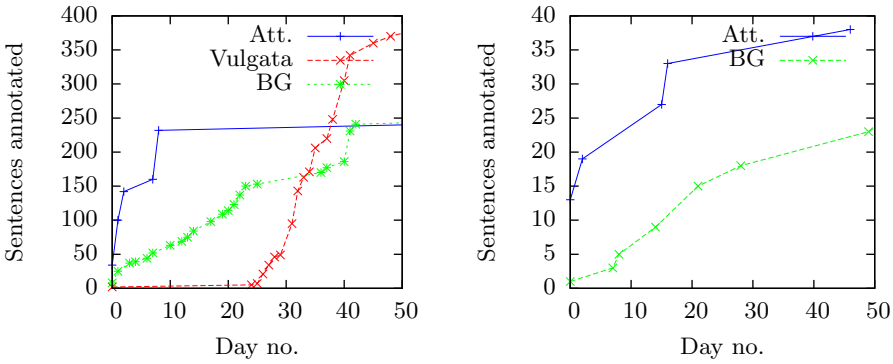
Table 4: Tagging accuracy (in percent) on Latin. Token accuracy (TA), out-of-vocabulary (OOV) and in-vocabulary (IV) accuracy.

3.2 Annotation speed

Annotation speed is harder to gauge accurately. The only information available from the PROIEL DB dump is a timestamp, the date and time when the annotator saved the annotation. Thus, information like time spent per sentence or how many sentences were annotated in a single sitting is not explicitly represented in the data. One could conceivably synthesise a number of sentences per sitting, for example by grouping sentences annotated within some threshold, say five or ten minutes, as belonging to the same annotation session. We have not done this however. Instead we truncate the timestamps to the date portion and count only the number of sentences annotated by each annotator each day. No matter which approach is used to synthesise such a statistic from the timestamps would involve drawing some arbitrary line in the sand, and we believe this approach to be the least problematic approach.

There are two annotators, Aulus and Gaius, both master's students, who have annotated the *Att.* corpus. Aulus has done the majority of the annotation (420 sentences) while Gaius has annotated 38 sentences. Aulus is the more experienced annotator, having previously annotated 4 207 sentences of the *Vulgata* corpus and 852 sentences in *BG*, while Gaius has annotated 332 sentences of *BG* before the *Att.* annotation. This gives us something to compare the *Att.* effort with. Figures 2 (a) and (b) show the total number of sentences by each annotator along the y-axis and number of days since his first annotation along the x-axis.

In the case of Aulus, it seems quite clear that annotation of the *Att.* corpus is helped by the pre-annotation; compared to the *BG* corpus, 10 days of annotation has produced what took 40 days without pre-annotation. The *Vulgata* graph has roughly the same slope, without pre-annotation, as the *Att.* effort, but this text is significantly simpler



(a) Aulus (b) Gaius

Figure 2: Annotation speeds. Note that the y-axes of the two figures are different, due to the difference in number of sentences annotated.

and not as comparable to Cicero's text as Caesar's. For Gaius the picture is less clear, but it seems that the assisted annotation of the *Att.* corpus is a bit faster than the unassisted annotation of *BG*.

A more rigorous way to test these notions is to apply hypothesis testing to the data. Table 5 shows the sufficient statistics to apply Student's t-test to the dataset: the number of datapoints (n), the mean number of sentences annotated per day (μ) and the sample standard deviation (s). Applying the t-test to these data, we find that Aulus' annotation of *Att.* is significantly different from his annotation of *BG*, but not his annotation of *Vulgata*, nor are Gaius' two annotation series significantly different (all at the 95% level).

3.3 Annotator error

Another important metric is annotator error. To quantify this we extract all the sentences that have been reviewed by the expert annotator (as per the PROIEL annotation procedure outlined in section 2.2); of Aulus' sentences 25 have been reviewed, and 32 of Gaius'. Using audit data tracking changes done to the tokens, we counted the number of tokens whose morphology had been changed since the sentence was annotated, the results of which are summarised in table 6. The numbers presented are token error (TE), sentence error (SE), the number of sentences with errors (n), the mean number of mistagged tokens per sentence with errors (μ) and the sample standard deviation of the number of tokens per mistagged sentence (s).

In the case of Aulus, we cannot meaningfully compare the number of errors per mistagged sentence in *Att.* with the others since there are only two such sentences, each

Annotator	n	μ	s
Aulus, *Vulgata*	96	43.8	39.8
Aulus, *BG*	55	15.5	13.6
Aulus, *Att*	9	46.7	23.6
Gaius, *BG*	44	7.55	4.61
Gaius, *Att*	5	7.60	3.21

Table 5: Annotation statistics. Number of days with annotation (n), number of sentences annotated per day mean (μ) and standard deviation.

Annotator	TE	SE	n	μ	s
Aulus, *Vulgata*	2.80	18.8	545	1.28	0.628
Aulus, *BG*	8.27	70.3	415	2.17	1.35
Aulus, *Att*	0.529	8.00	2	1.00	0.00
Gaius, *BG*	7.44	66.9	222	2.52	1.89
Gaius, *Att*	2.11	9.38	3	2.67	1.53

Table 6: Annotator error. Token error (TE) and sentence error (SE) in percent, number of mistagged sentences (n), number of erroneous tokens per mistagged sentence mean (μ) and sample standard deviation (s).

with a single mistagged token; this gives a standard deviation of zero, which again means that no matter the confidence level, the bound on the mean will always be ±0. Gaius' data one the other hand, do not have this problem; his numbers of wrong tokens per mistagged sentence are not significantly different at the 95% level.

Taking a somewhat broader perspective yields a quite pleasing view as well. Even though the number of errors per sentence once an error is made appears to be relatively unchanged, the number of errors made is reduced dramatically: Annotation error at the token level is reduced by almost a factor of four, from 7.44% to 2.11%, for the junior annotator and more than a full order of magnitude to a mere half percent for the more experienced Aulus. Sentence-level error is likewise encouraging, reduced by almost an order of magnitude for both annotators, from the neighbourhood of 70% to slightly less than 10%.

4 Conclusion & future work

All in all, the results of our study are very encouraging. Our experienced annotator, Aulus, benefits the most from assisted annotation. His annotation speed increased dramatically, with our proxy for annotation speed tripling compared to the unassisted annotation of the *BG* corpus, which is comparable in terms of linguistic complexity; his error rate was reduced by an order of magnitude, on both token and sentence level. Gaius, the less experienced annotator, had no measurable change in annotation speed, but he too made far fewer errors both in terms of tokens and sentences.

Based on this evidence we believe that assisted annotation is an excellent tool, even for annotation tasks with huge tagsets, and that if data is available to train a tagger, the assisted approach is preferable to unassisted annotation, both in terms of annotator error and annotation speed. For annotation error both annotators had a sizeable decrease in error, but for speed only one of the two annotators showed an improvement; however, given that our proxy measure for speed was tripled in the case of Aulus, and the assisted value is almost two and a half standard deviations from the unassisted value, we believe this to be indicative of a real improvement.

4.1 Future work

This work is a good start, but questions remain that we would like to see answered. First of all, a more in-depth study of assisted annotation using this kind of large tagset would be welcome. We have obtained good preliminary data, but certain metrics are unavailable to us given the nature of our dataset; chief among these are inter-annotator agreement and direct measurement of annotation speed. It would also be interesting to investigate further the influence of tagger accuracy on the various metrics. Fort and Sagot (2010) suggest that tagger accuracy in the 66–82% range is sufficient, and our tagger accuracy in the high seventies is consistent with these results; but since their work uses the fairly small Penn Treebank tagset one should verify that their results hold for extremely large tagsets as well. We would also like to investigate further if

the kinds of error the annotators make differ qualitatively from the errors made with unassisted annotation.

References

Bamman, D. and Crane, G. (2008). Building a Dynamic Lexicon from a Digital Library. In Larsen, R., editor, *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries1*, pages 11–20, New York. ACM.

Bar-Hillel, Y. (1960). The Present Status of Automatic Translation of Languages. *Advances in Computers*, 1:91–163.

Brants, T. (2000). TnT: a statistical part-of-speech tagger. In Nirenburg, S., editor, *Proceedings of the sixth conference on applied natural language processing3*, ANLC '00, pages 224–231, Stroudsburg. Association for Computational Linguistics.

Crane, G. (1991). Generating and Parsing Classical Greek. *Literary and Linguistic Computing*, 6(4):243–245.

Dandapat, S., Biswas, P., Choudhury, M., and Bali, K. (2009). Complex Linguistic Annotation - No Easy Way Out ! A Case from Bangla and Hindi POS Labeling Tasks. In Stede, M. and Huang, C.-R., editors, *Proceedings of the third linguistic annotation workshop*, pages 10–18, Stroudsburg. Association for Computational Linguistics.

Fort, K. and Sagot, B. (2010). Influence of Pre-annotation on POS-tagged Corpus Development. In Xue, N. and Poesio, M., editors, *Proceedings of the fourth linguistic annotation workshop*, pages 56–63, Stroudsburg. Association for Computational Linguistics.

Halácsy, P., Kornai, A., and Oravecz, C. (2007). HunPos: An open source trigram tagger. In *Proceedings of the 45th Annual Meeting of the ACL*, pages 209–212, Stroudsburg. Association for Computational Linguistics.

Haug, D. T. T. (2010). PROIEL Guidelines for Annotation.

Haug, D. T. T. and Jøhndal, M. L. (2008). Creating a parallel treebank of the old Indo-Eurpoean Bible translations. In Sporleder, C. and Ribarov, K., editors, *Proceedings of the Sixth International Language Resources and Evaluation1*, pages 27–34.

Kay, M. (1997). The Proper Place of Men and Machines in Language Translation. *Machine Translation*, 12(1-2):3–23.

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Passarotti, M. (2000). Development and perspectives of the Latin morphological analyser LEMLAT. *Linguistica Computazionale*, 3:397–414.

Passarotti, M. (2010). 7th SaLTMiL Workshop on Creation and use of basic lexical resources for less-resourced languages Workshop programme Additional Referees. In Sarasola, K., Tyers, F. M., and Forcada, M. L., editors, *Proceedings of the 7th SaLTMiL Workshop on Creation and use of basic lexical resources for less-resourced languages*.

Poudat, C. and Longrée, D. (2009). Variations langagières et annotation morphosyntaxique du latin classique. *Traitement Automatique des Langues*, 50(2):129–148.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing.*

Skjærholt, A. (2011). *Ars flectandi: Automated morphological analysis of Latin.* Master's thesis, University of Oslo.

Maria Sukhareva, Zahurul Islam, Armin Hoenen, Alexander Mehler

A Three-step Model of Language Detection in Multilingual Ancient Texts

Ancient corpora contain various multilingual patterns. This imposes numerous problems on their manual annotation and automatic processing. We introduce a lexicon building system, called Lexicon Expander, that has an integrated language detection module, Language Detection (LD) Toolkit. The Lexicon Expander post-processes the output of the LD Toolkit which leads to the improvement of f-score and accuracy values. Furthermore, the functionality of the Lexicon Expander also includes manual editing of lexical entries and automatic morphological expansion by means of a morphological grammar.

1 Introduction

For more than a decade, ancient languages have been an object of research in computational humanities and related disciplines (Smith et al., 2000; Bamman et al., 2008; Bamman and Crane, 2009; Gippert, 2010a). This relates to building morphosyntactic resources (Passarotti, 2000; Koster, 2005), co-occurrence networks (Büchler et al., 2008; Mehler et al., 2011a) and dependency treebanks, which often focus on texts in Latin or Greek (Bamman and Crane, 2009; Passarotti, 2010). As these efforts concern dead and, thus, low-resource languages, the morphological, syntactic and semantic analysis of them is a challenging task. As a consequence, *manual* annotation is an indispensable companion of building resources out of ancient texts that can be used as reliable input to the various tasks of NLP. This holds especially for ancient languages as, for example, Avestan, Old High German (OHG) or Old Georgian (Gippert, 2006), which – unlike Latin and Greek – are less common objects of computing in the humanities. In these cases, corpus annotation is often accompanied by the manual generation of a full-form lexicon as a prerequisite of building lemmatizers and taggers for these languages.

A central challenge of annotating such corpora together with building corpus-specific lexica is the multilingualism of ancient texts. This relates to texts that contain word forms of different languages as a result of, for example, ancient annotations (Migne, 1855) or of fragments of different translations (Gippert, 2010b). An example of a corpus that mixes source texts with notes of different languages is the *Patrologia Latina (PL)* (Section 3). Another example are documents (e.g., in OHG) that contain a multitude of words borrowed from another language (e.g., Latin). In all these cases, corpus building and lexicon formation are faced with the task of detecting and separating the corresponding source languages correctly – *starting from the level of tokens via the level of sentences up to the level of whole paragraphs.* This is needed since any subsequent

step of preprocessing (e.g., lemmatization or PoS tagging) is sensitive to the underlying language. Thus, in order to apply different preprocessors in a language-specific manner, one needs to know the language of any segment of the input texts. Moreover, the build-up of full-form lexica is by and large a task of manual annotation since taggers are hardly available for low-resource languages such as Old Georgian or OHG. In these cases, one needs to prevent any human annotator from handling, for example, thousands of *French* word forms in a corpus such as the Patrologia Latina if the target language is *Latin*.

In this paper, we introduce a software system called *Lexicon Expander* that supports the build-up of full-form lexica for ancient languages. The Lexicon Expander is part of the *eHumanities Desktop* (Mehler et al., 2011b) (Gleim et al., 2009a) that has been built as an online system of corpus processing in the digital humanities. The Lexicon Expander provides an online interface for lemmatizing and expanding unknown words morphologically. A central ingredient of the Lexicon Expander is a three-step language detector that annotates each input word by the name of the language that it probably manifests. Using these annotations, any human annotator can select language specific subsets of unknown words to handle them separately while leaving behind all words that do not belong to the target language of the lexicon to be built. We evaluate this model in two experiments: the one being based on ancient corpus data, the other being based on samples of modern languages.

The paper is organized as follows, Section 2 briefly discusses other projects that deal with building historical lexica. Section 3 gives an overview of multilingualism in ancient corpora and problems connected with it. Section 4 describes corpus preprocessing. It introduces the Language Detection Toolkit and the Lexicon Expander. Section 5 provides evaluation results. Section 6 discusses findings and draws a conclusion.

2 Related Work

This section gives an overview of systems developed for creating and processing of historical lexica. An extensive work on building historical lexica was done within the framework of the *Improving Access to Text (IMPACT)* project (Balk, 2010). The aim of the project is to digitalize printed documents created before the 19^{th} century. Amongst others, the IMPACT project provides tools for named-entity recognition and lexicon building. As the main task of this project is to heighten the accessibility to historical corpora and to simplify the search process, they provide a toolbox that assigns modern lemmata to historical word forms in order to avoid search problems, caused by historical spelling variations and changes in inflectional morphology. Presently, they provide historical lexica for Dutch and German and a graphical user interface for named entity attestation.

ToTrTaLe (Tokenisation, Transcription, Tagging and Lemmatization) (Erjavec, 2011) is a tool, developed for automatic linguistic annotation and applied to 19^{th} century Slovene. The input of the tool is a TEI encoded corpus. The tool automatically does a morphosyntactic annotation, assigns lemmata and modern equivalents of the

words. Lemmatization is also done by means of assigning modern lemmata to historical word forms. The output of the system is a TEI document containing the annotation information. The historical word forms were manually verified (Erjavec et al., 2011). A specialized editor, called LexTractor (Gotscharek et al., 2011), was used for processing of the historical corpus of Slovene. This web-tool, introduced by (Gotscharek et al., 2009), builds historical lexica with word forms mapped to modern lemmata. The GUI allows to work with unknown words, found in the corpus, and to manually annotate them. A user is asked to accept or to reject readings, proposed by the system. The tool was applied to build a historical German corpus, which currently contains ca. 10,000 entries.

Unlike the lexicon builder, which is presented in this paper, none of the aforementioned tools enables a user to annotate the unknown words with their language. We are not aware of any lexicon building tool, which also applies language detection to historical corpora. In this sense, we provide a tool to "fill this gap".

3 Multilingualism in Ancient Texts

There are various sources of multilingualism in ancient corpora, starting with mere borrowings and ending with comments in foreign languages added by corpus editors. This section briefly discusses the challenges, which multilingualism imposes on the annotation of ancient corpora.

Patrologia Latina (PL) is a collection of documents dating from the 4th till the 13th century. The Patrologia Latina was published in the 19th century. The original printing plates were destroyed in 1868, but lately restored and new editions were published. The Patrologia Latina is comprised of 8,508 documents written by 2,004 authors. The corpus includes over 100 Mio tokens (Mehler et al., 2011a).

The PL Corpus was lemmatized, tokenized and tagged with parts-of-speech (Mehler et al., 2011b). Nevertheless, there are ca. 700 000 tokens in the PL corpus that are marked as unknown. In this case, all the unknown tokens are to be annotated manually, but a more precise look at the unrecognized tokens reveals that a great number of them are not Latin words. The reason of it is that we deal with editions of the Patrologia Latina.

(1) Reliquiae antiquae Scraps from ancient manuscripts illustrating chiefly early english litterature and the english language, edited by Thomas Wright Esq. London, in 8, 2 vol. Remling, Urkundliche Geschichte der ehemaligen Abteien und Klöster in Rheinbaiern Neustadt a.d. Haardt, 1838, in 8,1 - 2. Reschius, Annales ecclesiae Sabinensis. Augustae Vindelic., 1765 in folio, 1-3.

These editions include frequent editors' comments in French, English, German, Italian, Portuguese etc. (1) is an example taken from the PL Corpus, containing editors' comments in English and German and the name of a document in Latin. Filtering out all the unknown words of the foreign origin saves a great amount of annotator's efforts.

The multilingualism problem is found in other corpora as well. The OHG corpus, created in the framework of the TITUS project[1] (Gippert, 2001), is composed of 101 texts with over 400,000 tokens. Some OHG texts are direct translations of Latin texts. Therefore, Latin words, phrases and sentences occur often in such texts. All in all, we found that approximately 9% of the words in the OHG corpus are Latin and 67% of them are single words within a context of OHG. (2) is an example of a bilingual sentence found in the OHG corpus.

(2) **Ein relatio ist patris ad filium ánderiû ist filii ad**
One.OHG relation.L is.OHG father.L to.L son.L other.OHG is.OHG son.L to.L
patrem[. . .]
father.L[. . .]
"One relation is between the Father and the Son, the other one - is between the Son and the Father[. . .]"

We find similar examples in the Avestan corpus (Example (3)). The Avestan corpus features ca. 30,000 words. The texts are written in Avestan, but some text segments are directly translated into one of the following languages: Middle Persian, New Persian, Gujarati, Sanskrit and included in some documents as comments, translations or instructions.

(3) **az xvarəθəm miiazdəm tā ānōh 2 bār guft[. . .]**
from.MP food.AV food sacrifice.AV till.MP there.MP two.MP time.MP speak.MP
"from "food sacrifice" up to here it is to be said twice..."

Summing up, multilingualism in ancient texts is a phenomenon found on lexical, phrasal, sentential and textual levels. In other words, the language can vary from verse to verse or sentence to sentence. Single foreign words and phrases also occur in texts. Filtering out foreign words that do not belong to the target language would simplify numerous tasks of automatic and manual corpus processing, such as lexicon building, corpus annotation, collocation extraction etc.

4 Approach

This section introduces a system for language detection, called *Language Detection (LD) Toolkit*, and the Lexicon Expander, an application module for the eHumanities Desktop (Gleim et al., 2009b). Section 4.2 describes the integration of the LD Toolkit into the Lexicon Expander and the system architecture. Corpus preprocessing is described in Section 4.1. The LD Toolkit is presented in Section 4.3. Finally, we describe the Lexicon Expander, which processes the output of the LD Toolkit in 4.4.

4.1 Preprocessing

The input corpora are annotated with the help of the *PrePro2010 - Text Processing Tool*[1], a text preprocessing tool that automatically does lemmatization, tokenization,

[1] Thesaurus of Indo-European Text and Language Materials (TITUS), http://titus.uni-frankfurt.de.
[1] PrePro2010-Text Processing Tool:http://api.hucompute.org/preprocessor/

sentence boundary detection, stemming, parts-of-speech tagging and named entity recognition (Waltinger, 2010). The resulting TEI P5 files are uploaded in a repository on the *eHumanities Desktop*.

4.2 System Design

The *Lexicon Expander* is an application module, implemented in the framework of the *eHumanities Desktop*, which enables a user to create, organize and annotate lexica.

Figure 1 shows the architecture of the system. A multilingual ancient corpus is converted in a TEI P5 (Sperberg-McQueen and Burnard, 2009) format and saved in a repository in the eHumanitiesDesktop. The user chooses the preprocessed corpus from the repository (Figure 2b). The TEI P5 corpus contains information about lemmata and PoS. Words that were not analysed during the preprocessing are marked with the attribute *"function = unk"*. The Lexicon Expander processes the TEI P5 marked up corpus and extracts such unknown words.

When the unknown words are extracted, the system builds up the lexicon out of them. In order to simplify manual annotation by filtering out words, that do not belong to the target language, the language of the words can be detected before building up the lexicon. The language detection is implemented by means of the LD Toolkit 4.3. The LD Toolkit runs as a background process and the Lexicon Expander can send various input to it. Previous studies (Islam et al., 2011) showed that the LD Toolkit has low *f-score* and *accuracy* if it takes a single word as an input, but if the toolkit is applied to sentences, *f-score* and *accuracy* are high.

In terms of the language detection, the user can choose from three options. First of all, the user can decide not to apply any language detection. In this case lexical entries are added directly into the lexicon without any language assigned. The lexicon is saved into a MySQL database. The second option presupposes that the user can choose to apply the LD Toolkit without the Lexicon Expander model. Then the LD Toolkit gets unknown words as an input. The output of the LD Toolkit is directly saved into the MySQL database. After the GUI of the Lexicon Expander (Figure 2a) is refreshed, the language column in the Lexicon Expander will be filled. The last option is to apply the Lexicon Expander model. In this case, the input of the LD Toolkit is unknown words and sentences, in which these words occur. Finally the output of the LD Toolkit is post-processed by the Lexicon Expander as described in Section 4.4. The language of the unknown words is re-assigned and saved into the MySQL Database.

When lexical entries are saved, the GUI of the Lexicon Expander (Figure 2a) is refreshed and the lexicon is available for further editing (Figure 2c). It is possible to edit the language of a lexical entry manually, specify its part of speech and assign values to grammatical categories. The user can also apply morphological expansion by means of a morphological grammar, defined by a finite-state compiler FOMA (Hulden, 2009). FOMA can assign parts of speech and respective grammatical features. Once the user finished editing the lexical entry, it is saved into the lexicon.

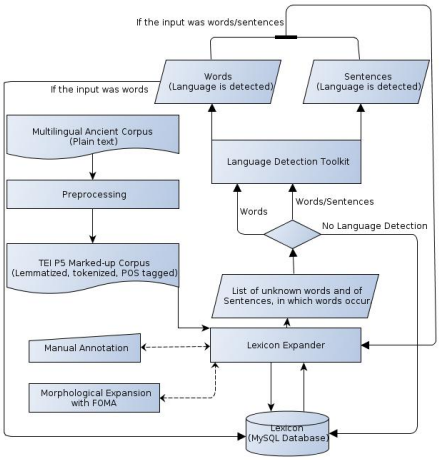


Figure 1: System diagram of the Lexicon Expander

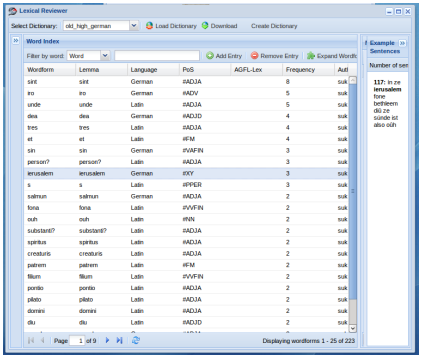
4.3 Language Detection Toolkit

The Language Detection (LD) Toolkit is based on Cavnar and Trenkle (1994) and Waltinger and Mehler (2009). Recently, Islam et al. (2011) have applied this technique to ancient languages. For every target category Islam et al. (2011) learn an ordered list of most frequent n-grams in descending order. The same is done for any input text stream so that categorization occurs by measuring the distance between n-gram profiles of the target categories and n-gram profiles of the input data. The idea behind this approach is that similar texts share features that are equally ordered.

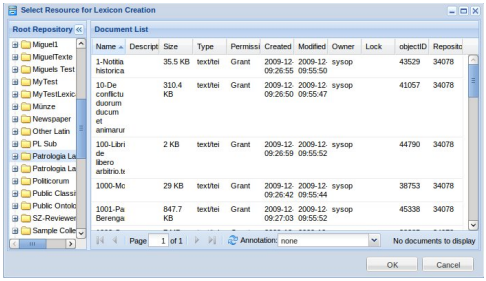
More specifically, classification is done by using the range of corpus features listed in (Waltinger and Mehler, 2009). Predefined information is extracted from the corpus to build sub-models based on these features. Each sub-model consists of a ranked frequency distribution of subsets of corpus features. The corresponding n-gram information is extracted for $n = 1$ to 5. Each n-gram gets its own frequency counter. The normalized frequency distribution of relevant features is calculated according to:

$$\widehat{f_{ij}} = \frac{f_{ij}}{\max_{a_k \in L(D_j)} f_{kj}} \in [0, 1] \tag{1}$$

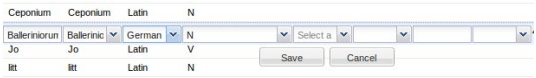
$\widehat{f_{ij}}$ is defined as the frequency f_{ij} of feature a_i in D_j divided by the frequency of the most frequent feature a_k in the feature representation $L(D_j)$ of document D_j (Waltinger and Mehler, 2009). To categorize any document D_m, it is compared to each category C_n using the distance d of the rank r_{mk} of feature a_k in the sub-model of D_m with the corresponding rank of that feature in the representation of C_n:



(a) Lexicon Expander



(b) Repository window



(c) Manual editing of lexical entries

Figure 2: The GUI of the Lexicon Expander

$$d\left(D_m, C_n, a_k\right) = \begin{cases} |r_{mk} - r_{nk}| & a_k \in L(D_m) \wedge a_k \in L(C_n) \\ max & a_k \notin L(D_m) \vee a_k \notin L(C_n) \end{cases} \quad (2)$$

$d\left(D_m, C_n, a_k\right)$ equals max if feature a_k does not belong to the representation of D_m or to the one of category C_n. max is the maximum that the term $|r_{mk} - r_{nk}|$ can assume.

To detect the language of a document, the LD toolkit traverses the document sentence by sentence and detects the language of each sentence (Islam et al., 2011). If the document is homogeneous, (i.e., all sentences belong to the same language), then sentence level detection suffices to trigger other processing tools (e.g., Parsing, Tagging and Morpho-syntactic analysis) that require language detection. In the case that the sentences belong to more than one language (i.e., in the case of a heterogeneous document), the toolkit processes the document word by word and detects the language of each token separately. This step is necessary in the case of multilingual documents that contain words from different languages even within the same sentences. For example: in a scenario of lemmatization or morphological analysis of a multilingual document, it is necessary to trigger language specific tools to avoid errors. Just one tool needs to be triggered for further processing of a homogeneous document, whereas for a heterogeneous document the same kind of tool has to be triggered based on the word level.

The LD toolkit is used as a web service component of the *Lexicon Expander* framework. Currently, the toolkit is able to detect 70 languages. It calculates the distances between

the input text and each of these 70 models. It returns the model name that minimizes distance. The input can be a document, a sentence or a word. The user can restrict the number of models to be used to detect an input. For example, in the case of a multilingual document, which contains sentences from OHG, Latin and Old Saxon languages, the toolkit can be restricted to these three models. In the case of ambiguity, the LD toolkit returns a list of languages with scores that can be used by the *Lexicon Expander* for further processing to assign the correct target language of the input.

4.4 The Lexicon Expander

The LD toolkit can be applied to a word, a sentence or a whole document. It yields high classification results, when its input is a document or a sentence (Islam et al., 2011). By contrast, if the input of the language detector are single words, *f-score* and *accuracy* results are low and not reliable. The output of the LD Toolkit for word input cannot be helpful for the annotator due to the numerous erroneous assignments. Nevertheless, for building a full-formed lexicon, we need to solve the problem of language detection for single unknown words.

In this section, we introduce a model, that processes the output of the LD toolkit and re-assigns a language to each unknown word, reaching higher *f-scores* and *accuracies*. The idea behind this model is that the target word's language is likely to be the same as the language of the sentence, in which this word occurs. In this way, many unknown words in the PL corpus come from editors' comments, containing several sentences in Italian or French. Not only languages assigned to the sentences, in which the word occurs, but also languages assigned to the co-occurring unknown words are important to detect the language of the target word.

The formal model of the calculation looks as follows. Let W be a set of word forms that occur in texts of the corpus C and let $W' \subseteq W$ be the subset of word forms whose language membership is unknown. Further, let S be the multiset of all sentences that occur in texts $x \in C$. Suppose that $f{:}S \rightarrow \mathrm{Pot}(W)$ is a function that maps each sentence $s \in S$ onto the multiset of the word forms that occur in it. The set of all sentences in which any given word form $w \in W$ occurs is defined as

$$S(w) = \{s \in S \mid w \in f(s)\} \tag{3}$$

We proceed by defining the language detection function L_1 that assigns to each word form (known or unknown) its language:

$$L_1{:}\, W \rightarrow \{l_1, \ldots, l_m\} = \mathbb{L} \tag{4}$$

$\mathbb{L}$ is the set of target languages to be detected. Note that we assume m target languages. In our experiments in Section 5, $|\mathbb{L}| = 2$. Note further, that we implement L_1 by means of the LD Toolkit (Section 4.3).

Now we are in a position to make the selection of the target language by means of our lexicon expander model: for any unknown word w it is defined as the language that is

```
Data: The set of unknown words W' = {w_1, ..., w_n}
Result: The language L(w) of any word w_i ∈ W'
for i = 1..n do
    L(w_i) ← {l ∈ L | ∃s ∈ S(w_i) : l = L_1(s)};
    if |L(w_i)| = 1 then
        L(w) ← L_1(w_i);
    end
    else
        L(w) ← L_2(w_i);
    end
end
```

Algorithm 1: Lexical Expander language assignment algorithm

most frequently assigned to those unknown words with which w co-occurs in sentences out of $S(w)$. Formally, we define the language detector function $L_2{:}\, W \rightarrow \mathbb{L}$ as follows:

$$L_2(w) = \arg\max_{l \in \mathbb{L}} \{|\{w' \in W' \mid \exists s \in S(w){:}\, w' \in f(s) \wedge \ L_1(w') = l\}|\} \tag{5}$$

Note that $L_2(w)$ is based on L_1 for guessing the language of unknown words.

In order to enlarge our *tertium comparationis*, we consider a third language detector function $L_3{:}W \rightarrow \mathbb{L}$. It assigns that language l to an unknown word $w \in W$ that is most frequently assigned to the sentences in which w occurs:

$$L_3(w) = \arg\max_{l \in \mathbb{L}} \{|\{s \in S(w) \mid L_1(s) = l\}|\} \tag{6}$$

Note that L_3 is also based on L_1, but is now applied to whole sentences (as input strings) instead of single words.

Algorithm 1 summarizes the choice between L_1 and L_2 (or L_3).

We applied this algorithm to several bilingual corpora (Section 5). Each time the assignment was done in three steps. The first step was to assign a language to an unknown word by the LD Toolkit. The second step was to assign a language to all the sentences in which the unknown word appears and find the most frequently assigned language. The third step was to assign the language to all the unknown words which co-occur with the target unknown word. Finally, when all the assignment steps successfully passed, we applied our decision algorithm that makes the final assignment.

5 Evaluation

The evaluation was run on four test corpora of various size and topics. (Table 1). Three test sets contain bilingual texts and the fourth test set is multilingual. For bilingual test sets, we used three language pairs: English and Turkish, German and French, OHG and Latin. The first test set is comprised of English Wikipedia articles (e.g. Atatürk, Istanbul etc.), which contained numerous Turkish words. The second one was a collection of German Wikipedia articles, containing French words. The third test set was composed of OHG sentences, containing Latin words. Sentences were manually

Language	Tokens	Sentences	Unknown
German - French	5893	315	460
English - Turkish	14022	724	438
OHG - Latin	1397	217	499
Multiling. German Text	1547	177	344

Table 1: Test corpora

extracted from the OHG corpus[2]. The gold standard includes all the tokens, found in the texts, which are manually annotated according to their language.

For the multilingual test set we used a text by Austrian author Hugo von Hofmannsthal. The basis of this text is an excerpt from his essay *"On the physiology of modern love"* (Hofmannsthal, 1891). It contains long French passages which are mainly quotes of contemporary French authors. Furthermore, Hofmannsthal comments on single sentences and constituents in German. Into this text English translations from the German original have been inserted by the authors in much the same manner, as sentences or constituents. Somewhere in between a ficticious commentary has been added. Additionally, very few sequences, all shorter than three words are Latin. So, except for German, English, French and Latin are also found in the text. In all test sets, all the tokens were manually annotated.

Language	F-Score	Accuracy
German - French	0.40	35.43%
English - Turkish	0.36	38.13%
OHG - Latin	0.79	70.34%
Multiling. German Text	0.37	41.2%

(a) Evaluation of LD toolkit: word level

Language	F-Score	Accuracy
German - French	0.49	49.13%
English - Turkish	0.5	52.51%
OHG - Latin	0.94	96.12%
Multiling. German Text	0.73	74.25%

(b) Evaluation of LD toolkit: sentence level

Language	F-Score	Accuracy
German - French	0.58	53.5%
English - Turkish	0.52	51%
OHG - Latin	0.95	91.78%
Multiling. German Text	0.73	72.96%

(c) Evaluation of the Lexicon Expander

Table 2: Evaluation Results

Table 2a shows the performance of the LD Toolkit for the word-by-word input. The LD toolkit yielded the highest results for the OHG/Latin test corpora and the lowest for the English-Turkish test set among bilingual test sets. The LD toolkit performed poorly on the multilingual test set. Table 2b presents the results of the assignment of the most prominent sentence language to the word. In other words, we calculated the most frequently assigned language of the sentences, where the word occurs, and re-assign this language to the word. *Accuracy* for the OHG text is higher than the one our model reached, but *f-score* is lower in all the cases and for the German text the

[2]TITUS, http://titus.uni-frankfurt.de.

difference in *f-score* between our model and sentence-wise detection is quite big. Table 2c shows results that were reached after the LD Toolkit output was postprocessed by the Lexicon Expander model. For all the test corpora, this model reached higher *f-score* and *accuracy* values than the LD toolkit. The best result is achieved for the OHG text. The average improvement for the bilingual test sets achieved by the postprocessing of the LD toolkit output is around 19% of *accuracy*. As for the multilingual test set, the improvement achieved by the model is even larger. The *f-score* is almost twice as big as the *f-score* of the LD Toolkit and *accuracy* grew for approximately 32%.

6 Discussion and Conclusion

We have shown that the Lexicon Expander as a postprocessing tool improved the performance of language assignment, based on the LD Toolkit, for each of the chosen language pairs. Herein, the Lexicon Expander showed consistent improvement of the *f-score* and *accuracy* values irrespective of the initial performance of the LD Toolkit. Though the language pairs for evaluation were heterogeneous in terms of language relationships (differences of lexical overlap) and text genres, this did not affect the performance of the Lexicon Expander. Further the size of the test corpora and the proportion of unknown words did not influence the improvement. We consider this a hint to the potential of the model in dealing with multilingualism in ancient corpora not only on sentential but also on the lexical level. The improvement of the *f-score* and *accuracy* values suggests that the Lexicon Expander is a first step in developing a functioning toolkit for multi-facetted language detection on various levels (e.g. lexical and sentential levels) and can be of help in saving annotator effort and in preprocessing ancient corpora.

References

Balk, H. (2010). IMPACT annual report 2009, version 1.0. http://www.impact-project.eu.

Bamman, D. and Crane, G. (2009). Structured knowledge for low-resource languages: The Latin and Ancient Greek dependency treebanks. In *Proc. of the Text Mining Services 2009, Leipzig, Germany*. Springer.

Bamman, D., Passarotti, M., Busa, R., and Crane, G. (2008). The annotation guidelines of the Latin Dependency Treebank and Index Thomisticus Treebank. In *Proceedings of LREC 2008*, Marrakech, Morocco. ELRA.

Büchler, M., Heyer, G., and Gründer, S. (2008). eAQUA – bringing modern text mining approaches to two thousand years old ancient texts. In *Proceedings of e-Humanities – An Emerging Discipline, workshop at the 4th IEEE International Conference on e-Science*.

Cavnar, W. B. and Trenkle, J. M. (1994). N-gram-based text categorization. In *In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175.

Erjavec, T. (2011). Automatic linguistic annotation of historical language: Totrtale and xix century slovene. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 33–38, Portland, OR, USA. Association for Computational Linguistics.

Erjavec, T., Ringlstetter, C., Zorga, M., and Gotscharek, A. (2011). A lexicon for processing archaic language: the case of XIXth century Slovene. In *Proceedings of the International Workshop on Lexical Resources at ESSLLI*.

Gippert, J. (2001). TITUS — Alte und neue Perspektiven eines indogermanistischen Thesaurus. *Studia Iranica, Mesopotamica et Anatolica*, 2:46–76.

Gippert, J. (2006). *Essentials of Language Documentation*, chapter Linguistic documentation and the encoding of textual materials, pages 337–361. Mouton de Gruyter, Berlin.

Gippert, J. (2010a). Manuscript related data in the TITUS project. *ComSt Newsletter*, 1:7–8.

Gippert, J. (2010b). New prospects in the study of old georgian palimpsests. In *Proceedings of the 1st International Symposium on Georgian Manuscripts, October 19–25, 2009, Tbilisi*.

Gleim, R., Mehler, A., Waltinger, U., and Menke, P. (2009a). eHumanities Desktop — an extensible online system for corpus management and analysis. In *5th Corpus Linguistics Conference, University of Liverpool*.

Gleim, R., Waltinger, U., Ernst, A., Mehler, A., Esch, D., and Feith, T. (2009b). The eHumanities Desktop — an online system for corpus management and analysis in support of computing in the humanities. In *Proceedings of the Demonstrations Session of the 12th Conference of the European Chapter of the Association for Computational Linguistics EACL 2009, 30 March – 3 April, Athens*.

Gotscharek, A., Neumann, A., Reffle, U., Ringlstetter, C., and Schulz, K. U. (2009). Enabling information retrieval on historical document collections: the role of matching procedures and special lexica. In *AND '09 Proceedings of The Third Workshop on Analytics for Noisy Unstructured Text Data*.

Gotscharek, A., Reffle, U., Ringlstetter, C., Schulz, K. U., and Neumann, A. (2011). Towards information retrieval on historical document collections: the role of matching procedures and special lexica. *IJDAR*, 14(2):159–171.

Hofmannsthal, H. v. (1891). Zur physiologie der modernen liebe. *Die Moderne*.

Hulden, M. (2009). FOMA: a finite-state compiler and library. In *Proceedings of the Demonstrations Session at EACL 2009*, pages 29–32, Athens, Greece. Association for Computational Linguistics.

Islam, Z., Mittmann, R., and Mehler, A. (2011). Multilingualism in ancient texts:language detection by example of old high german and old saxon. In *GSCL conference on Multilingual Resources and Multilingual Applications (GSCL 2011), Hamburg, Germany*.

Koster, C. H. A. (2005). Constructing a parser for Latin. In Gelbukh, A. F., editor, *Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2005)*, volume 3406 of *Lecture Notes in Computer Science*, pages 48–59. Springer.

Mehler, A., Diewald, N., Waltinger, U., Gleim, R., Esch, D., Job, B., Küchelmann, T., Pustylnikov, O., and Blanchard, P. (2011a). Evolution of Romance language in written communication: Network analysis of late Latin and early Romance corpora. *Leonardo*, 44(3).

Mehler, A., Schwandt, S., Gleim, R., and Jussen, B. (2011b). Der eHumanities Desktop als Werkzeug in der historischen Semantik: Funktionsspektrum und Einsatzszenarien. *Journal for Language Technology and Computational Linguistics (JLCL)*. Accepted.

Migne, J.-P., editor (1844–1855). *Patrologiae cursus completus: Series latina*, volume 1–221. Chadwyck-Healey, Cambridge.

Passarotti, M. (2000). Development and perspectives of the latin morphological analyser LEMLAT (1). *Linguistica Computazionale*, 3:397–414.

Passarotti, M. (2010). Leaving behind the less-resourced status. the case of latin through the experience of the index thomisticus treebank. In *Proceedings of the 7th SaLTMiL Workshop on the creation and use of basic lexical resources for less-resourced languages, LREC 2010, La Valletta, Malta*, Malta. ELDA.

Smith, D. A., Rydberg-Co, J. A., and Crane, G. R. (2000). The Perseus Project: A digital library for the humanities. *Literary and Linguistic Computing*, 15(1):15–25.

Sperberg-McQueen, C. and Burnard, L. (2009). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. TEI Consortium.

Waltinger, U. (2010). *On Social Semantics in Information Retrieval*. Phd thesis, Bielefeld University, Germany.

Waltinger, U. and Mehler, A. (2009). The feature difference coefficient: Classification by means of feature distributions. In *Proceedings of the Conference on Text Mining Services (TMS 2009)*, Leipziger Beiträge zur Informatik: Band XIV, pages 159–168. Leipzig University, Leipzig.

Author Index

Eduard Barbu
Universita' di Trento
eduard.barbu@unitn.it

Kristin Bech
University of Oslo
kristin.bech@ilos.uio.no

Francesca Bonin
Universita' di Trento
francesca.bonin@gmail.com

Federico Boschetti
University of Pavia - CNR of Pisa
federico.boschetti@yahoo.com

Fabio Cavulli
Universita' di Trento
fabio.cavulli@lett.unitn.it

Gregory Crane
Tufts University
gregory.crane@tufts.edu

Stefanie Dipper
Bochum University
dipper@linguistics.rub.de

Kristine Eide
University of Oslo
k.g.eide@ilos.uio.no

Asif Ekbal
IIT Patna
asif.ekbal@gmail.com

Voula Giouli
Institute for Language and Speech Processing, R.C. "Athena"
voula@ilsp.gr

Christian Girardi
FBK
cgirardi@fbk.eu

Iris Hendrickx
Universidade de Lisboa
iris@clul.ul.pt
Stefan Höfler

University of Zurich
hoefler@cl.uzh.ch

Armin Hoenen
Johann Wolfgang Goethe-Universität Frankfurt
hoenen@em.uni-frankfurt.de

Ryu Iida
Tokyo Institute of Technology
ryu-i@cl.cs.titech.ac.jp

Anton Karl Ingason
University of Pennsylvania
anton.karl.ingason@gmail.com

Zahurul Islam
Johann Wolfgang Goethe-Universität Frankfurt
zahurul@em.uni-frankfurt.de

Britta Juska-Bacher
University of Basel
britta.juska-bacher@unibas.ch

Dain Kaplan
Tokyo Institute of Technology
dain@cl.cs.titech.ac.jp

Timo Korkiakangas
University of Helsinki
timo.korkiakangas@helsinki.fi

Cerstin Mahlow
University of Basel
cerstin.mahlow@unibas.ch

Francesco Mambrini
Università Cattolica del Sacro Cuore, Milan
f.mambrini@gmail.com

Massimo Manca
University of Venice
massimo.manca@unive.it

Rita Marquilhas
Universidade de Lisboa
rita.marquilhas@gmail.com

Paolo Mastandrea
University of Venice
mast@unive.it

Alexander Mehler
Johann Wolfgang Goethe-Universität Frankfurt
mehler@em.uni-frankfurt.de

Filippo Nardelli
Expert System / Cogito
fnardelli@expertsystem.it

Kikuko Nishina
Tokyo Institute of Technology
nishina.k.aa@m.titech.ac.jp

Marco Passarotti
Università Cattolica del Sacro Cuore, Milan
marco.passarotti@unicatt.it

Massimo Poesio
University of Essex and Universita' di Trento
poesio@essex.ac.uk

Michael Piotrowski
Law Sources Foundation of the Swiss Lawyers Society
mxp@ssrq-sds-fds.ch

Eiríkur Rögnvaldsson
University of Iceland
eirikur@hi.is

Sriparna Saha
IIT Patna
sriparna.saha@gmail.com

Einar Freyr Sigurðsson
University of Iceland
einarfs@gmail.com

Arne Skjærholt
University of Oslo
arnsholt@gmail.com

Linda Spinazzè
University of Venice
linda.spinazze@gmail.com

Caroline Sporleder
Saarland University
csporled@coli.uni-sb.de

Egon Stemle
Universita' di Trento
Egon.Stemle@cimec.unitn.it

Maria Sukhareva
Johann Wolfgang Goethe-Universität Frankfurt
sukhareva@em.uni-frankfurt.de

Luigi Tessarolo
University of Venice
luigi.tessarolo@fastwebnet.it

Takenobu Tokunaga
Tokyo Institute of Technology
take@cl.cs.titech.ac.jp

Joel C. Wallenberg
Newcastle University
joel.wallenberg@gmail.com