Language Resources and Technologies for Processing and Linking Historical Documents and Archives- Deploying Linked Open Data in Cultural Heritage – LRT4HDA

Workshop Programme

09:00 - 09:30 – Opening and introduction by Workshop Chairs

09:30- 10:30 – Invited Talk
Eiríkur Rögnvaldsson, *Old languages, new technologies: The case of Icelandic*

10:30 – 11:00 Coffee break

11:00 – 11:30 – Session Language Resources
Patrick Schone, *A Personal Name Treebank and Name Parser to Support Searching and Matching of People Names in Historical and Multilingual Contexts*

11:30 – 12:00 – Session Language Resources
Elaine Uí Dhonnchadha, Kevin Scannell, Ruairí Ó hUiginn, Eilís Ní Mhearraí, Máire Nic Mhaoláin, Brian Ó Raghallaigh, Gregory Toner, Séamus Mac Mathúna, Déirdre D'Auria, Eithne Ní Ghallchobhair and Niall O'Leary, *Corpas na Gaeilge (1882-1926): Integrating Historical and Modern Irish Text*

12:00 – 12:30 – Session Language Resources
Ásta Svavarsdóttir, Sigrún Helgadóttir and Guðrún Kvaran, *Language resources for early Modern Icelandic*

12:30 – 12:50 – Session Language Resources
Dominique Ritze, Caecilia Zirn, Colin Greenstreet, Kai Eckert and Simone Paolo Ponzetto, *Named Entities in Court: The MarineLives Corpus*

12:50 – 14:00 Lunch break

14:00 – 14:20 – Session Language Resources
Stephen Tyndall, *Building Less Fragmentary Cuneiform Corpora: Challenges and Steps Toward Solutions*

14:20 – 14:40 – Session Language Resources
Thorhallur Eythorsson, Bjarki Karlsson and Sigríður Sæunn Sigurðardóttir, *Greinir skáldskapar: A diachronic corpus of Icelandic poetic texts*

14:40 – 15:00 – Session Historical Newspaper Archives
Oliver Pfefferkorn and Peter Fankhauser, *On the Role of Historical Newspapers in Disseminating Foreign Words in German*

15:00 – 15:20 – Session Historical Newspaper Archives
Örn Hrafnkelsson and Jökull Sævarsson, *Digital libraries of historical Icelandic newspapers, periodicals, magazines and old printed books*

15:20 – 15:50 – Session Historical Newspaper Archives
Susanne Haaf and Matthias Schulz, *Historical Newspapers & Journals for the DTA*

16:00 – 16:30 Coffee break

16:30 – 17:00 – Session Tools for analysis of historical documents
Jón Friðrik Daðason, Kristín Bjarnadóttir and Kristján Rúnarsson, *The Journal Fjölnir for Everyone: The Post-Processing of Historical OCR Texts*

17:00 – 17:20 – Session Tools for analysis of historical documents
Ludger Zeevaert, *IceTagging the "Golden Codex". Using language tools developed for Modern Icelandic on a corpus of Old Norse manuscripts*

17:20 – 17:40 – Session Tools for analysis of historical documents
Federico Boschetti, Andrea Cimino, Felice Dell'Orletta, Gianluca Lebani, Lucia Passaro, Paolo Picchi, Giulia Venturi, Simonetta Montemagni and Alessandro Lenci, *Computational Analysis of Historical Documents: An Application to Italian War Bulletins in World War I and II*

17:40 – 18:00 – Session Tools for analysis of historical documents
Cristina Vertan, Walther v. Hahn, *Discovering and Explaining Knowledge in Multilingual Historical Documents*

18:00 – 18:30 Discussions and Closing

Editors

Kristín Bjarnadóttir	The Arni Magnusson Institute for Icelandic Studies, Iceland
Mathew Driscoll	Arnamagnean Commission, Copenhagen, Denmark
Steven Krauwer	CLARIN ERIC, Netherlands
Stelios Piperidis	ILSP, Athens, Greece
Cristina Vertan	University of Hamburg
Martin Wynne	University of Oxford, UK

Workshop Organizers/Organizing Committee

Kristín Bjarnadóttir	The Arni Magnusson Institute for Icelandic Studies, Iceland
Mathew Driscoll	Arnamagnean Commission, Copenhagen, Denmark
Steven Krauwer	CLARIN ERIC, Netherlands
Stelios Piperidis	ILSP, Athens, Greece
Cristina Vertan	University of Hamburg
Martin Wynne	University of Oxford, UK

Workshop Programme Committee

Lars Borin	University of Gothenburg, Sweden
Rafael Carrasco	University of Alicante, Spain
Paul Doorenbosch	National Library of the Netherlands, Netherlands
Þórhallur Eyþórsson	University of Iceland, Iceland
Alexander Geyken	BBAW, Germany
Günther Görz	University Erlangen, Germany
Walther v. Hahn	University of Hamburg, Germany
Erhard Hinrichs	University of Tuebingen, Germany
Guillaume Jacquet	JRC, Italy
Marc Kupietz	IDS, Germany
Éric Laporte	Université Paris-Est Marne-la-Vallée, France
Piroska Lendvai	Hungarian Academy of Sciences, Hungary
Thierry Paquet	LITIS, France
Gábor Prószéky	MorphoLogic, Hungary
Bente Maegaard	University of Copenhagen, Denmark
Christian Emil Ore	University of Oslo, Norway
Eiríkur Rögnvaldsson	University of Iceland, Iceland
Petya Osenova	IICT, Bulgarian Academy of Sciences, Bulgaria
Manfred Thaller	Cologne University, Germany
Tamás Váradi	Hungarian Academy of Sciences, Hungary
Matthew Whelpton	University of Iceland, Iceland
Kalliopi Zervanou	University of Tilburg, the Netherlands

Table of contents

Author Index

Foreword

Recently, the collaboration between the NLP community and the specialists in various areas of the Humanities has become more efficient and fruitful due to the common aim of exploring and preserving cultural heritage data. It is worth mentioning the efforts made during the digitisation campaigns in the last years and within a series of initiatives in the Digital Humanities, especially in making Old Manuscripts available through Digital Libraries.

Having in mind the number of contemporary languages and their historical variants, it is practically impossible to develop brand new language resources and tools for processing older texts. Therefore, the real challenge is to adapt existing language resources and tools, as well as to provide (where necessary) training material in the form of corpora or lexicons for a certain period of time in history.

Another issue regarding historical documents is their usage after they are stored in digital libraries. Historical documents are not only browsed but together with adequate tools they may serve as basis for re-interpretation of historical facts, discovery of new connections, causal relations between events etc. In order to be able to make such analysis, historical documents should be linked among themselves, on the one hand, and with modern knowledge bases, on the other. Activities in the area of Linked Open Data (LOD) play a major role in this respect.

A particular type of historical documents are the newspaper collections and archives. Newspapers reflect what is going on in society, and constitute a rich data collection for many types of humanities research, ranging from history, political and social sciences to linguistics, both synchronic and diachronic, and both national and cross-national. They represent an important resource for analysis of changes at all levels which emerged in Europe with begin of the industrialization period.

Most digital libraries are made available not only to researchers in a certain Humanities domain (e.g. classical philologists, historians, historical linguists), but also to common users. This fact has posited new requirements to the functionalities offered by the Digital Libraries, and thus imposed the usage of methods from LT for content analysis and content presentation in a form understandable to the end user.

There are several challenges related to the above mentioned issues:

- Lack of adequate training material for real-size applications: although the Digital Libraries usually cover a large number of documents, it is difficult to collect a statistically significant corpus for a period of time in which the language remained unchanged.
- Historical variants of languages lack firmly established syntactic or morphological structures thus the definition of a robust set of rules is very difficult. Historical texts often constitute a mixture of multilingual paragraphs including Latin, Ancient Greek, Slavonic, etc.
- Historical texts contain a large number of anon-standardized abbreviations.
- The conception of the world is somewhat different from ours, which makes it more difficult to build the necessary knowledge bases.

For newspaper collections there are specific questions related to different stages of the whole cycle starting from acquisition of the digital data, conducting the research until publication of the final research results like use of incomplete OCR, available selection of digital newspapers, or copyright limited access to digital newspapers. Other relevant issues include: recommended standards and tools for structural (and, perhaps, lexical) annotation; methods for quality control of every step in the process (i.e., digitisation, transcription, annotation, tools, web services, etc); specific tools for layout analysis, for enabling access to the inline images/figures, etc.
This workshop brings together researchers working in the interdisciplinary domain of cultural

heritage, specialists in natural language and speech processing working with less-resourced languages as well as key players among Linked Open Data initiatives. They are expected to analyse problems and brainstorm solutions in the automatic analysis of historical documents, uni- or multimedia, their deep annotation and interlinking. The workshop builds on successful previous initiatives in this domain at LREC 2010, 2012, and RANLP 2011.

We received a considerable number of papers from which we selected 13, grouped in three sections, centered around following topics:

1. Language Resources for historical documents
2. Historical Newspaper Archives
3. Tools for analysis of historical documents

We are particulary grateful to our invited speaker, Eiríkur Rögnvaldsson, who will present the case of Old Icelandic as historical language and how modern technologies can be used to process it.

We would like to thank all members of the Programme Committee who reviewed in very short time a large number of papers and gave very useful feedback

The workshop is endorsed by the CLARIN Infrastructure Project http://www.clarin.eu

Kristín Bjarnadóttir, Mathew Driscoll, Steven Krauwer, Stelios Piperidis, Cristina Vertan,
Martin Wynne

Invited Talk:

Old languages, new technologies: The case of Icelandic

Eiríkur Rögnvaldsson
Department of Icelandic, University of Iceland
Árnagarði við Suðurgötu, IS-101, Reykjavík, Iceland
E-mail: eirikur@hi.is

Abstract

In the past few years, interest in developing language technology tools and resources for historical languages or older stages of modern languages has grown considerably, and a number of experiments in adapting existing language technology tools and resources to older variants of the languages in question have been made. The reasons for this increased interest can vary. One is that more and more historical texts are becoming digitized and thus amenable to language technology work. As a result, researchers from many disciplines are starting to realize that they could benefit from being able to search these texts and analyze them with all sorts of language technology tools. Another reason is that since historical texts often exhibit considerable variation in spelling and morphology, they pose great challenges to existing language technology tools and methods developed for modern standardized texts. Thus, many language technology researchers see historical texts as a good test bed for developing and enhancing their methods and tools.

In many ways, Icelandic is well suited for being such a test bed. Icelandic has a relatively large corpus of texts from all stages in its recorded history, starting with a number of narrative texts from the 13th and 14th centuries such as the well-known Family Sagas and the so-called Contemporary Sagas. Most importantly, however, Icelandic has changed less during the last thousand years than most or all other languages with a recorded history. True, the sound system has changed radically, especially the vowel system, but these changes are for the most part not reflected in the spelling. There are a number of changes in the syntax, especially as regards word order, but importantly for language technology work, morphological changes are minimal – the Modern Icelandic inflectional system is almost identical to the Old Icelandic system. Of course, the vocabulary has changed considerably, but since the changes are mainly due to new words being added rather than to old words becoming obsolete, these changes do not pose problems for the adaptation of language technology tools to older stages of the language.

Like most other historical languages, older stages of Icelandic show great variation in spelling, even though it may be mentioned that in the 12th century, shortly after Icelanders started writing in Latin letters, an unknown person usually referred to as the First Grammarian made an attempt to standardize Icelandic spelling in a famous essay called the First Grammatical Treatise. This attempt was not successful, and up to the beginning of the 19th century, everyone used his or her own spelling, usually reflecting a mixture of their own pronunciation and the spelling of the manuscripts they had been exposed to, accompanied by considerable intra-scribal variation. With the advent of periodicals around 1800, and especially after the advent of weekly newspapers around 1850, the spelling gradually became more and more uniform and around 1900, a commonly agreed standard had emerged.

Since serious work on Icelandic language technology started some 12 years ago, several important resources and tools for Modern Icelandic have been built, such as the Database of Modern Icelandic Inflections, the Tagged Icelandic Corpus, and the IceNLP package including a POS tagger, a shallow parser and a lemmatizer, to name the most important ones. As a result of the META-NORD project, most of the existing tools and resources are now open and free for everyone to use under standard licenses (GNU and Creative Commons).

In my talk, I will give an overview of the experiments that have been made in adapting and developing language technology tools for older stages of Icelandic. This includes the building of a parsed historical corpus (IcePaHC) spanning almost ten centuries; adapting POS taggers developed for Modern Icelandic to tagging Old Icelandic texts; developing tools to correct and normalize OCR scanned text of 19th century periodicals and newspapers; and using the IceNLP package in preparing electronic editions of Old Icelandic texts. These experiments have approached similar problems in different ways and I will compare their methods and assess their success in solving the problems.

Section 1:
Language Resources

A Personal Name Treebank and Name Parser to Support Searching and Matching of People Names in Historical and Multilingual Contexts

Patrick J. Schone
FamilySearch
50 E North Temple St. Salt Lake City, UT 84150
E-mail: boisebound@aol.com

Abstract

Personal names are often key elements desired from the processing of historical documents. An understanding of name syntax can be very valuable in aiding automation and analysis. Yet current grammatical parsers classify personal names as merely noun phrases with proper nouns as constituents. To that end, we have created a *Personal Name Treebank* (PNTB) and associated *Statistical Personal Name Parser* (SPNP) which are designed to carefully analyze syntactic structure of personal names. The PNTB consists of almost 10 million instances of constituency-parsed personal names attached to genealogically-related contexts. These instances are drawn from almost 200 different countries across millenia. The SPNP leverages the PNTB to achieve 94.4% parse constituency accuracy on a huge held-out set of names. The PNTB and SPNP represent significant new resources which we intend to make available to the research community. To our knowledge, comparable resources have never previously been created.

Keywords: Parsing, Treebank, Onomastics

1. Introduction

Many applications in industry and government require careful analyses of personal names. These applications have led to the emergence of various research areas of NLP which require in-depth understanding of personal names and of worldwide naming conventions. Examples of such areas (with a far from comprehensive list of references) include *person search* (Weerkamp, et al., 2011), *automatic knowledge-base population* (McNamee, et al., 2010), *person name disambiguation* (Hirschman and Chinchor, 1997; Cucerzan, 2007; Liu, et al., 2011), and *name matching* (Freeman, et al., 2006; MITRE, 2011; Schone, et al., 2012). These techniques sometimes need to be applied from one language into another, but the focus of current efforts has largely been toward analysis of modern names.

In recent years, there has been a surge in genealogical interest, where there is a critical need to perform matching and search on names which extend to many languages across historical contexts. FamilySearch, referred to by Wikipedia ("FamilySearch," 2014) as "the largest genealogy organization in the world," has hundreds of thousands of patrons who are seeking their ancestors. These often-amateur genealogists typically need to search and analyze personal names from billions of historical records and pedigrees which span many centuries and linguistic boundaries. Optimal processing of names in these contexts requires general-purpose name parsers and analysis tools which function well -- regardless of time period, language, or culture.

Creation of such tools is a clear challenge given that personal names have rich syntactic structure. Moreover, currently-available syntactic parsers do not delve into the structure of personal names. Instead, when applied to names, parsers tend to assign names to noun phrases with proper noun constituents. For example, consider the name "Maria del Carmen Hernandez Rodriguez." A reader familiar with Spanish might expect that "Maria del Carmen" is a multipart given name phrase; and "Hernandez" and "Rodriguez" represent the father's and mother's family lines, respectively. Canonical parsers (using the Cognitive Computation Group's "Curator" (2012) Demo) applied to this name yield:

Stanford Parser (Klein and Manning, 2003):
[NP [NNP Maria] [NNP del] [NNP Carmen]
[NNP Hernandez] [NNP Rodriguez]]

Charniak Parser (Charniak, 2000):
[S1 [NP [NNP Maria] [FW del] [NNP Carmen]
[NNP Hernandez] [NNP Rodriguez]]]

Such parses are adequate for generic text processing, but they are likely insufficient for in-depth name analysis.

For this reason, we have developed tools and resources whose goal is to provide deep syntactic understanding of personal names – especially across cultures, languages, and time frames. These include: (1) an extensive Personal Name Treebank (PNTB) which serves as training material; and (2) a statistical personal name parser (SPNP) which leverages the PNTB to provide broad coverage automatic analysis of personal names.

In the PNTB, the name above, whose constituent tags will be explained in this paper, would be parsed based on its genealogical-syntactic structure as

```
[NAME [GNP [GNP [G1 Maria]]
      [PPGN [INDT1 del] [GNP [NNPR1 Carmen]]]]
      [SNP  [FNPF   [FNF1 Hernandez]]
            [FNPM [FNM1 Rodriguez]]]]
```

The PNTB contains specific parse information on personal names from almost 200 countries (including some which are now defunct) and across over 20 centuries. This PNTB attempts to identify the origin of family; it distinguishes patronymics from family names; handles segmentation; and addresses other personal name issues which are referenced later.

Since most treebanks are drawn from a corpus which could potentially be used for context, we likewise supply context for our personal name parses. This context may include a date and place of the individual's life, the individual's gender, and names of close relatives like the father or spouse. These data are mined from historical records and genealogies, so we also specify the record type from which they are drawn (e.g., "BIRTH," "CENSUS", etc). The PNTB contains 10.0 million instances of name-plus-context of which 2.67M are unique parses. The PNTB represents a significant, novel resource to language processing.

To us, one of the PNTB's principal values is to serve as training material for our statistical personal name parser, *SPNP*. We have separated the PNTB into a train and test set and developed a multi-method, feature-based statistical parser which achieves a constituency accuracy of **94.4%** when tested on the linguistically and temporally diverse test set of 301K names (157K unique parses) which were never observed in the training data.

It is our intention of making the PNTB and SPNP available. Our hope is that they can have widespread value, and that others can perhaps extend them to new languages or time frames that we have not covered adequately. We therefore provide this description of these interesting resources.

2. Background on Treebanks and Names

Treebanks appeared in the 1990s (Marcus, et al. (1993); Garside and McEnery (1993)). A treebank is a collection of linguistic parses. In the case of the earliest treebanks, these collections involved constituency parses of well-formatted English text. Over the past 20 years, development of treebanks has extended to many different languages (including German (Brants, et al. 1999), Chinese (Xue, et al., 2005), Korean (Han, et al., 2002), and others; to multiple genre of data such as transcripts of speech (Taylor, 1996); and to other parse constructs including dependency parses (such as the Danish dependency treebank (Kromann, 2003)). To our knowledge, there has never been a treebank which focuses exclusively on personal names (though there are some rulesets available such as for handling names in bibliographies (BibTex, 2006)).

We have interest in searching personal names in genealogical settings that span centuries and that cross country and linguistic boundaries. To provide an optimal search experience, our algorithms must be aware of name syntax across these very diverse settings. Thus, the creation of a personal name treebank seems warranted.

The syntax of names is non-trivial when considered from a global context. People may think of names as sequences of *n* tokens where the first (*n*-1) tokens refer to given names and the final token is a "surname". Even if this were always true, our desire to automatically segment the names into their proper tokens and identify the origin of family name(s) would make even this simple view of names more complicated. Yet as Table 1 shows, name structures that vary from this simple intuition are ubiquitous (slashes in the table indicate "surnames").

Syntax Issue	Example 1	Example 2
Surname ordering	/*Зимин*/ Иван	/*SMITH*/ John
Multipart surnames	Ian /*Mc Carthy*/	Ely /*St. John*/
Multipart given names	*Maria de Jesus* /Hernandez/	*De Witt* /Jones/
Relational Indicators	*Mrs.* John /Brown/	*Daughter* of Earl of Kildare
Embedded titles	Erzsebet *Baroness* /Banffy/	*Al-hajji* Nasser /Sebaggala/
Whole-phrase names	*Crow Flies High*	*Kah mo to chi man*
Namesakes	*Martin van Buren* /Babcock/	*Benjamin Franklin* /Johnson/
Generational terms	Joseph /Smith/, *Junior*	Henry *VIII*
Patronymics	Petter /*Johan*sson/	Morgan /*William*/
Broken names	And. /*And.Sson*/	Drua /*Ols Datter*/
Embedded Weekdays	Quin-se-da *Monday* /Tawul/	*Wednesday* Udo-Udo Akpan /Ebong/
Non-Names	*Concubine* 1	*Blank Field* /*Blank Field*/
Conjunctions/ Disjunctions	Benj /Bez *Or* Bezer/	Diedr. /Pohl *Or* Borgmann/

Table 1: Examples of Atypical Name Forms

In addition to interesting name syntax and the difficulties of segmentation and determining family origin, some name parts have multiple different syntactic roles. In personal names, most personal name tokens can represent given and family names. Other words like "le," "van", "baron," "bishop", "mac", etc. can have multiple syntactic roles. Table 2 shows that "de" has at least six:

Syntactic Role of "de"	Examples
Preposition	Antonio /*De* la Cruz/
Determiner	Corn. /*De* Beer/
Spouse relation	Agustina /*de* Castro/
Particle	*De* Willis /Clark/
Family name	Chua /*De*/
Given name	*De* /Moss/

Table 2: Multiple name functions of "De"

Lastly, there are name syntax issues that have changed over time (such as patronymics becoming surnames) or as migration has occurred (such as surnames that previously reflected gender being reduced to a singular masculine form). An intelligent name syntax parser should be aware of these temporal and migratory syntax changes.

3. Identifying Syntactic Classes

Before describing the PNTB's construction, we need to define *personal name parts of speech* (*PNPOS*) that form the foundation of our parses. Then, we define the parse constructs that are used to wrap PNPOS into phrases.

3.1 PNPOSs

Parts of speech (POS), such as nouns and verbs, are common constructs of grammar. Wherever possible, we would like for the PNTB to take advantage of POSs as observed elsewhere. In particular, we use IN (preposition) and DT (determiner) frequently; and we use proper nouns (NNP), adjectives (JJ), and verbs and past tense verbs (VB, VBD) where appropriate.

PNPOS	Description	Examples
G	Given Name	*John,Maria*
GG	Generic nickname	*Buddy, Slim*
ABBRGN	Abbeviation	J*n.,Ma.*
ABBRI	Initialism	*J., M*
ABBRWC	Wild Card	*J*n, M…a*
G1a, G1b,..	Multipiece	*Ah Sing*
GNF	Father's Given Name	*Sven*-sson
GNM	Mother's Given	*Helga*-sson
GNPGF, GNPGM	Paternal grandfather/ grandmother's given	*Jon*-ssonar
FN,FNF, FNM, FNS	Family Name: Unattributed, of Father, of Mother, of Spouse	*Hernandez*
ABBRFN{F, M, S,}	Abbreviated family name of father, mother, spouse, other	*Hern.*
ABBRWCFN {F, M, S,}	Wild card family name	*H*z*
PARTRB	Right-bound particle	*Mc* Vey
REL{S,D,C, GC,GS,GD, W,A,U,V}	Relational particle: son, daughter, child, grandchild, grandson, granddaughter, wife, aunt, uncle, servant	*-sson,* *-sdottir* *-ssonar*
OCC	Occupation	*Soldier*
H, Hf	Honorifics	*Mr., Mrs.*
T	Title	*Captain*
ORDG	Generationals	*Jr., Sr.*
ORDR	Roman Ordering	*I, III, IV*
ORDM	Ordinal	*Third, 2nd*
M	Number	Two,Three
ORG	Organization indicator	*Spec. Forces*
NNPA, NNPN, NNPR	Proper nouns: animals or nature, or religious terms	*Crow* *Blanket,* dela*Trinidad*
NNPD	Demonyms	the *German*
NNPL	Location	of *Scotland*
WD	Weekday names	*Wednesday*
X	Non-name pieces	*Unnamed*
INDT	Preposition-Determiner	*Dos* Santos
PARTFEM	Feminizing particle	Первушин *-a*

Table 3: Added Personal Name Parts of Speech

However, in personal names, there are many interesting phenomena that one might like to recognize but which are not accounted for using typical POS. For examples, given names (like "John" or "Maria") and family names (like "Smith" or "Brown") are key components of names which have no distinguishing POS in the literature. Furthermore, since we would like to be able to determine family name origins (such as if the name derives through the father's family line, the mother's, the spouse's, etc), we need to be able to subdivide these major categories. Honorifics (like "Mrs." Or "Ui buin"), generational indicators (such as "Junior"), relation components ("So-and-so's *son*"); non-name components (such as "*Stillborn* Jones"), and others need to be accounted for.

Space limitations prevent us from describing all of the PNPOS in depth. Details for most of these labels are described elsewhere (see Schone and Davey, 2012). We therefore identify them in Table 3 in addition to providing brief descriptions and examples of each.

3.2 Wrapping PNPOSs into Parses

Identification of the PNPOS is only a first step for providing constituency parses. To complete a parse, we need to identify the phrase types that are used to group PNPOSs into ever-larger chunks.

Instead of a sentence, we structure personal names into "**NAME**" constructs. It is also possible for an overall NAME to include an embedded NAME, as in Mrs. *Mary Jones* or *Benjamin Franklin* Smith. Since surnames and given names are core pieces of many personal names, we use **SNP** and **GNP** to identify surnominal and given name phrases. We treat SNPs as any combination of family name phrases (which encapsulate family names and are marked as **FNPF, FNPM, FNPS,** and **FNP** for a father's, mother's, spouse's or unattributed family name phrase).

For patronymic phrases, we use **PNP** instead of SNP or GNP and we tag its sub-constructs. The patronymic Сергеевич (Sergeevich), for example, will be represented by the parsing fragment: [PNP [RELPC [GNPF [GNF1 Сергей]] [RELS +-й+евич]]]. This indicates that Сергей (Sergey) is the father, and that his name, when coupled with an attachment which drops the letter й and then adds евич, forms a relational phrase referring to the child (which would be roughly equal to "Sergey's son.")

Beyond these, there are additional phrasal constructs that are designed to address special issues that occur in names. For example, we accommodate option phrases (*John or Jon*); other kinds of relational phrases ("*Mrs. John*

Smith"); toponymic phrases ("*of England*") ; non-name phrases ("*unnamed child"*); titular phrase ("*her royal highness*"); attributional phrases ("*the forked beard"*); adverbial phrases ("*found dead"*); and within-name equations ("Eagle Horn *Hewanbliwin"*).

3.3 Chomsky Normalization

Since many statistical parsers are trained from Chomsky Normal Form, we incorporate additional parse structure to ensure that no more than two non-terminals are encapsulated in the same bracket. For instance, "John or Jon" involves three constituents, so we introduce a tag CCD_GNP around "or Jon" to limit constituents to a maximum of two. So a name like this will parsed as:

```
  [NAME [GNP [GNPO [GNP [G1 John]]
          [CCD_GNP [CCD1 or] [GNP [G2 Jon]]]]]].
```

These additional constructs are provided for convenience, but one should be able to deterministically remove them if non-binary form is desired.

4. Building the Personal Name Treebank

After having defined the PNTB tag set, we now describe our methodology for identifying and tagging the personal name examples. It should be commented first that all of the linguistic processing described in this section was performed by a single annotator with multiple-language processing ability. Therefore, there should be no sources of error in the PNTB that are due specifically to *interannotator* disagreement. On the other hand, the annotator could have made decisions that others might deem as incorrect for certain circumstances. Likewise, the entire annotation process consists of a mix of human labelling and automation, and the automated portion could be in error. It is therefore expected that the PNTB *training* data has 2-4% error; however, the test set has been repeatedly vetted, so the error there is expected to be on the order of 1% or less.

4.1 PNPOS Patterns

We started the process of converting personal names into parses by identifying interesting PNPOS structures. We initially mined an existing name matching corpus (Schone, et al 2012) and roughly categorized the PNPOS pattern structure of each name, identified new patterns, added additional rules, and repeated. For example, commonly-expected name patterns might be "G1 FN1", "G1 G2 FN1," and "G1 ABBRI1 FN1." Starting with these seed patterns, we would find that *Mr. John Smith* does not conform to these patterns. So we might add "H1 G1 FN1" and "H1 G1 G2 FN1," and repeat. After adding over 5000 patterns, we could categorize the majority of the names (without yet attempting to determine family name origin).

Next, we applied these rules to the unique names from the collection and grouped those with common PNPOS patterns. So "*Mary Smith*," "*John Brown*," and "*Fred Jones*" were grouped under G1 FN1; whereas names like "*S A Horrocks*" and "*P J Green*" were collected under ABBRI1 ABBRI2 FN1. Each name under a PNPOS-pattern grouping was reviewed by the annotator to determine if the predicted PNPOS-pattern was correct, incorrect, or interesting but wrong in association with the specific name. Over 700K names were evaluated by hand using this procedure.

4.2 Providing Context

We then sought to identify actual instances of each name for use as context. We distilled information from a multi-billion record collection of information derived from censuses, vital records, and genealogical pedigrees (records courtesy of FamilySearch, 2014). We retained at most 100 contexts for each unique name in training and only up to 5 contexts for testing. For example, one instance was:

Abigail /WILEY or WYLIE/; Goshen Twp., Sullivan, N.H.;F;(1791,≥1791);PED; 852088440; F:Benjamin /WILEY or WYLIE/|M:Abigail /Hurd/

These contextual details suggest that this instance came from a pedigree ("PED"). There is a specific birth date of 1791, but no specific death date (since the death is marked as ≥(birth_date)). The person is a female (F) from New Hampshire with her father being "Benjamin /Wiley or Wylie/" and mother named "Abigail /Hurd/".

4.3 PNPOS to Final Parses

Lastly, we mine the contextual clues to determine family name origins, patronymics, etc.; and then we apply parsing rules to convert to parses. The details of this process are extensive, but we attempt here to summarize.

4.3.1 From PNPOS to XPNPOS

If the individual's name intersects the place and/or relatives' names, these facts are added by rule to that name's PNPOS-pattern in order to create an "extended PNPOS" (XPNPOS). For example, "Jane Smith" may have originally been given a PNPOS-pattern of G1 FN1; but if her father's name was "William Smith," the XPNPOS-pattern would become "G1 FN1++FNF".

4.3.2 From XPNPOS to Parses

We grouped XPNPOS patterns by count and we specified likely parse rules for each kind of XPNPOS pattern. These parse conversion rules were applied to the name+context instances to form parses. For the XPNPOS pattern above regarding *Jane Smith*, the parse might be [NAME [GNP [G1 Jane]] [SNP [FNPF [FNF1 Smith]]]]. Although there may be 100 instances of Jane Smith in the training (or 5 in the test set), only the first instance per XPNPOS type would be presented to the annotator. If the annotator judged the parse to be correct, that parse would be treated as the correct one for all instances of Jane Smith with similar XPNPOS properties.

This procedure does lead to some error in instances where the automation did not correctly identify the extended PNPOS features. For example, if "Jane Smith"'s father was recorded as "William Smiht," the extended PNPOS may not have been augmented with the "++FNF" component which in turn could yield an improper parse. The shear volume of parses makes by-hand vetting of 10M parses intractable, so errors are corrected as they are discovered (either when flagged as suspect by parsers that will be mentioned later or through visual inspection).

4.3.3. Names May Have Multiple Parses

The reader should recognize that when certain context fields are absent, the proper origin of a name or the proper syntax may not be known. To accommodate this fact, the PNTB is designed to list the most likely parses. For example, consider the name "Anna Johanson." With limited context, there could actually be three commonly-observed parses for this name. The name could be patronymic (because it was written by genealogy patrons who did not realize that it should be "dotter" instead of "son" – which happens frequently); or it could be that Johanson is Anna's father's or spouse's family name. The PNTB would therefore store each of these possibilities, and a system would have to produce all three in order to be treated as perfectly correct:

[NAME [GNP [G1 Anna]] [PNP [RELPC [GNPF [GNF1 Johan]] [RELD +son]]]],
[NAME [GNP [G1 Anna]] [SNP [FNPF [FNF1 Johanson]]]],
[NAME [GNP [G1 Anna]] [SNP [FNPS [FNS1 Johanson]]]].

4.4 Treebank Statistics

Over the course of time, we have added to the data as new names and phenomena have been observed. As was mentioned previously, the PNTB currently has 10.0 million parses with corresponding contexts. We here present some additional interesting PNTB statistics.

Table 4 depicts the PNTB countries with their percentages of occurrence. Note that there are almost 200 countries represented, 22 of which contribute at least 0.5% of the entries. This coverage is large, but it biased toward Latin-script languages (with some limited Cyrillic and CJK representation). This bias was not intentional, but is an artifact of being drawn from FamilySearch data whose patrons have heritage from these countries.

Country	%	Country	%	Country	%
U.S.	35.6	Denmark	1.4	Bolivia	0.5
England	11.4	Finland	1.1	China	0.5
Mexico	6.7	Italy	1.0	Chile	0.4
Germany	5.8	France	1.0	Philippines	0.3
Sweden	2.6	Japan	0.9	Wales	0.3
Brazil	2.5	Netherlnd.	0.8	Jamaica	0.3
Norway	1.9	Ireland	0.7	Portugal	0.3
Scotland	1.8	Russia	0.6	Costa Rica	0.2
Canada	1.5	Argentina	0.5	Hungary	0.2
Spain	1.4	Peru	0.5	160 others	17.3

Table 4: Most-represented PNTB countries by percent

Figure 1 below shows the counts of instances drawn across historical time frames. Information from BC dates comes heavily from East Asian pedigrees which are often preserved for thousands of years.

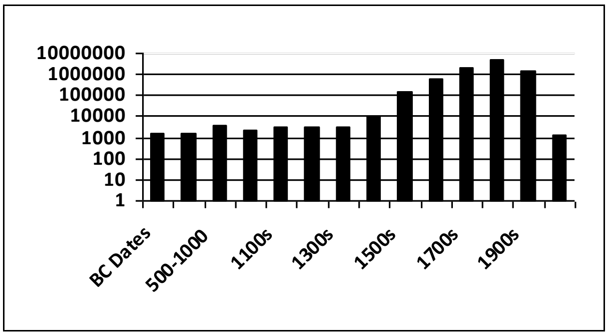


Figure 1: PNTB name frequencies by time frame

Table 5 provides the count of the unique *k*-long token parses in the PNTB. The longest name in the treebank, once segmented, has 15 tokens.

0: 1	**1**:22302	**2**:1373K	**3**:830769	**4**:193528
5:33721	**6**:6300	**7**:4263	**8**:22	**9+**:9

Table 5: Number of Terminals in Unique Parses

Lastly, Table 6 shows the frequent PNPOS patterns on the PNTB test set. *Family origin information is stripped from these rows.* Statistics from the training set are comparable. Yet the training set has many more of the "easy" names parses – which it has been provided for name coverage. It also has about 1500 additional PNPOS patterns that were not observed in the test set.

PNPOS Pattern	%	PNPOS Patterns	%
G1 FN1	36.2	G1 GNF1 RELS FN1	1.4
G1 ABBRI FN1	9.1	ABBRI1 ABBRI2 FN1	1.3
G1 G2 FN1	9.1	FN1 , G1	1.3
FN1 G1	5.2	G1 G2 IN FN1	1.2
G1 PARTRB FN1	2.1	G1 GNF1 RELS	1.0
FN1	1.8	G1 G2 INDT FN1	1.0
G1 IN FN1	1.7	ABBRGN1 FN1	1.0
G1 GNF1 RELD	1.5	**1124** Others	25.0

Table 6: Highest-frequency PNPOS patterns

5. Personal Name Parsing

The PNTB's data could potentially be useful for various NLP applications. We are principally interested in using the PNTB as training material for building a statistical personal name parser (**SPNP**) to facilitate name search. Parsing is a well-explored research area, and our desire is to use common parsing techniques. However, there are several key ways in which name parsing differs from typical, sentence-level parsing: (1) the names have a very restricted number of tokens and structures; (2) the same input string may require a different parse depending on the context in which the name is found; (3) out of vocabulary rates are high; and (4) multiple tokenizations of the same name may need to be considered to find the optimal segmentation. We address these issues here.

5.1 Parsing Options in Few-Token Scenarios

The first issue to consider for name parsing is the types of parser options that are available and which would be best suited to the task. Given the expectation that many names are of the form G1 G2 ...G(*n*-1) FNF1, we should at least begin with a default parser, **Baseline #1,** which parses names according to this premise. We can refine this by making a second **Baseline #2** where if a token is an initial or ends with a period, we should use ABBRI or ABBRGN in place of G or ABBRFNx for FNx; and if for a female, we should allow for both FNF1 and FNS1 parses. Yet as was observed in Table 6, many names do not follow these patterns and will require more in-depth analysis.

It was observed during our acquisition of PNPOS patterns that the parses for many of the *n*-token names were slight variations of the (*n*-1)-token names. This suggests that a dynamic program would likely be more effective in covering the name space. Since the names frequently contain out-of-vocabulary tokens, a statistical unlexicalized PCFG-based parser could be useful to allow for desired generalization.

One concern with PCFGs is that they have independence assumptions which are usually violated by the data. For names, this is also the case (though to a lesser degree). Klein and Manning (2003) suggest a PCFG modification which is simple but leads to improved-accuracy unlexicalized parsers (Manning (2012)) and helps to overcome some of the independence assumption. If we apply their process to our situation, we need to convert rules of the form *"NAME←GNP SNP"* into context-dependent rules like:

"*NAME←GNP|^NAME SNP|^NAME,<GNP*."

This means that the NAME is created by coupling a GNP (whose parent is a NAME) together with a SNP (whose parent is a NAME and left sibling is a GNP). We use this and refer to it as **KM Context**.

Another parsing strategy which is viable for names but would not be for most sentence-level parsing is to do statistical pattern-based or template processing. Since names are constrained in length (with the vast majority having fewer than eight tokens, as seen in Table 5), the training algorithm can count all of the naming patterns for a particular length and associate with each its probability of occurrence. For example, for two-token names, the following pattern would be expected:

[NAME [GNP [G1 ?]] [SNP [FNPF [FNF1 ?]]]].

If we have a name like "John Smith," we need only multiply Pr(Pattern)*Pr(John|G1)*Pr(Smith|FNF1) to associate a probability. We will refer to this as the **Patterns**-based approach.

5.2 Challenges of Multiple Possible Parses

Previously we saw that "Anna Johanson" could possibly have three parses in the absence of context. If context *is* provided, however, we should be able to narrow down the best parse. For example, if we knew that "Anna Johanson" was from Kentucky (United States) and born in the 1920s, then the patronymic possibility would be eliminated. If we knew that she married "Carl Johanson," and has father "John Smith", then the spouse-surname (FNS) option would be the only reasonable conclusion.

We automatically mine any provided contexts to obtain features which can help make these kinds of decisions. These features generally fall into two classes: (1) whole-name features, and (2) name piece features. In the case above, "1920s" and "Kentucky" suggest that the whole name is non-patronymic. The fact that "Anna" is female is a whole-name feature which suggests that FNS is a possible tag for Johanson. The observation that the "Johanson" of Anna's name intersects her husband's name would be a name-piece feature attached to "Johanson."

To identify whole-name features, we mine dates, places, and the character script that is used for the name to determine if the name is from a patronymic country/timeframe (PN); if the name is from a Spanish (ES), Portuguese (PT), or CJK-speaking (CJK) country; and if it is male or female. We use statistics of the name pieces and their intersections with family member names to try to determine if the name is in little-endian (given names appear before surnames) or big-endian (surnames are present and precede given names).

Determining the name-piece features is in some cases more difficult. If a specific name piece or set of name pieces are pre-marked (by patrons or record-creators) with slashes, such as "/Smith/," this mark is added as a feature. Otherwise, if a name piece (from the name to parse) is a match, a NYSIIS(1965) equivalence, or a known variant of a name piece from a family name (F,M,S,CH), a feature is added to the name piece corresponding to the family member with the intersection (such as "Johanson%%S" in the previous case). If there are no intersections with a particular family member, then we attempt to assess if that family member's family name is actually present. If statistics suggest that the family member's name likely contained a surname, we mark the name pieces with "!!" to say that the surname does not match (such as "Johanson%%!!F"). These features are induced at training and testing time. During testing, they are supplied to the parser to hopefully eliminate undesirable parse structures or to favor desirable ones and reduce cost.

Since personal names can have multiple valid parses, we fit the parser with such ability. It produces an up-to-*N*-best list where the worst result has to have at least 10% of the best's probability.

5.3 Out of Vocabulary and Backoff

In sentence parsing, a typical challenge is dealing with out of vocabulary words (OOVs) since name pieces (like “Giovanni” or “Hadamard”) are not seen in training but need to be processed at test time. The same is still true for name parsing and it plays a significant role in performance. For instance, despite our huge training set, 10.1% of the test *parses* have OOVs and 3.8% of all name *pieces* are OOV. An additional 1.3% of all name pieces (3.6% of the parses) are not OOV but have PNPOS classes in the test set which were not observed in training. As we add KM contexts and features, the challenge for unobserved situations increases dramatically. We use several strategies to try to overcome this.

First, we use Kneser-Ney (1993) smoothing where appropriate in backoff conditions. Also, to train, we divide the data into two halves. We use half to train and the other half to estimate OOV behavior, then vice versa. We count these OOVs as generic words (w1 if the first word OOV, w2 if the second is, etc) and separately as w1of *T*, w2ofT (where *T* is the number of tokens in the name). This wildcarding is beneficial for accommodating OOVs, but it has the potential of allowing parses that are illegal and must be incorporated judiciously.

Backoff comes at a computational price. This is especially true of PCFGs in that that they have O(m^3T^3) computational complexity, where *m* is the average number of grammar rules that have to be tested at any point in the computation. Backing off results in larger values of *m*. So we want to limit backoff and reduce *m* as much as possible.

We implement a **Multi-Stage** system that allows for successive levels of backing off, where each level becomes less restrictive. We also implement a filter at the end of parsing to ensure that the features have not been overlooked across the stages and that bogus parses have not been introduced through the backoff procedures.

5.4 Challenges with Tokenization

Lastly, we need to consider tokenization. The PNTB attempts to treat McLean and “Mc Lean” consistently by turning the first into “Mc+ Lean”; it decomposes patronymics; it works to re-segment falsely concatenated names like Rodriguezhernandez; it supports Chinese and Japanese tokenization; and so forth. Thus, the parser has to allow for multiple possible name segmentations. It is provided several kinds of known suffixes and prefixes, and it learns from the PNTB training data about other token sets that it should attempt to fuse or single tokens that it should try to subdivide.

The parser then needs to consider all possible segmentations. Longer phrases often have lower probabilities than shorter ones, so we multiply PCFG probabilities by $(2*T\text{-}1)^{1/2}$ to compensate where *T* is the number of tokens. We also allow for multiword unit parsing to favor multi-token segmentations that are often observed in training. We use “**NoModel**” to indicate parsing which applies the segmentation described above plus all final filters and Baseline2, but which uses none of the training data.

6. Parsing Performance Tests

We evaluate the parser on the PNTB test set which was created so that no full names that exist in the test are in the training. As mentioned, this results in a set of **301K full names** (157K different parses). Up to 5 repeat names can appear in the test set to allow for feature variations.

Constituency-level scores are usually reported in the literature for parsing, so we show them plus whole-parse scores. Since multiple parses may be observed in truth or hypotheses, we use per-name F-score as a means of giving partial credit to the specific name’s parse(s). If a reference has three parses, and the hypothesis finds one plus a separate spurious result, it would get an F-Score of 0.400. Averaging across all cases provides *average F-score*, our metric. (Note that this means that even though the system finds one proper parse for a name, it will not get full credit for its parsing of that name unless the truth set also only allows for a single parse.)

Table 7 shows average F-scores for the various parse stages we discussed in Section 5. The best score is obtained by fusing all of the parsers together where multistage KM with features is first applied, and failing that, by pattern-based parsing, and lastly, by a NoModel default. The **94.4%** constituency score is a very usable result, though we are still working to improve upon that.

Parser Configuration	Top *N*	Average F-Score on Constituents	Average F-Score on Whole Parse	Time Cost on 2.7GHz Intel i7
Baseline 1	1	0.6084	0.3840	7s
Baseline 2	1 2	0.6351 0.6353	0.4658 0.4927	7s
NoModel Baseline	1 2	0.6924 0.6928	0.5338 0.5600	200s
NoModel + Features	1 2	0.7160 0.7170	0.6012 0.6141	504s
No Backoff KM Context, else NoModel	1 2	0.8439 0.8410	0.6904 0.7095	5137s
MultiStage & KM, else NoModel	1 2	0.8538 0.8508	0.7028 0.7211	7232s
Multi+KM+Features, else NoModel	1 2	0.8892 0.8989	0.8206 0.8464	5324s
Patterns+Features, else NoModel	1 2	0.9233 0.9311	0.8610 0.8820	3041s
Multi+KM+Feat else Patterns else NoModel	1 2	0.9336 **0.9438**	0.8758 **0.9014**	5863s

Table 7: Average F-Score: Constituents or Whole Name

7. Synopsis and Conclusions

We have described the creation of our PNTB: a huge, deep-analysis, multilingual, historical tree-bank which is the first treebank to focus on personal names. By itself, this is a novel and valuable resource, which we desire to share with the research community. In addition, we have leveraged the PNTB to develop a high-accuracy personal name parser. This is also novel, and we believe it will have myriad applications. We intend to use these tools for personal name mark-up, name matching, and name search, and it is our hope that this work will motivate others to identify new applications. We would also especially hope that researchers with interests in linguistic or temporal domains that are under-represented in this collection will extend the PNTB to these domains and thus enable even more comprehensive understanding of personal names.

8. Acknowledgements

The author wishes to thank the anonymous reviewers for their time and their constructive comments.

References

BibTex (2006). www.bibtext.org

Black, E.W., Garside, R., and Leech, G. N. (1993) Statistically-driven computer grammars of English: The IBM/Lancaster approach. # 8. Rodopi.

Brants, T., Skut, W., and Uszkoreit, H. (1999). Syntactic Annotation of a German Newspaper Corpus. *Proc. of ATALA Treebank Wkshp.* Paris.

Charniak, E. (1993) *Statistical Language Learning*. MIT press, p 50.

Charniak, E. (2000) A maximum-entropy-inspired parser. *Proc. of NAACL*, pp. 132-139. Morgan Kaufmann Publishers Inc., 2000.

Curator (2012). Cognitive Computation Group. University of Illinois At Urbana-Champaign. http://cogcomp.cs .illinois.edu/curator/demo/

FamilySearch (2014) https://familysearch.org/search/collection/list

Freeman, A., Condon, S., Ackerman, C. (2006) Cross linguistic name matching in English and Arabic: a “one to many” extension of the Levenshtein edit distance algorithm. HLT-NAACL 2006, pp. 471-478

Garside, R., and McEnery, A. (1993) Treebanking: The compilation of a corpus of skeleton parsed sentences. In Black et al (1993) 17-35.

Han, C-H., Han, N-R., Ko, E-S., Yi, H., and Palmer, M. (2002) Development and Evaluation of a Korean Treebank and its Application to NLP, *LREC-2002*.

Hirschman, L., and Chinchor, N. (1997). MUC-7 Coreference Task Definition. *Proc. of MUC-7*.

Klein, D. and Manning, C. (2003) Accurate Unlexicalized Parsing. *Proc. of the 41st Meeting of the ACL*, pp. 423-430.

Kromann, M. (2003). The Danish Dependency Treebank and the DTAG treebank tool. *Proceedings of the Second Workshop on Treebanks and Linguistic Theories (TLT)*, p. 217. 2003.

Liu, Z., Lu, Q., Xu, J. (2011) High performance clustering for web person name disambiguation using topic capturing. *First Int’l Wkshp of Entity-Oriented Search*, Beijing.

Manning, C. (2012). http://www.youtube.com/watch?v= xtvP0YbO2Gc

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

McNamee, P., Dang, HT., Simpson, H., Schone, P., Strassel, S. (2010) An Evaluation of Technologies for Knowledge Base Population. LREC 2010, pp. 369-372

MITRE (2011) The MITRE Challenge. Mitrechallenge. Mitre.org/NameMatching

NYSIIS : New York State Identification and Intelligence System (1965). Fingerprint Study Group. *NYSIIS Fingerprint Classification and Identification System: Status Report*.

Kneser, R., and Ney, H. (1993) Improved clustering techniques for class-based statistical language modelling. *Third European Conference on Speech Communication and Technology*.

Schone, P., Cummings, C., Davy, S., Jones, M., Nay, B., Ward, M. (2012). Comprehensive evaluation of name matching across historic and linguistic boundaries, *Proc. from Family History Technology Workshop*, 2012.

Schone, P., Davy, S. (2012) A multilinfual personal name treebank to assist genealogical name processing. Proc. from Family History Technology Workshop, 2012.

Petrov, S. and Klein, D. (2007) Improved inference for unlexicalized parsing. *Human Language Technologies 2007: NAACL*. 404-411.

Taylor, Ann (1996) Bracketing Switchboard: An Addendum to the Treebank II Bracketing Guidelines. *Linguistic Data Consortium*.

“Treebank” (2012) http://en.wikipedia.org/wiki/ Treebank, 4 Dec 2012.

Weerkamp, W., Berendsen, R., Kovachev, B., Meij, E., Balog, K., de Rijke, M. (2011) People searching for people: analysis of a people search engine log. SIGIR2011, pp. 45-54.

Wikipedia (2014) “FamilySearch”, en.wikipedia.org

Xue, N., Xia, F., Chiou, F-D., Palmer, M. (2005) The Penn Chinese TreeBank: Phrase Structure Annotation of a Large Corpus. *Natural Language Engineering,* 11(2)207-238.

Corpas na Gaeilge (1882-1926): Integrating Historical and Modern Irish Texts

Elaine Uí Dhonnchadha[3], Kevin Scannell[7], Ruairí Ó hUiginn[2], Eilís Ní Mhearraí[1], Máire Nic Mhaoláin[1], Brian Ó Raghallaigh[4], Gregory Toner[5], Séamus Mac Mathúna[6], Déirdre D'Auria[1], Eithne Ní Ghallchobhair[1], Niall O'Leary[1]

[1]Royal Irish Academy, Dublin, Ireland
[2]National University of Ireland Maynooth, Ireland
[3]Trinity College Dublin, Ireland
[4]Dublin City University, Ireland
[5]Queens University Belfast, Northern Ireland
[6]University of Ulster, Northern Ireland
[7]Saint Louis University, Missouri, USA
E-mail: uidhonne@tcd.ie, kscanne@gmail.com, ruairi.ohuiginn@may.ie, e.nimhearrai@ria.ie,
nicmhaol@hotmail.com, brian.oraghallaigh@dcu.ie, g.toner@qub.ac.uk, s.macmathuna@ulster.ac.uk,
d.dauria@ria.ie; e.nighallchobhair@ria.ie, nialloleary.dho@gmail.com

Abstract

This paper describes the processing of a corpus of seven million words of Irish texts from the period 1882-1926. The texts which have been captured by typing or optical character recognition are processed for the purpose of lexicography. Firstly, all historical and dialectal word forms are annotated with their modern standard equivalents using software developed for this purpose. Then, using the modern standard annotations, the texts are processed using an existing finite-state morphological analyser and part-of-speech tagger. This method enables us to retain the original historical text, and at the same time have full corpus-searching capabilities using modern lemmas and inflected forms (one can also use the historical forms). It also makes use of existing NLP tools for modern Irish, and enables integration of historical and modern Irish corpora.

Keywords: historical corpus, normalisation, standardisation, natural language processing, Irish, Gaeilge

1. Introduction

This paper describes the preparation of *Corpas na Gaeilge (1882-1926)*, a corpus of historical texts, to be used in the first instance for lexicography in the Royal Irish Academy's *Foclóir na Nua-Ghaeilge[1]* [Dictionary of Modern Irish] Project. *Corpas na Gaeilge (1882-1926)* complements *Corpas na Gaeilge 1600-1882* (2004) a corpus of earlier Irish texts, which was published in 2004. The aim of *Foclóir na Nua-Ghaeilge* is the provision of a corpus-based dictionary arranged on historical principles to cover the period from 1600 to the present. The texts found in the period 1882-1926 vary in terms of orthography, morphology and syntax; therefore the processing combines both manual and automatic elements. Manual elements include the development of specific wordlists by a panel of language experts and the automatic elements include spelling standardisation, lemmatisation and part-of-speech tagging. The work which began in 2012 (directed by a management committee), is a collaboration between staff of the Royal Irish Academy (RIA), language experts and natural language processing experts. The corpus texts have been made available online in raw text format and TEI format since December 2013[2].

2. Background

The written history of the Irish language extends back to the seventh century, and perhaps up to two centuries earlier than that if we include the Ogham monument inscriptions, consisting of personal names written in a highly archaic form of Irish. As with any language that has such a long history, it is normal to divide it into periods, and for Irish the following are recognised: Old Irish (c.600-900), Middle Irish (c.900-1200), Early Modern Irish (c.1200-1650) and Modern Irish (c.1650-present).

The lexicography of Irish has been served in an uneven manner. For its earliest stages, c.600-1650 we have the *Dictionary of the Irish Language* (1976) which appeared in a series of fascicles between 1913 and 1976. It is a dictionary compiled in broadly historical terms, giving earliest attestations, variant forms and meanings and sometimes also etymologies. Due to its long period of compilation, the standard and quality of fascicles vary and, as may be expected in a work of this nature, certain entries are incomplete or out of date. However, a digitised version of the dictionary which was published online in 2007[3] has made it possible to update certain aspects of the work through supplements and additional entries. In 2013 a second revised electronic edition containing over 4,000 amendments and additions was made available online.

[1] *Foclóir na Nua-Ghaeilge:* http://www.ria.ie/research/focloir-na-nua-ghaeilge.aspx

[2] RIA Corpus: http://research.dho.ie/fng/index.php; http://research.dho.ie/fng/cuardaigh.php

[3] The electronic Dictionary of the Irish Language (eDIL): http://www.dil.ie/about.php

2.1 *Foclóir na Nua-Ghaeilge*

In the case of Modern Irish, it will be helpful to recognise two broad periods, 'revival' Irish dating from roughly 1900 to the present, and an earlier period stretching from c.1650-c.1900. Revival Irish has been relatively well served with both English-Irish and Irish-English dictionaries. A modern online English-Irish dictionary appeared in 2013[4] and an online Irish-English dictionary is scheduled to appear in 2015. Several printed Irish-English and English-Irish dictionaries appeared in the course of the twentieth century. All of these works, it should be noted, are functional dictionaries that offer English equivalents for Irish terms, or vice versa. They do not have a historical dimension and most do not give sources.

The earlier period, 1650-1900, is devoid of any modern dictionary. On the completion of the *Dictionary of the Irish Language* (600-1650) in 1976, the Royal Irish Academy established a new project, *Foclóir na Nua-Ghaeilge*, which was to provide a dictionary arranged on historical principles to cover the period from 1600 to the present, and thus continue the *Dictionary of the Irish Language*. Work on this project has been in progress since that time.

The challenges facing the compilers of this dictionary are quite daunting. In the period in question, the Irish language underwent many changes. Following the downfall of the Gaelic aristocracy in the early seventeenth century and subsequent colonisation and plantation, the language became in the course of the next centuries mainly the language of a rural peasantry. Despite its reduced status, the growth in population meant that there probably were more people speaking Irish in the seventeenth and eighteenth centuries than at any other time in its history. This, coupled with the widespread availability of paper and, to a lesser extent printing, has left us with a large body of material – devotional texts, historical tracts, songs, poems and tales from this period. The electronic corpus of material being compiled in the RIA, to be used in drafting the dictionary entries, is drawn from both written and oral sources and when complete will comprise 90+ million words, it is estimated.

The broadly standardised written language which obtained down to the seventeenth century was replaced by a written language that varied quite widely in its orthography, morphology and grammar. Dialectal forms came very much to the fore, and due to the expansion of English many words were borrowed from that language. In 1958, the spelling and orthography of Irish were again standardised and this standardised language is now used in most published works in Irish, including dictionaries, where the headwords are in a form appropriate to this standard. This standard can be at a considerable remove from forms found in our corpus of material. Processing these historical texts to enable their efficient use in lexicography presents many challenges.

In 2004, 705 texts from the period 1600-1882 were published in CD form as *Corpas na Gaeilge 1600-1882*[5]. This 7.2 million word corpus comes with a concordance of all forms occurring in these texts, but as the texts have not been lemmatised or annotated with part-of-speech tags, the variants, inflected forms, etc., are not grouped together under one headword and may be widely dispersed.

3. Corpas na Gaeilge (1882-1926)

The corpus being described in this paper, *Corpas na Gaeilge (1882-1926)* also contains approximately 7 million words. This corpus consists of books, published by more than twenty publishers, covering a wide range of topics and genres as well as representing the three major dialects of Modern Irish. (Newspapers and periodicals are currently in preparation). A breakdown of topics and genres are given in Table 1.

Text classification	**Number of texts**	**% of total**
Informational works		
Folklore	75	26
Textbooks	12	4
Linguistics	20	7
Other non-fiction	65	23
Sub-Total	**172**	60%
Creative works:		
Poetry	17	6
Drama	18	6
Short story collections	57	20
Novels	20	7
Essays	2	1
Sub-Total	114	40%
Total	**286**	**100%**

Table 1: Classification of texts in Corpas na Gaeilge (1882-1926)

Modern Irish has three distinct dialects (and further sub-dialects). It is necessary for a corpus-based historical dictionary to have sufficient representation of each of the major dialects in its corpus. The dialectal composition of the corpus is given in Table 2.

Dialect	**Number of texts**	**% of total**
Connacht	54	20
Ulster	35	12
Munster	81	28
Translations from other languages	31	11
Non-dialect	85	30
Total	**286**	**100%~**

Table 2: Dialectal composition of texts in Corpas na Gaeilge (1882-1926)

[4] Foras na Gaeilge's New English Irish Dictionary: http://www.focloir.ie/

[5] Corpas na Gaeilge (1600-1882): http://www.ria.ie/Research/Focloir-na-Nua-Ghaeilge/Foilseachain--Publications.aspx

As all of these texts predate computerisation, the texts had to be captured in electronic form, either by typing or by scanning. In the earlier stages of the *Foclóir na Nua-Ghaeilge* project, texts were typed, followed by a period where optical character recognition (OCR) methods were tested in parallel with typing. In recent times, the majority of texts are scanned and OCR is performed using Optopus OCR software[6] . The decision regarding whether texts are typed or scanned is based mainly on the quality of the print or the condition of the book. Close attention is paid to the accuracy of the data capture process. Through a series of experiments it emerged that the most effective method of ensuring high accuracy involves a combination of close reading and enhanced spellchecking.

4. Corpus Processing

4.1 Corpus Processing Tools for Irish

In order to process *Corpas na Gaeilge (1882-1926)*, a survey of existing natural language processing (NLP) tools for Irish was carried out. A finite-state morphological analyser and part-of-speech (POS) tagger for the modern standardised language (post 1958) were available (Uí Dhonnchadha & van Genabith, 2005). In this rule-based system, POS tagging and lemmatisation are carried out in two stages. In the first stage, each token in the text is analysed using the finite-state morphological analyser (both *xfst*[7] and *foma*[8] versions are available), and a set of possible morphological analyses and lemmas are assigned to each token. In the second stage, context specific rules (Constraint Grammar[9]) are used to determine the most likely POS for the token based on its surrounding context in the sentence. This POS tagger has 95-96% accuracy on unrestricted text.

The morphological analyser's lexicon incorporates all of the 50K headwords in the Ó Dónaill (1977) dictionary and generates all inflected forms of the headwords. It also includes a set of morphological guessers which use morphological clues (e.g. distinctive inflectional suffixes) in unrecognised words to guess the likely part of speech and features of unknown words.

These tools however could not deal with the older and varied spellings prevalent in this historical corpus. One option would be to extend the existing POS tagger's lexicon to incorporate older forms. This would probably reduce efficiency and result in increased ambiguity in this rule-based tagger. Another possibility would be to create a specialized version of the tagger for the time period in question. This would require substantial development and would have limited reusability.

A better solution would be to standardise the texts in such a way as to enable them to be processed with the modern POS tagger. Fortunately, there was a prototype standardiser available; *An Caighdeánaitheoir* (Scannell, 2009). With further development and training (described in Section 4.2) this could be used to associate a modern standard (inflected) form with pre-standard words in the texts, and thereby enable these texts to be lemmatised and POS tagged with minimal adjustments to the existing POS tagger. This solution has the advantage that sub-corpora from different historical periods can be POS-tagged in the same way and all pre-standard inflected word-forms can be united under the modern lemma. This will enable lexicographers to search the corpus for examples using the modern spelling (either lemma or inflected form) and retrieve all variant and historical forms as desired.

4.2 *An Caighdeánaitheoir* (The Standardiser)

The strategy employed by the standardiser is to treat spelling standardisation as a problem in machine translation between two very closely related languages: pre-standard and standard Irish. Indeed, our implementation uses well-known techniques in statistical machine translation and can be viewed as a variant of the word-based IBM model 1 (Brown *et al*, 1993). As such, the key elements are a *language model* for the target language (standard Irish), and a *translation model*, representing the conditional probability of a particular non-standard spelling corresponding to a particular standardised spelling. There were challenges to overcome in constructing both models.

We use a trigram language model for standard Irish, and training simply requires a sufficiently large corpus. The difficulty here is a philosophical one; namely, what is "standard Irish"? A major spelling reform was introduced in the 1940s and brought to completion in 1958, and operationalized through the publication of two major bilingual dictionaries in the second half of the twentieth century[10]. A simplified grammar was published in 1958, together with the spelling recommendations, as *Gramadach na Gaeilge agus Litriú na Gaeilge: An Caighdeán Oifigiúil* [Irish Grammar and Spelling: The Official Standard]. In practice, however, the story is quite complex; the published dictionaries do not adhere completely to the standard, nor do certain widely-used grammars, e.g. New Irish Grammar (The Christian Brothers, 1994). To add to the confusion, a major revision of the official standard was published in 2012, with the goal of simplifying the rules and bringing the standard more in line with the language as spoken by native speakers in the Gaeltacht. The result is that despite having access to a large corpus of texts (Scannell, 2007) (more than 100 million words)

[6] Optopus is a trainable OCR program from Makrolog, Germany. http://www.makrolog.com/ It is currently unsupported.

[7] Xerox Finite State Tools: http://www.stanford.edu/~laurik/fsmbook/home.html

[8] Foma Finite State Compiler: http://code.google.com/p/foma/

[9] VISL Constraint Grammar: http://beta.visl.sdu.dk/constraint_grammar.html

[10] (Ó Dónaill, 1977); (de Bhaldraithe, 1959)

published since these reforms were put into place, virtually none of the texts fully complies with any variant of the standard. How does one create a language model when there are *no* non-trivial texts written in that language?

Our solution is to employ a suite of rule-based proofing tools (spelling and grammar correction) developed by the second author to create a sub-corpus of about 40 million words consisting of the texts which are most compliant with the official standard, at least as it is implemented in the proofing tools. Additionally, we applied automated standardisations to a small number of recurring non-standard forms in order to produce a training corpus which best approximates to "standard Irish".

Our approach to the translation model is somewhat unusual in that we do not train the probabilities using a "bilingual" corpus and the Expectation Maximization (EM) algorithm (Koehn *et al,* 2007), as is typical in this context. The reason is, in short, that we can do substantially better with an *ad hoc* approach using resources we have at hand. First, we only have access to a relatively small number of texts written in both pre-standard and standard Irish (about 700,000 words), and there is a tremendous amount of variation among these pre-standard texts in terms of both dialect and time period. The statistical alignments generated from this corpus are quite noisy and unsuitable for the high-precision translation task at hand. Second, through ongoing lexicographical work, we already have a large, manually-curated database of about 22,000 pre-standard lemmas mapped to their standard forms, plus an additional 10,000 mappings taken directly from the Ó Dónaill (1977) dictionary.

Finally, whereas many spelling standardisations are essentially arbitrary choices of one form over others (for example, *eileastrom, feileastar, seileastram* are all treated as variants of *feileastram* 'wild iris'), many others are consequences of a number of general context-sensitive rules. For example, a word internal *-bhth-* is standardised to *-f-*, a word final *-ghail* standardises to *-aíl*, and *sg-* always becomes *sc-*. The current version of the standardiser implements 567 hand-written rules of this type.

Quite commonly, a standardised form is discovered through a combination of rule applications and lexical standardisations. For example, the hypothetical form *coimh-mheasguighthe* would undergo the following sequence of spelling changes; the first five represent applications of general rules, while the final change maps a non-standard second declension verb to its standard first-declension form, using a mapping found in the lexical database:

coimh-mheasguighthe → *comh-mheasguighthe* → *cóimheasguighthe* → *cóimheascuighthe* → *cóimheascaighthe* → *cóimheascaithe* → *cóimheasctha* ('coalesced')

The definition of the translation model is naive but effective in our context. All non-standard forms that are paired with a particular standard word in the lexical database are viewed as having the same conditional probability. Non-standard forms that are paired with a standard word through one or more rule applications are "penalized" for each rule that is applied; that is to say, the conditional probability is multiplied by a fixed factor $\beta < 1$ each time a rule is applied. A tuning process allows us to choose an optimal value for β; a small corpus of standard/non-standard sentence pairs was held out, and the performance of the standardiser was evaluated for different values of β.

Once the language model and translation model are in place, the decoding process is straightforward. Some standardisations map more than one word to a single word (*i mbárach* → *amárach*), but a pre-processing step treats these set phrases as a single token, so we can effectively decode wordfor word. Decoding proceeds from left to right, maintaining a data structure of all possible hypotheses and their probabilities. Since we use a trigram language model, when multiple hypotheses share the same final two words, we discard all but the highest probability candidate. At the end of each input sentence, the maximal probability hypothesis is output.

4.3 Initial Survey of *Corpas na Gaeilge (1882-1926)*

In order to establish the extent of the non-standard spelling in *Corpas na Gaeilge (1882-1926),* the finite-state morphological analyser was run on the seven million words of raw text before standardisation. As expected, many non-standard spelling were not recognised by the morphological analyser and therefore could not be accurately assigned tags automatically. Morphological guessers were not used as they would not be able to predict the modern equivalent lemma for non-standard spellings.

Of the 7 million words, 65% of word types were not recognised, and 35% of word types were recognised, i.e. words that are the same in the modern standard. Ignoring uppercase/lowercase distinctions there are 166K (approx.) different unknown words (types) which have non-standard spelling.

The 166K types were sorted according to frequency of occurrence in the corpus, e.g. the word *chuaidh* 'went' which is near the top of the frequency list occurs 5700 times in the corpus. Therefore, by adding this word to the Standardiser's lexicon (i.e. by pairing it with its modern equivalent *chuai**gh*** 'went'), 5700 instances of this word in the corpus will be automatically identified by the POS tagger.

The 1500 most frequent unknown types account for 50% (approximately) of the unknown tokens in the corpus. By manually assigning the modern equivalent to the 1500 most frequent non-standard word types, and adding these pairing to the Standardiser's lexicon, 50% of the unrecognised word forms can immediately be assigned the correct standard form, while the remaining unrecognised word forms can be processed by using rules and probabilities.

However, as we go down the list, the types occur less and less frequently in the corpus, so that the benefit of manually adding pairings to the lexicon makes less and less of an impact on the overall recognition rates. For example adding a further 500 words pairings would only improve the recognition rates by less than 3%, (e.g. 2028 most frequent types account for 53% approximately (in this sample).

The language experts on the team took the 1500 most frequent items on the list of non-standard types and associated them with the modern standard wordform. This became the basis of a database of non-standard to standard pairings. These pairings were then used directly by *An Caighdeánaitheoir.* In addition, approximately 2500 named entities (people, places etc.) which had been manually marked up in the first corpus, *Corpas na Gaeilge 1600-1882*, were also added to this database, and used in the same way.

4.4 Processing Stages

Our processing of the historical texts goes through a series of stages: tokenisation, standardisation, lemmatisation and POS tagging. We will briefly describe each stage.

Tokenisation: Firstly, the text is segmented into units called tokens. For many languages including Irish, tokenisation is mainly based on space between words, where a word equates to a token. But sometimes we may wish to divide a word into more than one token, e.g. a contracted form such as *I'm* is separated into two tokens: *I* (pronoun) and *'m* (= am verb). And sometimes we wish to keep two or more words together as one token, e.g. names or placenames such as *Finnegan's Wake*, *Baile Átha Cliath* 'Dublin' (proper nouns) or compound prepositions, e.g. *tar éis* 'after', *os cionn* 'above', where it does not make sense to analyse the parts individually.

Standardisation: The pre-standard inflected forms encountered in historical and dialectal texts are annotated with their modern standard inflected equivalents using *An Caighdeánaitheoir*.

Part-of-speech (POS) Tagging and Lemmatisation: The standard inflected word forms are processed using the POS tagger, enabling part-of-speech tag and lemma annotations to be added.

The following sentence (1), taken from the beginning of a short story printed in 1913, illustrates the process.

```
(1)  Bhí baintreabhach mhná  ann
     fad  ó shoin.
     Was widow         woman there
     long ago
     'There was a widow long ago'
```

Table 3 shows the sentence in vertical form, i.e. column 1 represents the original text showing one token per line. In column 2 we see the standard forms, in column 3 we see the PAROLE[11] POS tag and in column 4 we have the lemma.

Original	Std. form	POS	Lemma (base)
Bhí	Bhí	Vmis	bí
baintreabhach	baintreach	Ncfsc	baintreach
mhná	mná	Ncfsg	bean
ann	ann	Rl	ann
fad	fada	Rt	fada
ó	ó	Sp	ó
shoin	shin	Pd	sin
.	.	Fe	.

Table 3: Sample annotation of historical text

Formatting: The annotated corpus is formatted in both vertical form (similar to Table 3) and in XML Corpus Encoding Standard (XCES[12]) format as follows , using the `<w>` word tag with attributes `tag` for POS, `base` for lemma and `std` for standard form.

```
<p>
<s>
<w tag = "Vmis" base = "bí" std =
"Bhí">Bhí</w>
<w tag = "Ncfsc" base = "baintreach" std =
"baintreach">baintreabhach</w>
<w tag = "Ncfsg" base = "bean" std =
"mná">mhná</w>
<w tag = "Rl" base = "ann" std =
"ann">ann</w>
<w tag = "Rt" base = "fada" std =
"fada">fad</w>
<w tag = "Sp" base = "ó" std = "ó">ó</w>
<w tag = "Pd" base = "sin" std =
"shin">shoin</w>
<w tag = "Fe" base = "." std = ".">.</w>
```

This same vertical/XCES format is used for modern texts (e.g. the 30 million word *Nua-Chorpas na hÉireann*[13]) in which case the `std` value is usually the same as the original token (except for dialectal variants). In this manner historical texts and modern texts can be seamlessly integrated. Therefore, rather than normalising pre-standard forms, (e.g. as in Bollmann *et al* (2012), the

[11] Full specification of the Parole tags can be found at https://www.scss.tcd.ie/SLP/parole.htm
[12] XCES: http://www.xces.org/
[13] http://corpas.focloir.ie/ which can be queried using the SketchEngine interface (http://the.sketchengine.co.uk)

original form is kept and is annotated with modern standard form. Historical forms which do not have a modern equivalent are added to the finite-state POS tagger lexicon and associated with the modern lemma where one exists.

5. Evaluation

A detailed evaluation of three texts, (each one representative of one of the three major dialects), was carried out by the language experts. There were a number of difficulties. Firstly, the texts are in pre-standardised orthography. Secondly, there are differences, sometimes quite significant, between the dialects in terms of morphology, inflexion and syntax. Finally, there was small number of typographical errors and errors made in the scanning process. Consequently, there were problems noted at every stage of processing, including some errors in the original texts themselves.

Texts:
There were occasional typographical errors in the printed works, and some residual OCR errors.

Tokenisation:
There were many issues surrounding the non-standard use of hyphens to connect words which should be two tokens e.g. *oidche-sin* 'night-that' or even to connect suffixes to words, e.g. *táim-se* in place of *táimse* 'I am (emphatic)'. There were also difficulties regarding the widespread use of apostrophes for elided and contracted forms, particularly in stories using direct speech, again causing two tokens to be joined, e.g. *c'acu* in place of *cé acu* 'which of-them'.

Standardisation:
There were problems connected to residual non-standard spellings and non-standard inflectional morphology, e.g. *dubhras* 'I said' with non-standard spelling and incorporated pronoun, rather than the standard form *dúirt mé* 'I said' which has a separate pronoun. The most difficult problem to remedy (systematically) is where a non-standard root is used e.g. *dtearn* 'did' rather than the modern standard form *ndearna* 'did'.

POS tagging:
There were also problems at the POS tagging stage; the most common problem related to older verbal forms which used to have a preverbal particle *do* with the past-tense form. This preverbal particle takes the same form as the modern preposition *do* 'to', causing the POS tagger to tag some past tense verbs as nouns. There are also a small number of nouns which had a different gender in historical texts, based on inflection and accompanying definite article (or anaphoric references). These are tagged with the modern gender which will be wrong in those instances.

After these problems were addressed in the corpus processing tools, the same three texts were re-processed and the current accuracy of the standardiser's output is calculated to be approximately 95%. Other random samples have been selected from the corpus for detailed evaluation and initial calculations show accuracy rates ranging from 91-96%. The accuracy of POS tagging and lemmatisation has not yet been evaluated.

6. Conclusions

We believe this to be a very promising method of processing and integrating historical and contemporary documents, which makes maximum use of existing tools while developing specific standardisers for specific time periods. The original texts are not changed[14], rather additional information is added. The project relies on the collaboration of many different individuals each with different skills, without which the work could not be accomplished.

7. Future Work

In the immediate future, a substantial body of material from newspapers and periodicals from the period 1882-1926 will be processed. This work is currently in hand and it is envisaged that a further 3.5 million words from over twenty-five periodicals (many of which are of great historical interest) will be added to the corpus. Following this, the possibility of processing and integrating the earlier *Corpas na Gaeilge 1600-1882* with the current corpus will be investigated.

8. References

Bollmann, M., Dipper, S., Krasselt, J., and Petran, F. (2012). Manual and Semi-automatic Normalization of Historical Spelling - Case Studies from Early New High German. In *Proceedings of KONVENS 2012 (LThist 2012 workshop)*. Vienna, September, 2012

Brown, P., Della-Pietra, S., Della-Pietra, V. and Mercer, R. (1993). The Mathematics of Statistical Machine Translation. *Computational Linguistics*, 19(2), pp. 263-313.

Corpas na Gaeilge 1600-1882. (2004). Royal Irish Academy: Dublin.

Dictionary of the Irish Language (1976). Edited by E.G. Quin. Royal Irish Academy: Dublin.

de Bhaldraithe, T. (1959). *English-Irish Dictionary*. An Gúm.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., … & Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions* (pp. 177-180). Association for Computational Linguistics.

Ó Dónaill, N. 1977. *Foclóir Gaeilge-Béarla.* [Irish-English Dictionary]. Baile Átha Cliath: An Gúm.

Scannell, K. (2009) *Standardization of corpus texts for the New English-Irish Dictionary*, paper presented at the 15th annual NAACLT conference, New York, 22

[14] Except for obvious typographical and OCR errors. There is also an online archive which preserves the original printed layout http://research.dho.ie/fng/index.php

May 2009. (http://borel.slu.edu/pub/naaclt09.pdf)

Scannell, K. (2007). The Crúbadán Project: Corpus building for under-resourced languages, Cahiers du Cental 4 (2007), pp. 5-15, C. Fairon, H. Naets, A. Kilgarriff, G-M de Schryver, eds., "*Building and Exploring Web Corpora*", Proceedings of the 3rd Web as Corpus Workshop in Louvain-la-Neuve, Belgium, September 2007.

The Christian Brothers. (1994). *New Irish Grammar*. Baile Átha Cliath: C. J. Fallon.

Uí Dhonnchadha, E. and van Genabith, J. (2005) Scaling an Irish FST morphology engine for use on unrestricted text. *In* editor(s)A. Yli-Jyrä, L. Karttunen, J. Karhumäki, *Lecture Notes in Artificial Intelligence (LNAI): Proceedings of the FSMNLP 2005 Finite-State Methods in Natural Language Processing*, Berlin, Springer-Verlag, 2006, pp247-58

Language resources for early Modern Icelandic

Ásta Svavarsdóttir, Sigrún Helgadóttir, Guðrún Kvaran
The Árni Magnússon Institute for Icelandic Studies / University of Iceland
Reykjavík, Iceland
asta@hi.is,sigruhel@hi.is,gkvaran@hi.is

Abstract

This paper describes the compilation of a language corpus of early Modern Icelandic, intended for research in linguistics and lexicography. The texts are extracted from a digital library, accessible on the website *Tímarit.is*, containing scanned images of individual pages and OCR read text from all Icelandic newspapers and periodicals from that period. In its present form this resource does not fulfill all needs of linguists and lexicographers, due mainly to errors in the digitized texts, the lack of annotation, and limited search possibilities. To create a new language corpus from this text material with the time and money available, methods and tools for automatic or semi-automatic correction of OCR errors had to be developed. The text was to be corrected according to the originals, without any standardization, which poses various challenges in the construction of the corpus. These connect to the correction process itself, the possibilities of using available tools for tagging and lemmatizing, as well as the design of search functions and interface. The solution was to build a parallel corpus with two layers, one with diplomatic text and the other with a standardized modern version of the same text, with mapping between the two.

Keywords: historical corpora, automatic correction, standardization and mapping

1. Introduction

There is great demand for historical language resources for research purposes in lexicography and linguistics. In Iceland, there exist so far only two historical corpora designed primarily with such needs in mind, 'The Icelandic Parsed Historical Corpus' (IcePaHC; (Rögnvaldsson et al., 2012)), spanning the entire history of the Icelandic language from the 12th century onwards, and a corpus of Old Icelandic, i.e. medieval, narrative texts, mainly from the sagas (Rögnvaldsson and Helgadóttir, 2011).[1] The texts in both corpora have a standardized, modern spelling, which has made it possible to handle them with available language technological tools, and facilitates search. At the same time this feature limits the use of these resources. They have proved to be valuable for syntactic, and to a certain extent morphological, research, and they are also useful in lexicography within the limits set by their relatively small size. On the other hand, they can not be relied on in historical investigations of e.g. orthography or phonology.

This paper presents an ongoing project which aims at the building of a corpus of early Modern Icelandic, i.e. 19th and early 20th century language, with non-fictional prose. The texts are extracted from a digital library which includes all Icelandic newspapers and periodicals published in that period (Hrafnkelsson and Sævarsson, 2014). This resource, accessible on the website *Tímarit.is*, contains scanned images of individual pages and a raw version of an OCR read text, i.e. without corrections or adaptation of any kind. The existence of this digitized text material, however defective, together with the immediate needs of several research projects for resources of 19th century language, were the main motivations for the attempt to build the corpus. It was clear that to create a large enough corpus to serve the various needs of the research projects involved, methods and tools for automatic or semi-automatic correction of OCR errors in the texts had to be developed. Some of these projects, aimed at a broad investigation of 19th century language and language use, requested that the text would be corrected according to the spelling of the originals, without standardization of any kind. This poses various challenges in the preparation and construction of the corpus, with respect to the correction precess itself, the possibilities of using available tools for tagging and lemmatizing, as well as the design of the search functions and interface. The solution was to aim at a corpus with two layers, one with diplomatic text and the other with a standardized modern version of the same text, with mapping between the two. The result would be a parallel corpus, where the layers are not two different languages, but two stages of the same language. This would enable the application of language technological tools designed for the modern language, and allow the users to apply the well kown and standardized modern word forms to search the corpus, and get all variants of these word forms as part of the results. The construction of the corpus, due to be completed in 2014 or early 2015, is described in the article.

The organization of the paper is as follows: In chapter 2., we describe the main characteristics of the digital library, which supplies us with the OCR read text, and the reasons why this resource does not fulfil the needs of linguists and lexicographers. Chapter 3., the main part of the paper, deals with the corpus building process. Here we first discuss the general objectives and main requirements for the content and functions of the corpus, then we describe the selection of texts and their extraction from the archives, and after that we explain the correction procedure and the development of methods and tools for correcting the OCR read texts. The last section recounts the present state of the project and the remaining tasks. Finally, there is a short chapter where we draw the main conclusions from the experience of the procedure we have followed in the project.

[1] The corpus is accessible for search at the website `http://mim.arnastofnun.is/index.php?corpus=for`.

2. A digital library of Icelandic newspapers and periodicals

2.1. *Tímarit.is*: Description and objectives

The National and University Library of Iceland has compiled a digital library of Icelandic newspapers and periodicals, and made it available at the website *Tímarit.is*.[2] The collection covers the period from the late 18th century to the present, with a (nearly) complete coverage of Icelandic newspapers and periodicals published before 1920, and a great and increasing selection of titles from then on. No authorization is needed for the pre-1920 material, but later texts are added with the agreement of their publishers. The database currently contains a total of 866 titles, and pages available online are approximately 4.5 million.[3]

The digital library consists of scanned images of each page (pdf-files), with an OCR read text of the respective pages (txt-files) attached to them.[4] The tool applied for the OCR reading is *AbbyyFineReader*.[5] No corrections are made to the OCR read texts, and there are numerous errors in the text files, although their number varies considerably depending on the quality of the original. The metadata documented for each title in the collection includes the following: publication type (journal, newspaper, etc.), language, number of volumes, number of issues, publication period, location(s), publisher(s), (keyword(s)), description, etc.

All the material is contained in a searchable database, and displayed at *Tímarit.is*. Text search is limited to strings of letters (e.g. word forms), one or more at a time. Metadata can to some extent be used to delimit the search, i.e. to a particular title or a certain period. Results are presented Google-style, and the user can reorder them chronologically, or filter them by title and/or period. The user can also choose whether the results return short (<2 lines) or long (approx. 4 lines) snippets of text. The interface only allows the user to view one page at a time, whether it is selected by browsing (by title, year and issue) or as a result of a text search. This applies both to the images and the text attached to them.

The main objective of the *Tímarit.is* database is to make the newspapers and periodicals easily accessible to the research community as well as the general public.[6] It is especially useful as a research tool in many fields of social sciences and the humanities, including history, literature and language studies of various kinds.

2.2. *Tímarit.is* as a resource for language research

The *Tímarit.is* website has been successfully applied as a resource in linguistics and lexicography. It has, in particular, served as a valuable source of examples and citations. As an effective and reliable tool in language research this resource has, however, various limitations and shortcomings in its present form. Due to OCR errors in the texts, examples of words and structures can, for example, be easily missed even if they occur in the material. For the same reason, as well as the lack of annotation, it cannot be trusted that the search returns all and only relevant examples. The database can therefore not be applied in any kind of quantitative research, even if the results may give a vague indication of the existence, (in)frequency, or distribution of a particular word, word form or word combination. Furthermore, the form in which the search results are presented makes it difficult to get an overview of the results. Working with the database is very cumbersome as each page has to be retrieved separately to check the example, and, if it is relevant, it must be copied into another file for analysis. So even if the *Tímarit.is* gives access to much valuable language material, there is a lot to be wished for concerning the form and presentations of the data with respect to linguistic and lexicographic research.

3. A corpus of early Icelandic newspapers and periodicals for the purposes of language research

3.1. Objectives and requirements

On account of the shortcomings of *Tímarit.is* as a resource for language research, as described in 2.2., a separate corpus of early Icelandic newspapers and periodicals is under construction at The Árni Magnússon Institute for Icelandic Studies. This can be seen as a sub-corpus of the *Tímarit.is* archives in the sense that all the digitized text material is extracted from that. The main objectives of the project are to compile a corpus of early Modern Icelandic non-fictional texts, and construct a database and search interface for the corpus that serve the special needs of research projects in linguistics and lexicography, as well as practical tasks such as dictionary making.

The corpus will cover the 19th and early 20th centuries. Texts from that period, especially the first part of it, pose special problems that have to be solved in the process, as they are not standardized, neither with respect to orthography nor morphology, and the spelling can vary greatly from one title to another, or even within the same paper, e.g. from one time to another. One of the requirements for the corpus is that the texts should be presented with their original spelling and word forms, and in the development of tools for a semi-automatic correction of the OCR read text this has to be taken into account. Another requirement is that search in the corpus should be efficient, flexible, and preferably include possibilities that go beyond simple text search for particular words or word forms. The prerequisite for this is that the texts can be grammatically tagged

[2] The website also contains Faroese and Greenlandic newspapers and periodicals, and the compilation of the corpus and the construction of the database is a collaboration between the National and University Library of Iceland, The National Library of the Faroe Islands, and The National and Public Library of Greenland. In this article, however, we are only concerned with the Icelandic material.

[3] Cf. `http://timarit.is/about_init.jsp?navsel=3&lang=en`. The figures include Faroese and Greenlandic texts, but those are only a small minority of the collection (approx. 30 titles). Some of the Icelandic papers, esp. the older ones, were published in Denmark or the Icelandic settlements in Canada.

[4] Cf. `http://timarit.is/view_page_init.jsp?pubId=315&lang=en` for an example.

[5] Cf. `http://finereader.abbyy.com/`.

[6] Cf. `http://timarit.is/about_init.jsp?lang=en`.

and lemmatized, both in order to overcome the problems posed by the many spelling variants of words and word forms, and to make it possible to search not only for word forms but also for grammatical features. Language technological tools for tagging and lemmatizing Icelandic texts are available, but they have been developed for the modern language (Loftsson, 2008), and it is unclear how useful they would be for earlier language stages. They have, however, been applied successfully for the annotation of Old Icelandic texts (Rögnvaldsson and Helgadóttir, 2011), but these texts had already been standardized for publication according to Modern Icelandic standard orthography. In the construction of the present corpus, the plan is to map the corrected texts automatically to a standardized version, i.e. Modern Icelandic standard spelling, to enable the application of the available tools, and the corpus will thus consist of two layers of text, a diplomatic version and a standardized version.

The immediate needs of ongoing lexicographic and linguistic projects, i.e. studies of lexical borrowings in the 19th and early 20th centuries, [7] and an investigation of language variation and language standardization in 19th century Icelandic, [8] have influenced the content and structure of the corpus. Nevertheless the corpus is intended as a general resource for language research, and is not limited to these particular projects.

3.2. The selection of texts and scope of the corpus

The main criterion for the selection of texts for the corpus was that the collection would be sufficiently representative of the genre and the period in question, within the limits of the *Tímarit.is* archives. The genre is, as previously described, newspapers and periodicals. Such texts were, in fact, a substantial part of published Icelandic texts at the time, esp. in the 19th century, and they are a good representative for non-fictional texts in general, covering a variety of topics and sub-genres, such as narratives, news, discussions, advertisements, etc. (and some of them even include some fiction as well). The period that the corpus covers is the 19th and early 20th centuries, until around 1920. The *Tímarit.is* archives, which is the material available to choose from, contain approx. 35,000 issues of 300 Icelandic titles published in Iceland and Denmark during that period.[9] The distribution of published texts over time is, however, far from even, both with respect to the number of titles and the number and size of issues relating to these titles, and there is much less material to choose from in the first part of the 19th century, than there is later. This is in many ways reflected in the corpus as the selection of texts had to be denser for the early 19th century in order to cover the entire period, and it was difficult to ensure the desirable diversity in the older texts. The proportion of extracted texts in each decade over the period, estimated by title and issue, is displayed in Figure 1.

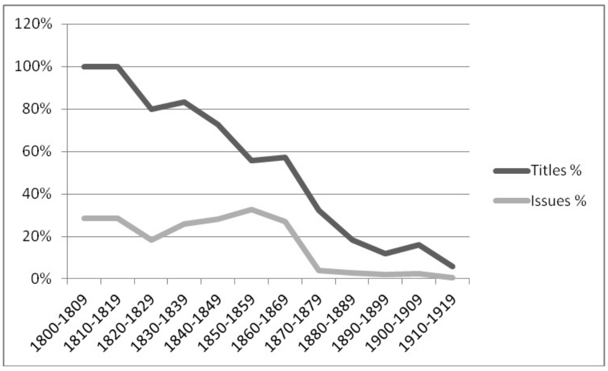


Figure 1: The proportion of titles and issues extracted for the corpus out of the available material in the *Tímarit.is*-archives by decades (1800-1919)

In the early period, text from the great majority of published papers has been extracted and these are around a quarter of available issues. In the late period, on the other hand, text from a small minority of available titles and issues has been extracted. This partly reflects the increase in published newspapers and periodicals in the last 4-5 decades of the period, as the amount of text extracted for the language corpus actually increases considerably over time (cf. Figure 2 below).

The selection of texts from the early period is further limited by the fact that material printed in the only press available in Iceland until about 1840 were in fraktur typesetting, which is practically unreadable by the OCR reading tools currently applied at The National and University Library. The amount of errors in these texts make it unfeasible to try to correct them (semi-)automatically, and most of the earliest texts in the corpus would therefore need to be from Icelandic papers printed and published in Denmark, where latin typesetting became frequent some decades before. To compensate for this, at least to a certain degree, text from a few issues of the earliest papers was entered manually to be included in the corpus.

The texts were selected according to the criteria described above, and extracted from the *Tímarit.is* database in two stages:

1. Texts from ca. 1870-1920
2. Texts from ca. 1817-1870

In addition, two sets of texts were handled separately:

4. Texts from ca. 1800-1840, printed with fraktur typesetting, and entered manually from the images
5. Texts from one periodical, 9 issues, published 1835-1847, which were handled separately in a related project (Daðason et al., 2014)

The selection was done by linguists and lexicographers at The Árni Magnússon Institute for Icelandic Studies, and the technicians at the National and University Library then

[7] A dictionary of 19th and 20th centuries loanwords in Icelandic (Guðrún Kvaran; cf. `http://www.arnastofnun.is/page/tokuord_19_old_20_aldar`).

[8] *Language Change and Linguistic Variation in 19th-Century Icelandic and the Emergence of a National Standard*; cf `http://www.arnastofnun.is/page/LCLV19_project`.

[9] This amounts to a total of approx. 270,000 pages of scanned and OCR read text, but figures for the number of running words could not be obtained.

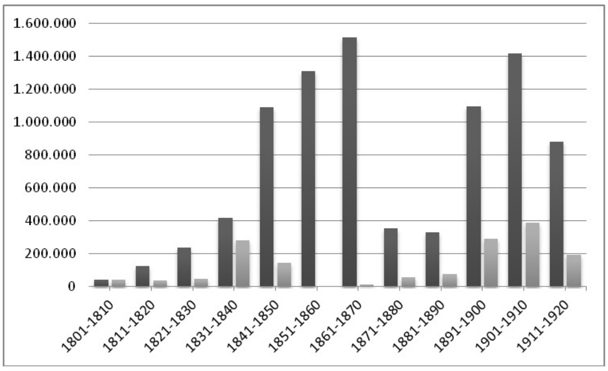


Figure 2: Overview of the amount of texts by decades, total of extracted texts (dark) and already corrected texts (light), measured in number of running words

extracted the selected OCR read texts from the *Tímarit.is*-archives. The selection was based mainly on the type and content of texts, as well as the place and year of publication, according to the criteria described above. The texts were extracted by issue, i.e. each text file contains one issue of a particular newspaper or periodical. The first collection of texts, acquired in 2010, consists of 625 issues of 29 different titles, a total of approx. 4.1 million running words. These texts were the basis for the development of tools for the (semi-)automatic correction, and, as a part of this process, texts from approx. 190 issues of 28 different titles (a total of 1.4 million running words) have been fully corrected, either manually or semi-automatically (cf. section 3.3. below). The second collection includes approx. 560 issues of 14 different papers, again a total of approx. 4.1 million words. These were extracted in 2013, and still await correction. In the third collection, there is manually entered text from the early period, 32 issues (only a part of a few of them) of 7 titles, a total of 256,000 running words. The text entry was done on the basis of the scanned pages displayed at the *Tímarit.is* website. The fourth collection contains the entire text of one particular periodical, published in the 1830s and 1840s (9 issues), a total of approx. 300,000 running words. An overview of the amount of material and its distribution over time, both the extracted text and the texts already corrected or entered manually, is shown in Figure 2.

The figure displays the uneven distribution of the amount of extracted texts over the period, and indicates that there might still be need to add texts from certain decades, and diminish the text material from others, in order to further balance the corpus. It also shows that the amount of texts already prepared is unevenly distributed as well.

3.3. Correcting the OCR read texts

The digitized text from the newspapers and periodicals contains a number of errors that are a result of the optical character recognition (OCR) process. The quality of the OCR reading depends on the quality of the original document, both of the paper and the printing. When lead was used for printing, the printing faces got worn over time, and it also happened that the printing face of individual letters got damaged. In earlier times, when there were few printing presses in Iceland, the same letters were used for a long time, and printed material from the last part of the life of a particular set of letters can be almost unfit for OCR reading. The quality of the paper of old newspapers and periodicals is sometimes very poor and the pages may contain creases that disturb the quality of the OCR. The Icelandic alphabet contains a number of letters with diacritics (*á*, *é*, *í*, *ó*, *ú*, *ý* and *ö*), as well as the letters *ð*, *þ* and *æ*. A review of over one million manually corrected words in a large-scale digitization effort at the Icelandic parliament revealed that over 56% of all word errors involved the misrecognition of one of these characters (Daðason, 2012). Furthermore, the spelling of old texts is not standardized, as there was no agreement on a common standard for Icelandic orthography before the 20th century, and the earliest official standard was only put forth in 1918 (Jónsson, 1959). As previously mentioned (cf. section 3.1.), the correction of texts within the present project aims at a diplomatic version of the original.

When a decision was made to build a corpus based on OCR read texts from the *Tímarit.is* database, it was clear that relying solely on manual correction was infeasible due to the scale of the project. Instead, methods and software for automatically correcting digitization errors would have to be developed in order to compile a corpus of a decent size. An experiment was therefore carried out in 2010, based on the first selection of texts (cf. section 3.2. above). A portion of the selection, 24 issues (in total about 150,000 running words) from various newspapers and periodicals released throughout the time period, were manually corrected with comparison to the scanned images on the *Tímarit.is* website. At the same time, the development of the software, which is based on the principles for spelling correction, was initiated. The software uses frequency information and a lexicon derived from the corrected texts, as well as lists of word forms extracted from lexicographic historical archives and text collections at The Árni Magnússon Institute for Icelandic Studies (Daðason et al., 2014). Based on the manually corrected issues it was estimated that the digitized texts had an average word accuracy of about 91%. The uncorrected versions of the texts were run through the software and the result compared to the manually corrected versions. It was estimated that about 60–65% of digitization errors were corrected with the software. A considerable number of uncorrected issues from the text collection were run through the software, and the remaining errors in these issues were then corrected manually by students. At the end of this phase 51 issues from various newspapers and periodicals, a total of about 290,000 running words, had been corrected.

Previous work (Cushman et al., 1990) has shown that for correction of digitized text to be profitable it needs to have at least 98% character accuracy. If it is assumed that the mean number of characters per word is 5[10] and that character errors are evenly distributed this amounts to 90% word accuracy. Based on these figures it seems to be profitable to correct the digitized text in our case, especially after it has

[10] No figures are available for Icelandic, this is a pure assumption.

been run through the correction software.

Work on the material was continued in 2011, by manual correction of issues that had already been automatically corrected. Halfway through this phase the lexicon used by the software was updated with the vocabulary of the texts corrected so far. Additionally, the spellchecker was supplied with a list of corrections to known word errors, which was derived from the manually corrected texts. After that it was estimated that the software automatically corrected about 77% of all word errors in the scanned texts. The cycle was repeated by running new issues from the text collection through the improved software followed up by manual correction.

In the last phase of the correction project, carried out in 2012, the development of an interactive spellchecker was undertaken. This was meant to replace the process of first running the text through the correction software and then correcting the remaining errors manually. The software (Daðason et al., 2014) was based on the previous work on the automatic correction of digitization errors in our text material, and on other related projects (Daðason, 2012). As well as updating the lexicons the software was supplied with noisy channel functionality for spelling correction (Brill and Moore, 2000). The program, now called *Skrambi*, is a web application, and it accepts uncorrected OCR read text, underlining possible errors and suggesting corrections. The user is given 5 suggestions ordered by probability of being correct, as determined by the noisy channel model. The first suggestion is the correct one in 72% of occurrences, and the correct word is among the top five shown in 84% of occurrences. If none of the suggestions is correct the user can ignore them and make his own correction. An image of the original document is always accessible to the user.

An experiment was performed to test the efficiency of the software. First the time taken to manually correct the errors in 9 issues, each from a different newspaper or periodical from the text collection, was measured. The texts were not run through the correction software first. Then comparable issues (wrt. length and word accuracy) were selected, typically the next issue of the same titles, and these were corrected semi-automatically with the aid of *Skrambi*. The results of the time measurements indicate that correction of the OCR read text is on average three times faster with the help of *Skrambi* than correcting text which has not been run through the software. It was also found that the higher the word accuracy the greater the efficiency of the correction process becomes. This agrees with the results of (Cushman et al., 1990). This is valid for corrections both with and without the help of *Skrambi*. If the OCR read text has high word accuracy the increase in efficiency can be almost fivefold.

3.4. Present state of the project and remaining tasks

To date the text in 188 issues of 28 newspapers and periodicals from the period 1870 to 1920, in total about 1.4 million running words, has been run through the correction process. All issues have been checked manually. The corrected texts are already available for search in the 'Icelandic Text Collection' (*Íslenskt textasafn*) at the website of The Árni Magnússon Institute for Icelandic Studies (under the heading "Blöð_og_tímarit_1860-1930" (Newspapers and periodicals 1860-1930)).[11] The search in this database is limited to strings of letters. The interface allows the user to enter the lemma (base form) of a word (or two adjacent words), and with linking to the 'Database of Modern Icelandic Inflection' (Bjarnadóttir, 2012) the search program seeks out all inflectional forms of the selected word(s), though only with the modern standard spelling. The user can search for all these forms, or deselect forms that he or she chooses to disregard. The texts are useful for various tasks, e.g. in lexicology and lexicography.

For a more focused search in the 19th and early 20th century texts, that could find all possible spelling variants of words and word forms, as well as allow the possibility of searching for grammatical features, a second, standardized layer of tagged texts would be needed, as described in 3.1. above. The next stage in the development of the corpus is therefore to add such information to the texts.

As previously mentioned, the spelling in the newspapers and periodicals, intended for the corpus, differs from the modern spelling norm in various ways, and as the orthography was not standardized in the period, there occur many variants of the same word form. This has consequences both for searching the texts, and for the application of available language technological tools, such as the tools that have been developed for the tagging of Modern Icelandic text (Loftsson, 2008). Rather than adapting the tagging tools to the different spelling variants appearing in Icelandic texts from the 19th and early 20th centuries, we will map the texts to the modern Icelandic spelling standard. The software *Skrambi*, already mentioned in connection with the correction of digitizing errors, will be used for this purpose. *Skrambi* has evolved into a multipurpose software that can be adapted to various tasks, such as spelling correction, the correction of digitized text, and for mapping text between different spelling variants. This is achieved by providing the software with different lexicons. For the mapping between a diplomatic version of 19th century text and a version with modern spelling, a lexicon based on the 'Database for Modern Icelandic Inflection' (Bjarnadóttir, 2012), together with lexicons based on the old texts already corrected, and other language data from the same period, will be used. If time and resources permit, the mapping will be performed semi-automatically with an interactive version of *Skrambi* which works in a similar way to the correction of OCR errors (cf. 3.3 above). After mapping to the modern spelling norm, the texts will be tagged with the system used for tagging the 'Tagged Icelandic Corpus' (*MÍM*) (Helgadóttir et al., 2012). The original, corrected diplomatic text from the newspapers and periodicals, and the tagged text with standardized modern spelling will form a parallel corpus. A search interface will be developed for this corpus using the *Glossa* system (Johannessen et al., 2008) which offers possibilities for search in multilingual corpora. The *Glossa* system has already been adapted to

[11] Cf. `http://corpus.arnastofnun.is/`; there are a few texts from the 1860s and 1920s, which explains the heading.

modern Icelandic for *MÍM*.[12]

Text from early newspapers and periodicals, amounting to about 8.8 million running words, has been extracted from the *Tímarit.is* database in several stages (cf. section 3.2.). Of these, close to 1.6 million running words have been corrected, or in a few cases entered manually. An overview of prepared texts is given in Table 1.

Corpus of early Modern Icelandic			
	Number of titles	Number of issues	Running words
1800-1820	3	7	78,427
1821-1840	5	21	327,357
1841-1860	2	13	146,439
1861-1880	7	29	71,045
1881-1900	17	92	364,124
1901-1920	10	65	579,427
Total	**—**	**227**	**1,566,819**

Table 1: An overview of the amount of texts already prepared for the Corpus of early Modern Icelandic (1800-1920) and their distribution over the period (the texts in the table have either been manually and/or automatically corrected, or manually entered)

A considerable number of texts from the first part of the 19th century has already been prepared, either by running it through the correction procedure (as part of a separate project; (Daðason et al., 2014)) or by manually entering the text (cf. section 3.2.). A few more issues from this early period will be run through the semi-automatic procedure in the next few months. The corrected or manually entered texts from the early part of the period will then be used to adapt the software and increase its effectiveness with respect to the oldest texts. After that the multipurpose program *Skrambi* will be applied to automatically correct the remaining texts and then map them to a standardized modern Icelandic version. After that these texts will be tagged automatically with the available tools. In the absence of any manual corrections, there will surely remain errors in the corrected texts and they will cause further errors in the mapping which again will cause errors in the tagging. It is, however, anticipated that search in these texts will give better results than search in the uncorrected OCR read texts available on the website *Tímarit.is*, and even the corrected but untagged texts in the 'Icelandic Text Collection'. The automatically corrected texts, which are an approximation to a diplomatic version, and the mapped and tagged texts will be a separate part of the intended Corpus of early Modern Icelandic.

4. Conclusions

In the paper, we have described and discussed an attempt to build a corpus of early Modern Icelandic, intended mainly to serve the needs of the linguistic and lexicographical research community, as economically as possible with respect to time and money. We have succeeded to lay the foundations of such a corpus, by using a selection of texts already digitized with OCR methods. This raw material is, however, deficient as it has not been controlled or corrected in any way. The need for a more efficient resource arose in connection to a number of ongoing research projects and the task of constructing a new corpus was undertaken by a group of linguists and specialists in language technology at The Árni Magnússon Institute for Icelandic Studies. Their task was to select the text material, develop methods and tools to correct such material automatically, and at the same time keeping the unstandardized and variant spelling and word forms of the original. This diplomatic text version then needs to be converted to a parallel version with modern standardized text, to enable the application of tools for text analysis, i.e. for automatic tagging of grammatical features and lemmatization. Both versions are then to be presented as a parallel corpus in an easily accessible database with effective and flexible search possibilities. Even if the project is not completely finished, we consider it to be advanced enough to show that the procedures applied have been successful, and we foresee that it will only have taken about four years, with very limited financial resources, to build a useful language corpus with 19th and early 20th century Icelandic.

5. Acknowledgements

Work on the corpus and the development of tools for correcting and mapping text has been a collaboration between linguists, lexicographers, and specialists in language technology at The Árni Magnússon Institute for Icelandic Studies and The University of Iceland, and a number of student assistants have worked on the preparation of texts and the development of methods and tools. Apart from the authors, who have been responsible for data selection and involved in the overall organization and supervision of the project, the following persons should be mentioned: Kristín Bjarnadóttir supervised much of the work on corrections, error analysis and development of methods, esp. in the early stages, as well as organizing an independent project which has enriched the corpus building, both as regards the contents and methods for the automatic mapping of a diplomatic text and a standardized version; Jón Friðrik Daðason started out as a student assistant for developing methods for (semi-)automatic corrections of OCR read texts, but carried on and made use of his experience in the project a strand of his MS-thesis, of which the *Skrambi*-program was a product (Daðason 2012); Kristján Rúnarsson, Hjördís Stefánsdóttir, Guðrún Línberg Guðjónsdóttir and Kristján Friðbjörn Sigurðsson worked on the manual and semi-automatic corrections of the texts at various stages of the project; Örn Hrafnkelsson and Kristinn Sigurðsson at The National and University Library of Iceland have provided the OCR read texts from the *Tímarit.is* archives.

A number of tools and resources used for the construction of the corpus, have been compiled and developed in previous projects. A great number of scholars and students have taken part in these and we want to acknowledge their contribution to the present project, even if it is impossible to list all their names.

The construction of the corpus has been connected to a

[12] Cf. `http://mim.arnastofnun.is/`.

number of ongoing research projects, and part of their funding has gone into the building of this resource: *The strengthening of the text archives at The Árni Magnússon Insitute* (Guðrún Kvaran and Sigrún Helgadóttir) in connection with the compilation of a dictionary of 19th and 20th century loanwords and *Foreign influence in late 19th and 20th century Icelandic* (Ásta Svavarsdóttir), both funded by grants from the University of Iceland Research Fund, as well as the project *Language Change and Linguistic Variation in 19th-Century Icelandic and the Emergence of a National Standard*, supported by the Icelandic Reasearch Fund (grant nr. 120646021/2/3 2012-14; PI: Ásta Svavarsdóttir). Jón Friðrik Daðason got a grant from the the Icelandic Student Innovation Fund 2011 for his program development, and in 2012 the same fund gave support to the project *Fjölnir fyrir hvern mann*, aimed at the correction of a particularily complicated text of one periodical, as well as the development of methods for automatic standardization and mapping between versions. Besides, the corpus construction has benefited from financial support from the Directory of Labour, in the form of summer wages to student assistants working on particular tasks within the project.

6. References

Bjarnadóttir, K. (2012). The Database of Modern Icelandic Inflection. In *Proceedings of the workshop Language Technology for Normalization of Less-Resourced Languages, SaLTMiL 8 – AfLaT, LREC 2012*, pages 13–18, Istanbul, Turkey.

Brill, E. and Moore, R. C. (2000). An improved error model for noisy channel spelling correction. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, Hong Kong.

Cushman, W. H., Ojha, P. S., and Daniels, C. M. (1990). Usable OCR: what are the minimum performance requirements? In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 145–152.

Daðason, J. F., Bjarnadóttir, K., and Rúnarsson, K. (2014). The Journal *Fjölnir* for Everyone: The Post-Processing of Historical OCR Texts. In *Proceedings of the workshop Language resources and technologies for processing and linking historical documents and archives-Deploying Linked Open Data in Cultural Heritage, LRT4HDA, LREC 2014*, Reykjavík, Iceland.

Daðason, J. F. (2012). Post-Correction of Icelandic OCR Text. MS thesis at University of Iceland, http://hdl.handle.net/1946/12085.

Helgadóttir, S., Svavarsdóttir, Á., Rögnvaldsson, E., Bjarnadóttir, K., and Loftsson, H. (2012). The Tagged Icelandic Corpus (MIM). In *Proceedings of the workshop Language Technology for Normalization of Less-Resourced Languages, SaLTMiL 8 – AfLaT, LREC 2012*, pages 67–72, Istanbul, Turkey.

Hrafnkelsson, Ö. and Sævarsson, J. (2014). Digital libraries of historical Icelandic newspapers, periodicals, magazines and old printed books. In *Proceedings of the workshop Language resources and technologies for processing and linking historical documents and archives-Deploying Linked Open Data in Cultural Heritage, LRT4HDA, LREC 2014.*, Reykjavík, Iceland.

Johannessen, J. B., Nygaard, L., Priestley, J., and Nøklestad, A. (2008). Glossa: a Multilingual, Multimodal, Configurable User Interface. In *Proceedings of LREC 2008*, pages 617–621, Marrakesh, Morocco.

Jónsson, J. A. (1959). Ágrip af sögu íslenzkrar stafsetningar. [An overview over the history of Icelandic orthography.]. *Íslenzk tunga/Lingua Islandica*, 1:71–119.

Loftsson, H. (2008). Tagging Icelandic text: A linguistic rule-based approach. *Nordic Journal of Linguistics*, 31(1):47–72.

Rögnvaldsson, E. and Helgadóttir, S. (2011). Morphosyntactic Tagging of Old Icelandic Texts and Its Use in Studying Syntactic Variation and Change. In Sporleder, C., van den Bosch, A. P. J., and Zervanou, K. A., editors, *Language Technology for Cultural Heritage: Selected Papers from the LaTeCH Workshop Series*, pages 63–76. Springer, Berlin.

Rögnvaldsson, E., Ingason, A. K., Sigurðsson, E. F., and Wallenberg, J. (2012). The Icelandic Parsed Historical Corpus (IcePaHC). In *Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012*, Istanbul, Turkey.

Named Entities in Court: The MarineLives Corpus

Dominique Ritze,[1] Cäcilia Zirn,[1] Colin Greenstreet,[2] Kai Eckert,[1] Simone Paolo Ponzetto[1]
[1] Research Group Data and Web Science, University of Mannheim; [2] MarineLives
{dominique,caecilia,kai,simone}@informatik.uni-mannheim.de, colin.greenstreet@googlemail.com

Abstract

In this paper, we introduce the MarineLives corpus. This consists of a collection of manually-transcribed historical records of the English High Court of Admiralty between 1650 and 1669. The transcriptions are obtained collaboratively in an open and transparent process, and are made freely available for the research community. We conduct first experiments with off-the-shelf state-of-the-art Natural Language Processing (NLP) tools to extract named entities from this corpus. In particular, we investigate to what degree the historical language and the highly specific domain of the document affects the results. We find that non-trivial challenges lie ahead and that domain-specific approaches are needed to improve the extraction results.

Keywords: Historical Texts, Named Entity Recognition

1. Introduction

Transcribed historical texts are invaluable resources in the digital humanities, where interdisciplinary approaches from computer science, linguistics and the humanities can be applied to gain new insights. For example data visualization techniques can be used to graphically represent timelines or graphs of traveling routes encoded within the texts (Strötgen and Gertz, 2012). Automatically detected cross-references between texts based on common named entities, or simply by means of location and time, can help scholars to take additional sources outside of his or her focused research topic into account. Another possibility is the creation of orthogonal databases like HISKLID (Glaser and Riemann, 2009), where accounts on weather information are gathered from any kind of historical texts. While providing a gold-mine of information, historical texts also pose many challenges for Natural Language Processing (NLP) research (Piotrowski, 2012). For example, most of existing state-of-the-art tools and approaches have been primarily developed for English newswire texts whose characteristics differ a lot from historical texts.

In this paper, we introduce the MarineLives corpus, consisting of transcriptions of records of the English High Court of Admiralty from the 17th century. In order to better understand the challenges and problems posed by such highly-specific domain text, we test state-of-the-art approaches for Named Entity Recognition (NER), in order to be able to automatically detect NEs on these historical data. Perhaps unsurprisingly, our results indicate that off-the-shelf NERs are not suitable for the task at hand, and call for either training statistical model on in-domain data, or the application of automatic domain adaptation techniques.

2. The MarineLives Project

MarineLives[1] is a project for the collaborative transcription, linkage and enrichment of primary manuscripts, which were originated in the English Admirality Court, London. This Court was a civil law court, which dealt with marine commercial disputes. The Courts records are preserved in the English National Archives[2].

The collaboration was formed in September 2012 by several academics and enthusiastic practitioners around Colin Greenstreet and Jill Wilcox (Greenstreet, 2012). Today, many more volunteers and academic partners from different institutions, countries, and domains contribute to and follow the project. Recently, MarineLives joined the project Digitised Manuscripts to Europeana (DM2E)[3], which not only connects its transcribed texts to other historical documents, but also will lead to the representation of the texts in the Europeana digital library[4].

The MarineLives project focuses on records produced in the period 1650–1669. Its long term goal is a semi-diplomatic full text transcription of all court records in this period, together with related metadata. The text will be enriched through collaborative, but mediated, annotation of a web available version of the edited full text transcription. The planned corpus will exceed 20 million words. The text, together with its annotations, will be made publicly available in stages under a CC by 3.0 license.

The completed and edited transcription will be a resource for social, material, and marine historians, who have previously lacked a full text transcription, as well as for corpus linguists. Several calendars of thematic Admiralty court material exist (Appleby, 1992; Murphy, 2011), but there has been no prior scholarly full text transcription for the 17th century records of the Court. It is anticipated that research historians, who use the MarineLives corpus in combination with specialized NLP tools, will be able to ask new questions, as well as to improve on prior answers to known questions. The extraction of weather related events serves as a good example, another one is the extraction and visualization of point to point transportation routes.

[1] http://www.marinelives.org; http://www.hastac.org/organizations/marinelives

[2] The National Archives, Kew, Richmond, Surrey TW9 3DU, United Kingdom. Document classmark: HCA. http://www.nationalarchives.gov.uk/records/research-guides/high-court-admiralty.htm

[3] http://dm2e.eu

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

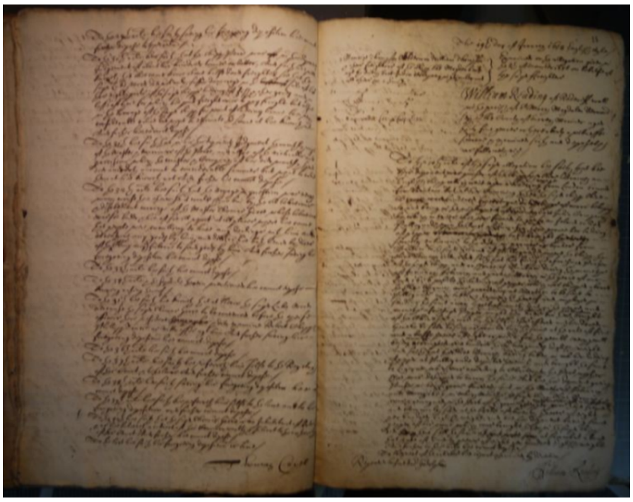


Figure 1: Admiralty Court deposition book, 1658-61 (HCA 13/73 f.14v & 15r). [6]

3. Corpus

In the following, we provide a brief introduction into the historical background of the corpus, the transcription process, as well as some general statistics.

3.1. Historical Background and General Corpus Characteristics

The 17th century English Admiralty Court generated a large volume of manuscript pages each year to record and support its activities. These documents fall into a number of record types, which relate to the Admiralty Court legal process. In rough chronological order, an action or cause is initiated by the lodging of a claim in the Court. It is followed by libel, which again might be lodged by one or both parties to a dispute. That libel might then be converted into one or more allegations, which contain articles to be addressed by the parties to the dispute. In addition to the articles of the allegations, interrogatories are also prepared and put to witnesses. Testimonies are recorded in the form of depositions. However, most causes started in the Admiralty Court were discontinued, or resolved, before they reached the stage of taking witness depositions. The stages in the legal process were recorded in the Act Books of the Court, and can be used to reconstruct the progression of individual cases across multiple record types contained in different physical volumes and document boxes.

The MarineLives corpus is currently formed from the document series HCA 13. Documents in this series take the physical form of large leather bound volumes of 600–750 folios. Facing verso and recto pages are illustrated from the volume HCA 13/73, which covers the period late 1658 to early 1661 (Figure 1). Other record types will be added to the corpus later in the project, and individual documents will be linked by the project team across record types to enable users to access all documentation for a given case.

Volumes in the document series HCA 13 contain written manuscript records created from oral responses to written questions. Witnesses (termed deponents) were invited to the Doctors' Commons in London, where the Court met. There they met with Court proctors (civil lawyers) assigned to the litigants and their respective witnesses, and the proctors reviewed the written allegations and interrogatories with the witnesses. The depositions of the witnesses were written down by notary publics in the presence of both the witnesses and the Court proctors. They were subsequently signed or acknowledged with a mark, by the deponent, and read (or repeated) into the Court record.

These written records are grouped in quires, which were subsequently bound in volumes. Manuscript pages within a quire are chronological, but successive quires are not strictly chronological. The layout of an individual deposition is standardized, though depositions themselves vary significantly in length, between half a manuscript page and twenty-five pages. The standard layout consists of front matter (date of the deposition; name of case in long or short form; and the name of the deponent, his or her occupation, place of residence, and approximate age), followed by responses to the articles contained in the allegations in numbered paragraph form, and then by answers to interrogatories and cross-interrogatories, again in numbered paragraph form.

3.2. Transcription

The transcription which forms the basis of the MarineLives corpus was created through defined processes of transcription and editing, following a defined set of guidelines, developed by the project team with the assistance of Dr. Charlene Eska (Virginia Tech). The guiding principle has been to create a semi-diplomatic transcription, without reproducing most of the layout of the original manuscript. In the final public edition, transcriptions will be made available close to high resolution digital images of the related pages. Orthographical variation in the original has been reproduced in the transcription, with significant variation in spelling and capitalisation between scribes, and even within a single page produced by just one scribe. Punctuation is minimal and inconsistent in the original. Although the punctuation has been reproduced as seen, it is of little use in machine processing the text. Grammar is largely consistent across scribes.

Prior to transcription, each target volume was digitally imaged and the images were loaded into a tailored version of the open access transcription software Scripto[7]. One volume in the corpus (HCA 13/71) was created through a collaborative effort involving 30 transcribers, grouped in teams of three to four, and supported by team facilitators.

[6] MarineLives transcription HCA 13/73 f.14v: http://annotatehca1373.wikispot.org/HCA_13/73_f.14v_Annotate and HCA 13/73 f.15r: http://annotatehca1373.wikispot.org/HCA_13/73_f.15r_Annotate. Crown copyright image reproduced courtesy of The National Archives.

[7] The software (http://scripto.org/) was developed with the support of an National Endowment for the Humanities grant by the Roy Rosenzweig Centre at George Mason University, Virginia. Giovanni Colavizza (now at École Polytechnique Fédérale de Lausanne) tailored the software to meet the specific requirements of the MarineLives project (MarineLives-Transcript, http://marinelives-transcript.org/scripto/).

Volume	Page count	Word count	Words/page
HCA 13/71	342	138,332	404 w/p
HCA 13/72	1,078	477,822	443 w/p
HCA 13/73	434	208,919	481 w/p

Table 1: Corpus statistics.

The transcriptions were then edited by the team facilitators and by a chief editor. Two further volumes (HCA 13/72 and HCA 13/73) have been transcribed by an individual transcriber, and subsequently edited.

Table 1 lists the number of pages transcribed in above mentioned corpora, as well as the number of words. It can be seen that the number of words per pages differs slightly, over all three volumes, a typical manuscript page has around 445 words.

4. Recognizing persons, locations and ships

There are a variety of applications in the domain of history or sociology the MarineLives data could be used for. For a historian, it might be interesting to analyze which route a ship was taking, and who was on board. Other scholars might be interested in the persons and organizations involved in a case and information about them. To most of those applications, named entities such as persons, locations, organizations and ships are essential. To extract them, we apply automated named entity recognition on the data. There are pre-built off-the-shelf named entity recognizers available, however, they are usually trained on modern language. The MarineLives corpus on the other hand covers 17th century English and a very particular domain.

Rayson et al. (2007) analyzed the limitations of applying state-of-the-art natural language processing tools to non-modern texts by running a Part-of-Speech tagger (CLAWS, developed by Garside and Smith (1997)) to Shakespeare texts. While the tagger achieves an accuracy of 96% on modern English texts, only 81% accuracy can be achieved when applied on the ancient texts.

In Grover et al. (2008), the authors build a prototype named entity recognizer for persons and places in British parliamentary proceedings from the late 17th and early 19th century. It is based on lexicons and hand-crafted rules especially customized to find monarchs, earls etc. An example is “earl of [capitalized word]”. They achieve precision and recall values around 70%.

In this paper, we run a small experiment to investigate how well an off-the-shelf named entity recognizer pre-trained on modern English performs on the MarineLives corpus. Our purpose is to investigate the limitations of already existing methods and the need to adapt approaches to the properties of the MarineLives corpus.

4.1. Experiments with a state-of-the-art named entity recognizer

For our experiments, we use the HCA-13/72 data[8]. Within this part of the data, the most salient persons and ships were marked by the transcribers by HTML tags. We make use of those manual annotations as gold standard data. Ships are marked by emphasis tags: <em>ship</em>, persons by strong tags : <strong>person</strong>. Locations and other entities are not annotated. Furthermore, the annotations of persons and ships are not complete. This is due to the fact that the annotations were originally intended as an emphasis for humans and not as gold standard. The strong tags are used to mark other phenomena as well, for example abbreviations that serve as paragraph headers, e.g. <strong>Rp. .j</strong>. We removed them from the experiment by a simple heuristic: if a complete line is surrounded by strong tags and does not contain more than four tokens, it is ignored. Within the whole data set, there are 5140 manually annotated instances, thereof 1350 being persons and 3790 being ships.

In our experiments, we use the Stanford Named Entity Recognizer (NER) (Finkel et al., 2005) which is a state-of-the art tool. It uses conditional random fields combining distributional similarity based features. The model we chose was trained on the CoNLL 2003 shared task English dataset (Tjong Kim Sang and De Meulder, 2003). It recognizes four classes of named entities: persons, locations, organizations and miscellaneous, whereof the latter class covers all type of named entities that do not fit into any of the other classes.

Applying the NER to the data, it returns 14762 persons, 6070 locations, 8896 organizations and 3109 entities of the miscellaneous class. The detected entities have been evaluated to see whether the NER is able to properly recognize the named entities. For two reasons, it is not possible to calculate precision and recall of the results straightforward. First, as mentioned before, the manual annotations in the gold standard are not complete: not all mentions of ships and persons are marked, and locations as well as organizations have not been annotated at all. Second, it is not easy to directly compare the output of the NER to the manual annotations: names (of persons as well as ships) often consist of several tokens, for example *William Smith* or the ship *Mary and Joyce*. The NER does not detect the beginning and end of named entities, but only classifies single tokens: *William*/person *Smith*/person. Another challenge is the classification of ships - do they count as locations, or should they rather be classified into miscellaneous? We evaluate the performance in the following way:

1) Persons. To calculate the recall of the recognized persons, we counted how many of the manually annotated persons of the gold standard were also found by the NER. For example if *William Smith* is listed in the gold standard, and the NER tagged *William* as well as *Smith* as person, we counted that as one match. For the precision, we asked human annotators to read the sentence and to count how many of the tokens which are tagged as person by the NER are persons indeed.

2) Locations. As we do not have a gold standard for locations, we did not compute the recall. The precision was calculated as described for persons: human annotators were asked to count the correct decisions of the NER regarding locations.

[8] http://annotatehca1372.wikispot.org/

	Precision	Recall
Persons	76.8 %	68.0%
Locations	77.0%	–

Table 2: Results of the Named Entity Recognition for persons and locations.

3) Organizations and Miscellaneous. We abstain from evaluating the detected organizations and miscellaneous as they are not the main focus of our use cases and they are difficult to be evaluated for non-expert human annotators.

4) Ships. Ships were evaluated in a special way, as our NER neither has a particular class for them nor classifies all of them consistently into the same class. We asked the human annotators to check for all ships contained by the gold standard and to state whether they were (N)ot found at all, found as (P)ersons, as (L)ocations, as (O)rganizations or as (M)iscellaneous or as (S)everal of those classes. The latter might apply to ship names consisting of several tokens.

Table 2 shows the results for persons and locations. The precision for recognizing persons and locations was surprisingly high, considering that in Grover et al. (2008) the hand-crafted rules achieved a comparable precision only. However, the performance of the NER on modern data is up to 86% F-measure. This suggests that if we adapt the NER to the MarineLives domain by training it on particular data, we might achieve even better results. The same applies to the recall: we are only able to retrieve 68% of the gold standard persons. The NER achieves a comparable precision for locations. As mentioned above, we did not evaluate recall, but we assume a similar recall as for persons. One of the challenges of the data consists in the spelling variations: a name, even when referring to the same entity, might be spelled with different variations even within the same page, for example *Katherine / Catherine* or *Lisbon / Lisborne / Lisbone*. As a consequence, gazetteers that just match names against a pre-compiled list will not perform successfully on this task: it would be impossible to create a list containing all possible (ancient) spelling variations. We expect better results from a NER that is based on a trained model which makes use of context features.

While the NER achieved at least a decent performance on recognizing persons and locations, it was barely able to recognize any ships, as can be seen in Table 3. About a third of the ships marked in the gold standard were not found at all. 20% were each found as persons and organizations, the rest as locations, miscellaneous or as a combination of two classes. The challenge in recognizing ships seems to lie in the variety in type of ship names: ships might be named after persons (*Elisabeth*), after abstract nouns (*success*, *hope*, *fortune*), or any other type, like *trade increase* or *six brothers*. Even more difficulties are posed by names consisting of whole phrases, like *Mary and Joyce* or *Nostra Seniora da Rosario*. Here it is not only necessary to detect the existence of a ship but also where the name ends. Furthermore, they are not capitalized consistently.

	N	P	L	O	M	S
#	93	61	19	61	38	13
%	32.6%	21.4%	6.7%	21.4%	13.3%	10.63

Table 3: Results of the Named Entity Recognition for Ships (#: absolute numbers; %: percentage).

Having a look at the classification results, we noticed that in the corpus there are many phrases such as *Richard Megin of Ratcliff*, which the NER recognizes as person. For such cases, before creating a gold standard it has to be decided whether the whole phrase refers to the person, or whether the last part consists of a location.

4.2. Analysis of ship mentions

Since the recognition of ships with a standard NER is not sufficiently robust, we further analyzed whether ships can be found by applying simple pattern techniques. When skimming through the transcribed manuscripts, it strikes that the word "ship" or "shipp" often occurs before the mention of a ship, e.g. "the shipp Catherine", "on the ship the Anna and Mary", "the shipp called the Recoverie". In some cases, one or more words are between the word "ship" and the ship name.

From the 3790 manually annotated ships, 1076 ones match the following pattern: *ship[p] [w1] [w2]* ***name***, where the alternative spelling variation of ship as well as the occurrence of maximal two words in between are considered. For the remaining ship mentions, it is quite difficult to find any pattern, since they often just occur in the text without any characteristic surrounding words. Applying the pattern, we can determine that a ship name will probably start within the next three words. However, it is not clear where the name actually starts and especially where it ends. A simple heuristic is to assume that ship names always begin with a capital letter. Identifying the end of a ship name is more difficult. Of all manually annotated ships, 2014 ship names consist of 1 word, 1054 of 2 words, 531 of 3 words and the remaining 101 ones of more than 3 words. For a first experiment, we focus on names that only consist of one word. From the 1076 manually annotated ship names that can be found by the pattern, 456 only consist of one word and start with a capital letter. If we apply this refined pattern to the whole text, we find 1320 potential ship names. Thus, a lot of incorrect names are detected, especially because the word "ship" does not only occur in combination with a ship name. Another reason is the irregular capitalization of words, e.g. "and" often starts with a capital letter. Examples of incorrectly found names are the following ones: "sayd shipp And further", "the shipp Interrogate called the Elizabeth". This first experiment shows that a pattern-based approach including a learning component could be a possible strategy to recognize ship names but especially the general use of characteristic words like "ship", the irregular capitalization of words as well as different lengths of ship names raise new challenges.

5. Conclusions

In this paper, we introduced the MarineLives corpus, and presented some initial experimental work on applying NLP tools for Named Entity Recognition to these historical data. The MarineLives corpus is created collaboratively and grows constantly. Besides, the manuscripts of the English Admiralty Court provide fascinating details of the marine life in the 17th century. Finally, the whole corpus of depositions uses a relatively consistent structure and grammatical style. These are all factors which make it a suitable, yet challenging textual data resource for the growing field of applications of NLP techniques to historical texts.

Our preliminary results on NER are mixed at best. By using a state-of-the-art NER tagger, trained on standard newswire text, we were able to obtain reasonable results for the automatic identification of persons and locations. On the other hand, the results also indicate clear challenges such as the automatic identification of ships, which are an obviously predominant kind of named entities in these texts, although rather uncommon in general text corpora. Our first experiments with a pattern-based approach suggest that the accuracy can be greatly improved, especially when more elaborated techniques are used. Possible directions for future work include either to concentrate on further annotation efforts, in order to provide labeled, in-domain data for retraining standard statistical models, or to explore domain adaptation techniques.

Regardless of the specific NLP techniques to be explored in the near future, our long-term vision is to provide the NLP community with a freely available, high-quality historical corpus to promote further research in the challenging field of NLP for historical data. We believe, in fact, that these kinds of data not only show the limitations of existing tools and techniques, but also call for the development of lightweight, easy-to-adapt methods that do not merely rely on standard newswire data, but rather address the complexity of fine-grained and specialized domains such as those provided by heterogeneous historical texts.

6. Acknowledgments

This work was supported in part by the EU funded project DM2E Digitised Manuscripts to Europeana (Ref. No. ICT-PSP-297274). We thank our annotators Orphee De Clercq, Daniel Fleischhacker, Jan Noessner, and Jörg Schönfisch for helping us with the evaluation of the Named Entity Recognition tagger.

7. References

Appleby, J. C. (1992). *A Calendar of material relating to Ireland from the High Court of Admiralty Examinations, 1536-1641*. Irish Manuscripts Commission, Dublin.

Finkel, J. R., Grenager, T., and Manning, C. (2005). Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 363–370, Stroudsburg, PA, USA. Association for Computational Linguistics.

Garside, R. and Smith, N. (1997). A hybrid grammatical tagger: CLAWS4. In Garside, R., Leech, G., and McEnery, T., editors, *Corpus Annotation: Linguistic Information from Computer Text Corpora*, pages 102–121. Longman.

Glaser, R. and Riemann, D. (2009). A thousand-year record of temperature variations for Germany and Central Europe based on documentary data. *Journal of Quaternary Science*, 24(5):437–449.

Greenstreet, C. (2012). On the Crest of a Wave. *History Today*, 62(9).

Grover, C., Givon, S., Tobin, R., and Ball, J. (2008). Named Entity Recognition for Digitised Historical Texts. In *LREC*. European Language Resources Association.

Murphy, E., editor. (2011). *A calendar of material relating to Ireland from the High Court of Admiralty, 1641–1660*. Irish Manuscripts Commission, Dublin.

Piotrowski, M. (2012). *Natural Language Processing for Historical Texts*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

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 abstract. In *Corpus Linguistics Conference*.

Strötgen, J. and Gertz, M. (2012). Event-centric Search and Exploration in Document Collections. In *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 223–232.

Tjong Kim Sang, E. F. and De Meulder, F. (2003). Introduction to the CoNLL-2003 Shared Task: Language-independent Named Entity Recognition. In *Proceedings of CoNLL-03*.

Building Less Fragmentary Cuneiform Corpora: Challenges and Steps Toward Solutions

Stephen Tyndall

University of Michigan
Lorch Hall, Tappan Street, Ann Arbor, Michigan
styndall@umich.edu

Abstract

This paper examines the particular problems of connecting fragmentary cuneiform material into larger corpora. There are several distinct layers of challenges in this kind of corpus connection project at the orthographic, morphological, and full-text levels of textual analysis. This paper will present these challenges and lay out a pathway by which they may be overcome. Further, specific techniques, particularly the creation of morphological schemata for eventual use as a preprocessing step for classification and clustering, are discussed. These challenges and techniques are presented with respect to their applicability and utility in building these corpora, with the eventual goal of creating fuller, more connected, and more useful cuneiform corpora for scholarly work.

Keywords: cuneiform, morphology, orthography, corpus, classification, clustering

1. Introduction and Purpose

This paper lays out some of the unique challenges involved in the creation and connection of text corpora of cuneiform languages. Because of their position at the dawn of literate history, the cuneiform languages provide an important view into the past, but the age of the material, and the consequent fragmentary nature of the extant corpora, make interpretation notoriously difficult. Indeed, connecting disparate fragments of cuneiform tablets into larger, more complete texts is one of the primary tasks of cuneiform studies. Horst Klengel succinctly summarizes some of the major difficulties of the cuneiform Hittite corpus, difficulties shared by the other two major languages of cuneiform scholarship, Akkadian and Sumerian.

> “Some general problems, affecting both philologists and historians, are caused by the Hittite textual tradition itself. First, the bulk of the cuneiform material is fragmentary. The tablets, discovered in various depots in the Hittite capital and in some provincial centers, normally were of a larger size. When the archives were destroyed, the tablets for the most part broke into many pieces. Therefore, the joining of fragments became an important prerequisite for interpretation” (Klengel, 2002).

Drawing connections between fragmentary elements of larger corpora has also been a longstanding problem of classical philology and ancient studies. In Ancient Greek philology, the recent discovery of ancient papyrus fragments has

> “offered new problems, since in many cases, where no attributed ancient quotation comes to help, it is difficult to decide from which book the different scraps, often rather badly preserved, came. Moreoever, even when part of a new papyrus text coincides with an attributed ancient quotation, or with an already attributed papyrus text, the solution might not be so easy. Papyrus fragments are normally grouped according to the appearance of the handwriting, and it was possible that the same scribe wrote more than one book of a single author, or even more than a single author” (D’Alessio, 1997).

The case of cuneiform corpora is yet more complicated than that of Greek, since no cuneiform writing practice survived beyond the first century CE, leaving the writing system, and the cultures recorded with it, lost to history until the 19th century.

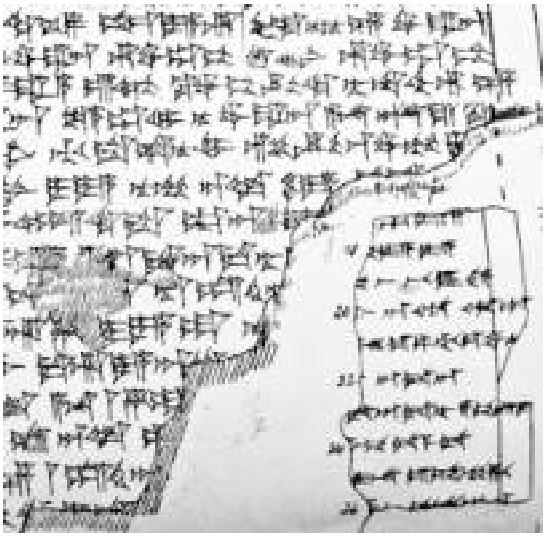


Figure 1: Hand Copy of Cuneiform Tablet, with Join

Cuneiform texts are typically published in one of two ways. Hand-drawn copies on paper are one method, with the drawings representing the actual tablets and the cuneiform signs closely adhering to the position on the structure of the tablet itself. One such copy can be seen in Figure 1. The other is transliterated texts, with the disambiguations already performed by an editor. Transliterated editions are typically only created for larger texts, composed of many distinct fragments. Cuneiform texts are typically composed of many distinct tablet fragments, with gaps from the disintegration of the pieces of the clay tablets themselves. The process of text assembly is mostly haphazard. Scholars reading tablets notice that the language or thematic material of a fragment seems similar to that of another, and proposed joins are built by cutting out a hand copy of one fragment and trying to place it against another.

Some amount of work exists in the digital representation and publishing of cuneiform tablets. Anderson and Levoy presented a method for representing tablet fragments in three dimensions and then 'unwrapping' the 3D model for printing purposes (Anderson and Levoy, 2002). Others have attempted cheaper and more portable solutions to digitization (Willems et al., 2005). Some attempts have been made to help with character recognition and disambiguation, both by machine, using 3D models (Mara et al., 2010), and through crowd-sourcing (Nurmikko et al., 2012). These tools, with improvement, may provide considerable improvement in the speed and accuracy with which cuneiform tablet fragments are published.

Further challenges are posed by the ambiguous nature of the cuneiform system itself. Cuneiform writing is logosyllabic in nature with some signs representing whole words, some signs representing syllables (typically CV, CVC, or VC), and some signs having different logographic readings based on context. In order to represent the text correctly, the correct readings must be discovered. Further, most cuneiform languages have a significant degree of morphological complexity - Sumerian has a very complex verb template, with slots for eight morphemes; Akkadian shows root-pattern ablaut morphology typical of Semitic languages; and Hittite shows a mix of ablaut processes and suffix morphology. Lemmatizing word forms is therefor a non-trivial task.

Thus, the sequence of events for the eventual linking of fragments into larger texts must be:

1. Transcribe and transliterate tablet fragments.
2. Resolve ambiguous signs into their correct readings
3. Lemmatize the forms in the corpus according to language-specific principles
4. Connect fragments with each other and with already-identified larger texts.

It is the second and third steps with which this paper is primarily concerned. Below I discuss a few rules for ambiguous sign resolution, and then a technique for morphological analysis and lemmatizing, one that can see through certain kinds of intraparadigmatic morphophonemic variation, and then the potential use of these processes in classification and clustering of fragments.

2. Cuneiform Languages and Their Current Corpora

Cuneiform writing was developed first for Sumerian, a language isolate spoken and written in modern-day Iraq between ca. 3500 BCE and 2500 BCE. The writing system was adapted for a number of other Ancient Near Eastern languages, including Akkadian, a Semitic language; Hittite, the earliest-attested Indo-European language; Hurrian, the language of the Mitanni empire, in modern-day Syria during the second millennium BCE; and some of the earliest Old Persion, among other languages.

Of these languages, Sumerian, Akkadian, and Hittite have the largest extant bodies of tablet fragments. Hittite tablet fragments number in the tens of thousands, and several hundred larger texts have been assembled from them. Sumerian and Akkadian tablets are far more numerous than Hittite tablets, as these languages were in common literary use for at least a thousand years longer than Hittite, and these languages have many more assembled texts than Hittite.

3. Sign Disambiguation Processes for Cuneiform

Cuneiform signs frequently have many values, and context determines the correct reading. For instance, the Sumerian sign AN can be the VC phonemic sequence /an/, the logogram for 'star', the logogram for 'god(dess)', or a noun classification prefix[1], indicating that the following sequence of signs is the name of a divine entity. Wordplay and puns are common as well, as in ^{URU}KA.DINGER.RA, phonologically in Sumerian 'ka dinger-ak', which is Sumerian meaning 'city of the gate of the god,' a phrase that, translated into Akkadian, is *bāb-ilī* 'gate of the god,' which sounds like *Babylon*, the city that the sequence names. A disambiguation process is therefor necessary to determine the correct reading for a particular sign, given its context.

- *Hittite*: Many signs are ambiguous. In Hittite, the sign *an*, for instance, has three distinct values. First, it represents the phonological sequence *-an-*. Second, it has two ideographic values. It can be the Sumerogram *DINGIR*, as a noun meaning 'god' or 'divine entity,' and it can be a deteminative sign, indicating that the sequence to follow the name of a god. For instance, the sequence *an-u-as* must be read as ^{d}*Ú-aš*[2], the ideographic writing of the name of the Hittite local storm god, *Tarhuntaš*. For the most part, Hittite ambiguities can be resolved with a small dictionary of rules, and errors will typically only occur with unseen personal (or divine) names.
- *Akkadian*: Akkadian uses a mix of Sumerian logograms, like AN for Akkadian *illum* 'god', and Sumerian syllabic signs, typically represented in transliteration with italic capital characters, as in the occasional phonemic spelling of the above form as *IL-LU-UM*.
- *Sumerian*: Sumerian presents the deepest challenge in sign disambiguation. The definitive text for keys

[1] Classification prefixes indicate human or divine names, certain animals (the sign MUSEN precedes names of bird species, for instance), and city names.

[2] In transliteration of Hittite texts, phonemic Hittite words are spelled with lower-case characters, Akkadian ideograms (and occasional Akkadian phonemic spellings) with italic capitals, and Sumerian logograms with plain capitals.

for the pronunciation of Sumerian ideograms is a table called Proto-Ea.[3] Sumerian spelling is quite variable, and signs are used by some scribes idiomatically, leading to much scholarly argument of the reading of Sumerian words.

The current working dictionary (kept in Prolog) for this project includes many rules of the form:

> lexicon([an,lum], [DINGER,LUM]).
> lexicon([an,zi,da,an,ta|REST],
> [m,zi,da,an,ta|REST]).

In these rules, the first argument is the strict phonological input, and the second argument is the correct reading for the signs, logograms in the first case, and the replacement of a personal name determinative marker in the second.

4. Morphology and Lemmatization

Morphological complexity in all three languages is at issue:

- Sumerian verbs are morphologically complex, with great numbers of agreement prefixes and suffixes. A morphological template, such as are often used with Native American languages, should prove useful for Sumerian verbs.
- Akkadian words show ablaut patterns typical of Semitic languages, and further uses significant suffix morphology for for person and number agreement with verbs. A number of prior studies have had success in lemmatizing modern Semitic languages, particularly Arabic and Hebrew. A particularly promising approach, and one that seems likely to be successful for lemmatizing Akkadian as well, is a clustering algorithm, presented in (De Roeck and Al-Fares, 2000)
- Hittite words carry significant inflectional morphology, both in the nominal and verbal systems. Hittite nouns are inflected in the singular and plural, with five cases for each number category. Verbs carry subject agreement and tense as a single suffix morpheme, and some verbs change stems between tenses.

For lemmatizing in a more automatic way, and for automatically identifying both roots and suffixes in Hittite and Akkadian, I have begun to use the following procedure, which, critically, sees through morphophonemic variation at morpheme boundaries to spot roots and suffixes through phonological processes like assimilation and deletion. As my corpora of Hittite and Akkadian fragments are still under construction, I used a corpus of Latin text - Vergil's *Aeneid* for initial tests of this procedure, the results of which are discussed below.
Tokenize the transcribed fragment at word boundaries. Assume that each word w_n is composed of a root r_n and a suffix s_n, i.e. $w_n = r_n s_n$. Note that r_n must be non-empty, and s_n must have a length of between one and five characters. One of the following processes applies to each word in the list.

1. If w_1 has already been decomposed, do nothing and remove it from the list.
2. If w_1 has a root r_1 and a suffix s_1 that are already in our list of decompositions but are not attested together, record the new combination.
3. If w_1 (composed of $r_1 s_1$) contains a suffix s_1 already in our list of decompositions but a root r_1 that is not, find another word w_2 such that $w_2 = r_1 s_2$. Record the decomposition of w_1 into $r_1 s_1$.
4. If w_1 contains a root r_1 already present in our list of decompositions but a suffix s_1 that is not, then find another word $w_2 = r_2 s_1$. Record the decomposition of w_1 into $r_1 s_1$.
5. If r_1 contains neither a root r_1 nor a suffix s_1 present in our list of decompositions, find a word $w_w = r_1 s_2$ and a word $w_3 = r_2 s_1$. Record the decomposition of w_1 into $r_1 s_1$.
6. If none of the above is possible, examine the word for sequences of the type *-ts-*, *-ps-*, and *-ks-*, replace any such sequences with *-ds-*, *-bs-*, and *-gs-*, and attempt the process again. Further, the replacement of orthographic *x* with *ks* was added to deal with the orthographic practices of Latin.

With this procedure, many Latin inflectional morphemes and roots are discovered correctly, as in the uncovered paradigm of 'king' with root *reg* plus distinct endings, including the first form, which presents both an assimilation process (anticipatory devoicing of the final root consonant) and an orthographic process:

- rex
- regis
- regi
- regem
- rege

Many forms are also decomposed incorrectly or partially incorrectly by this process, as shown by *tempestatumque*, decomposed into root *tempestat* and suffix *umque*. Regardless, these results show promise for eventual use in helping build dictionaries of roots and suffixes on less-regular cuneiform-language transliteration data, and I expect to perform these tests soon.

5. Cuneiform Corpora as a Classification Problem

Very little computational attention has been focused on cuneiform languages in the past. An early study of text similarity between standard Babylonian Akkadian and the variant of Akkadian used in the mostly Hurrian-speaking town of Nuzi, in which Smith used vector similarity measures to establish Hurrian substrate interference in the Nuzi Akkadian texts, rather than to connect tablet fragments into more complete texts (Smith, 2007).

[3] Available in transliteration at `http://cdli.ucla.edu/tools/lexical/proto_ea_sign.html`

Viewed as a classification problem, cuneiform corpus creation becomes a process of establishing whether a particular unknown tablet fragment belongs with any currently-known texts. This approach was taken with mixed success with a corpus of Hittite tablet fragments by (Tyndall, 2012). Tyndall used Naive Bayes and MaxEnt classifiers with simple, unanalyzed wordcounts as features, and pre-existing Hittite texts as the labels to which the unknown fragments were assigned.

I plan to build on this work, to extend the classification problem by preprocessing by the lemmatization schemes described above before using similar machine learning techniques, and hopefully improve upon the somewhat weak accuracy of these classifiers. Further, the addition of a ranked suggestion set of known texts to which a particular fragment might belong is in process.

6. Cuneiform Corpora as a Clustering Problem

Clustering is another natural method to use for connecting tablet fragments with other, similar fragments, one that's yet to be implemented anywhere in the literature. This technique shows much promise in dealing with the vast quantity of unlabeled tablet fragments stored in museums around the world - many cuneiform tablet fragments are poorly provenenced, and many belong to unknown texts. In particular, a large quantity of new Hittite material has appeared in recent excavations of the ancient city of Šapinuwa. Some 5000 new fragments have already come to light, and more are appearing (Süel, 2002). Clustering processes, while unlikely to be successful on their own, will likely help point scholars to collections of fragments with similar vocabulary and topics, and provide a useful tool for the establishment of new complete texts.

7. Conclusions and Further Work

In this paper, I have presented a number of challenges facing the study of cuneiform languages and have proposed solutions grounded in current literature to a few of them. One of the primary goals of this paper is, rather than to solve these issues herein, to stimulate discussion and connections between computational linguistics and natural language processing experts on one hand and cuneiform scholars on the other. Adapting more complex morphological, syntactic, and discourse-level machine-learning methods from the contemporary NLP and machine learning literature will be a great benefit to the discipline of cuneiform studies, and, eventually, to the understanding of this important era of early literature and cultural history.

8. References

Sean E. Anderson and Marc Levoy. 2002. Unwrapping and visualizing cuneiform tablets. *Computer Graphics and Applications, IEEE*, 22(6):82–88.

Giovan Battista D'Alessio. 1997. Pindar's "prosodia" and the classification of pindaric papyrus fragments. *Zeitschrift für Papyrologie und Epigraphik*, pages 23–60.

Anne N De Roeck and Waleed Al-Fares. 2000. A morphologically sensitive clustering algorithm for identifying arabic roots. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 199–206. Association for Computational Linguistics.

Horst Klengel. 2002. Problems in hittite history, solved and unsolved. In Simrit Dhesi K. Aslihan Yener, Harry A. Hoffner Jr., editor, *Recent developments in Hittite archaeology and history: papers in memory of Hans G. Güterbock*, pages 101–109. Eisenbrauns.

Hubert Mara, Susanne Krömker, Stefan Jakob, and Bernd Breuckmann. 2010. Gigamesh and gilgamesh:–3d multiscale integral invariant cuneiform character extraction. In *Proceedings of the 11th International conference on Virtual Reality, Archaeology and Cultural Heritage*, pages 131–138. Eurographics Association.

Terhi Nurmikko, Jacob Dahl, Nicholas Gibbins, and Graeme Earl. 2012. Citizen science for cuneiform studies.

S.P. Smith. 2007. *Hurrian Orthographic Interference in Nuzi Akkadian: A Computational Comparative Graphemic Analysis*. Ph.D. thesis, Harvard University Cambridge, Massachusetts.

A. Süel. 2002. Ortaköy-sapinuwa. In Simrit Dhesi K. Aslihan Yener, Harry A. Hoffner Jr., editor, *Recent developments in Hittite archaeology and history: papers in memory of Hans G. Güterbock*, pages 157–165. Eisenbrauns.

Stephen Tyndall. 2012. Toward automatically assembling hittite-language cuneiform tablet fragments into larger texts. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 243–247. Association for Computational Linguistics.

Geert Willems, Frank Verbiest, Wim Moreau, Hendrik Hameeuw, Karel Van Lerberghe, and Luc Van Gool. 2005. Easy and cost-effective cuneiform digitizing. In *The 6th International Symposium on Virtual Reality, Archaeology and Cultural Heritage (VAST 2005)*, pages 73–80.

Greinir skáldskapar: A diachronic corpus of Icelandic poetic texts

Þórhallur Eyþórsson, Bjarki Karlsson, Sigríður Sæunn Sigurðardóttir
University of Iceland
E-mail: tolli@hi.is, bjarki@fraedi.is, sss17@hi.is

Abstract

Greinir skáldskapar is a part-of-speech tagged, lemmatized and syntactically annotated corpus of historical Icelandic poetry (http://bragi.info/greinir/), which it is also annotated for phonological and metrical factors. The purpose of the corpus is to enable an integrated search for syntax (dependency grammar model), phonology and metrics. The database employs a preprogrammed query system, where the user chooses what to search for from an option window. Currently, the options include query possibilities concerning metrics, syllable structure, compound words, grammatical categories, clause structure, syntactic dominance and word search. Different factors can be combined to search for elements that fulfill more than one condition. It is also possible to conduct a search in a specific poem only, or particular poetic genre. A comparison of Greinir skáldskapar with a diachronic corpus of Icelandic prose (IcePaHC) reveals that despite differences they complement each other in a number of ways.

Keywords: poetic corpus, integrated search, Old Icelandic

1. Introduction

In this paper we present a part-of-speech tagged, lemmatized and syntactically annotated corpus of historical Icelandic poetry, Greinir skáldskapar (http://bragi.info/greinir/; Karlsson et al. 2012). This particular database was created in cooperation with other corpora projects, notably the Icelandic project Medieval Manuscripts of the Eddic Poetry: An Electronic Edition and a Lemmatized Concordance (Bernharðsson & Gunnlaugsson, 2013), and the Norwegian projects PROIEL, ISWOC and MENOTEC. In addition to being part-of-speech tagged, lemmatized and syntactically annotated, Greinir skáldskapar is also annotated for phonological and metrical factors. Unlike the Icelandic Parsed Historical Corpus (IcePaHC; Wallenberg et al., 2011), which was modeled on the Penn Treebank and annotated according to phrase structure rules, the text in Greinir skáldskapar makes use of dependency grammar. The initial motivation for choosing dependency grammar over phrase structure grammar in this particular corpus was the fact that it contains Old Norse-Icelandic poetic texts which exhibit a rather free word order, at least compared to that of the prose texts in IcePaHC. Moreover, dependency grammar has proven to be a suitable syntactic framework for highly inflected languages like Old Norse-Icelandic, and it is commonly employed in language technology.

The organization of the paper is as follows. First, we present a brief overview of the texts contained in Greinir skáldskapar and their main characteristics (section 2). Next, we discuss the syntactic analysis in the corpus which is based on the dependency guidelines of the Norwegian MENOTEC team (section 3). Thereupon, we sketch the phonological and metrical analysis employed in the corpus (section 4), before describing some of the major novel features of the database and the query system (section 5). Finally, we offer a brief comparison of Greinir skáldskapar with the IcePaHC corpus, which contains Icelandic prose texts from the 12th to the 21st century (section 6). Section 7 concludes the paper.

2. The Texts

Greinir skáldskapar is a database containing Old Icelandic poetry. The ultimate goal of our project is that it will include both Eddic and skaldic poetry, along with a selection of later epic poetry (*rímur*). At present, however, only the Poetic Edda is accessible fully annotated.

The text of the Poetic Edda was obtained from the Icelandic project Medieval Manuscripts of the Eddic Poetry: An Electronic Edition and a Lemmatized Concordance. The text is an accurate transcript of the Codex Regius manuscript (GKS 2365 4to) and comes in three different layers: as a facsimile text, a diplomatic text and a text using normalized orthography. To simplify matters, only the normalized text was used for syntactic annotation.

A complete list of Eddic poems available in Greinir skáldskapar can be seen in Table 1. The number of stanzas and words in each poem is also given. The total number of poems is 29. Since the purpose of the corpus is to enable an integrated search for syntax, phonology and metrics, a few words about the main Eddic meters is appropriate. *Fornyrðislag*, marked F in Table 1, is usually considered to be the oldest meter of Icelandic poetry. It is an alliterative verse with either one or two *stuðlar* in primary lines and a *höfuðstafur* in subsequent lines. There are two lifts per line. *Ljóðaháttur*, marked L in Table 1, is also an alliterative verse. A stanza in this type of meter can be divided into two parts, each containing three lines. The first two lines in each part alliterate, but lines 3 and 6 are usually longer and have their own internal alliteration (two *stuðlar*).

In addition to the Eddic poems from Codex Regius, a selection of skaldic poetry and lausavísur is available in the database, but only metrically annotated,. The text was modeled on various scholarly editions of skaldic poetry, including *Skjaldedigtning* edited by Jónsson (1912-15). At present, two complete *rímur* are in Greinir skáldskapar, *Ólafs rímur Haraldssonar* and *Ormars rímur.* Both these texts are normalized versions, edited by Þorgeirsson (2013).

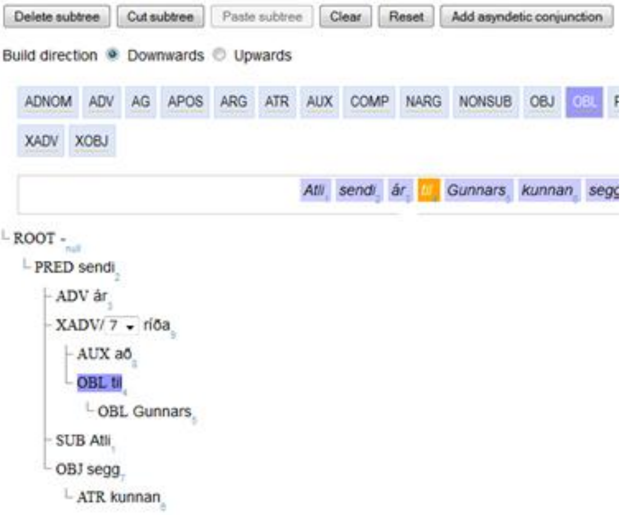
Poem	**Meter**	**Stanzas**	**Words**
Völuspá	F	62	1471
Hávamál	L	161	4028
Vafþrúðnismál	L	56	1179
Grímnismál	L	56	1214
Skírnismál	L	42	955
Hárbarðsljóð	F	60	1080
Hymiskviða	F	36	827
Lokasenna	L	66	1492
Þrymskviða	F	32	746
Völundarkviða	F	41	947
Alvísmál	L	35	756
Helgakviða Hundingsbana I	F	58	1219
Helgakviða Hjörvarðssonar	F	45	1035
Helgakviða Hundingsbana II	F	51	1222
Grípisspá	F	52	1246
Reginsmál	L	27	457
Fáfnismál[1]	L	45	991
Sigurdrífumál	L	28	720
Brot af Sigurðarkviðu	F	20	417
Guðrúnarkviða I	F	25	561
Sigurðarkviða in skamma	F	67	1586
Helreið Brynhildar	F	13	300
Guðrúnarkviða II	F	45	982
Guðrúnarkviða III	F	10	237
Oddrúnargrátur	F	34	730
Atlakviða	F	46	1114
Atlamál in grænlensku	F[2]	117	2696
Guðrúnarhvöt	F	23	477
Hamdismál	F	26	704
SUM =		**1379**	**31389**

Table 1: List of annotated Eddic poems in Greinir skáldskapar

3. Syntactic Annotation

The syntactic analysis of the Poetic Edda was carried out manually in the Norwegian PROIEL corpus, according to the dependency guidelines created by the MENOTEC team (Haugen and Øverland 2013). The analyzed text was then transported to Greinir skáldskapar. An example from *Atlakviða*, of the sentence in (1), with annotation, is shown in Figure 1, and a dependency tree of the same sentence is shown in Figure 2.

(1) Atli sendi / ár til Gunnars / kunnan segg at ríða.
Atli sent of old to Gunnarr known man to ride
'Atli sent long ago a known man riding to Gunnarr.'
(*Atlakviða* 1)



Build direction ● Downwards ○ Upwards

Figure 1: The annotation mechanism of the sentence in (1)

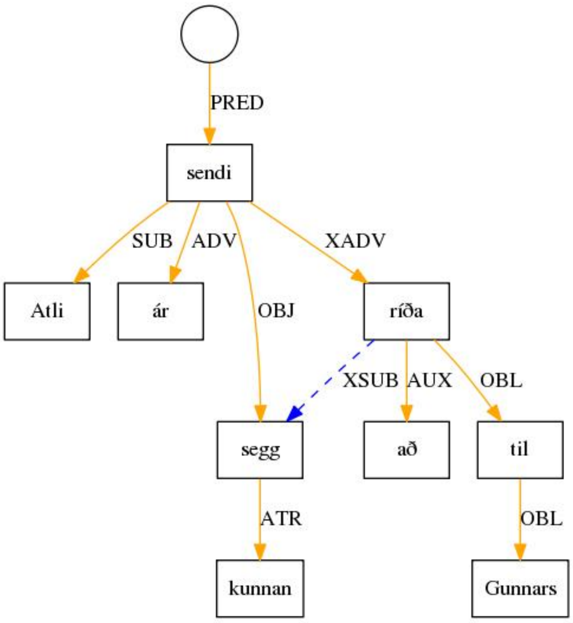


Figure 2: A dependency tree diagram of (1)

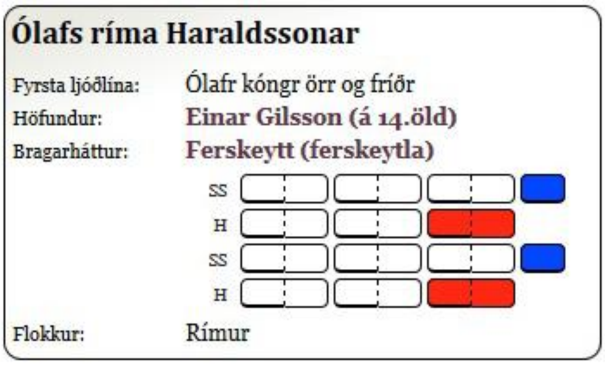
As expected, the syntactic annotation influences how material can be retrieved from the corpus. For example, in the case of the syntactic annotation of auxiliary verbs in the Eddic poetry, the Greinir skáldskapar team decided to follow the rules created by the Norwegian MENOTEC project. When they began the annotation there were no specific rules for how the auxiliaries should be annotated so initially they were all marked as PRED. Eventually,

[1] The text of *Fáfnismál* is currently under revision.
[2] The metrical analysis of *Atlamál* is under revision.

however, it was decided that it would be more advantageous to annotate the auxiliaries as AUX dependent on PRED. The result is that all auxiliary verbs can now be easily retrieved from the database – provided that the researcher is familiar with their marking. The search would only have to include a verb marked as AUX.

4. Phonological and Metrical Annotation

In phonological and metrical annotation of the texts, carried out in Greinir skáldskapar, every stanza was annotated for syllable structure, alliteration, lifts, feet and other variables.[3] In addition to all stanzas being annotated independently for phonological and metrical variables, a general description of the main meter of each poem is given in a "visual" way. Quadrangular fields are used to represent number of feet in line. Every field is divided into one or more parts according to number of syllables in each foot. Colors are used to indicate rhyme, and a bold line at the bottom of each box to indicate stressed syllables. Finally, the letters S and H are used to denote *stuðlar* (S) and *höfuðstafur* (H). This representation is shown in Figure 3 for *Ólafs ríma Haraldssonar*, which is an example of a *ríma*.

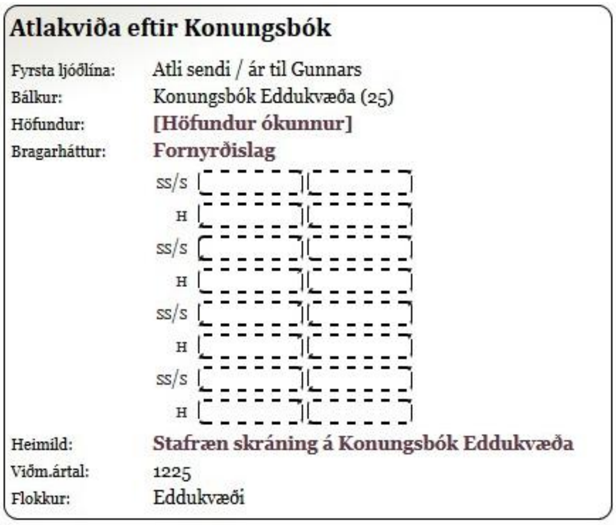


Ólafs ríma Haraldssonar

Fyrsta ljóðlína: Ólafr kóngr örr og fríðr
Höfundur: Einar Gilsson (á 14.öld)
Bragarháttur: Ferskeytt (ferskeytla)
Flokkur: Rímur

Figure 3: A representation of the meter in *Ólafs ríma Haraldssonar*

In the case of *Ólafs ríma Haraldssonar*, the following can be observed: Each stanza consists of four lines with four feet in primary lines and three in subsequent lines. The last foot in the primary lines is one syllable long and these lines rhyme. The last foot in lines 2 and 4 contains two syllables which rhyme. Two *stuðlar* are found in lines 1 and 3, and a single *höfuðstafur* is found in lines 2 and 4. Comparing this with the meter of *Atlakviða* (Figure 4), it becomes clear that *Ólafs ríma Haraldssonar* is metrically more complicated than the traditional *Fornyrðislag*. In the latter meter, which is an accentual verse, there are two lifts, with a varying number of syllables (indicated by broken lines), and there are either one or two *stuðlar* as against a *höfuðstafur*.



Atlakviða eftir Konungsbók

Fyrsta ljóðlína: Atli sendi / ár til Gunnars
Bálkur: Konungsbók Eddukvæða (25)
Höfundur: [Höfundur ókunnur]
Bragarháttur: Fornyrðislag
Heimild: Stafræn skráning á Konungsbók Eddukvæða
Viðm.ártal: 1225
Flokkur: Eddukvæði

Figure 4: A description of *Fornyrðislag* (*Atlakviða*)

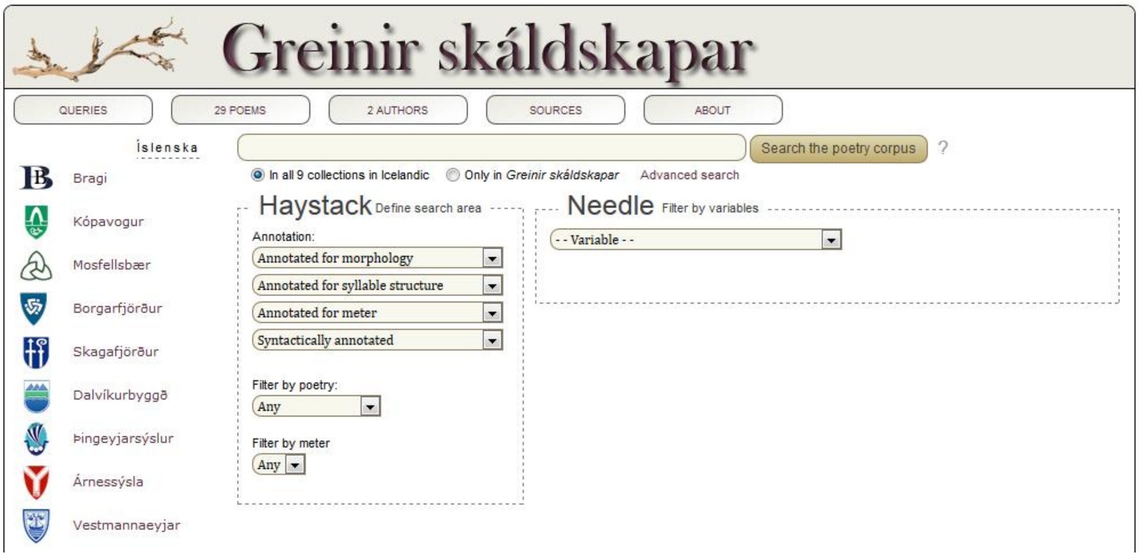
5. The database and the query system

The database itself is accessible online at http://bragi.info/greinir/. In order to reduce the time the researcher has to spend on learning how to make queries, the team decided that the database should be easily accessible and user friendly. This resulted in a preprogrammed query system, where the user chooses what to search for from an option window. Currently, the options include query possibilities concerning metrics, syllable structure, compound words, grammatical categories, clause structure, syntactic dominance and word search. Importantly, it is possible to combine different factors in order to search for elements that fulfill more than one condition. Moreover, it is possible to conduct a search only in a specific poem or a specific poetic genre.

The query system of Greinir skáldskapar is shown in Figure 5. The system is divided into two parts. First, there is the Haystack which defines the search domain. It allows the querent to decide whether to search for variables in the whole corpus, in a specific type of meter, a specific branch of poetry or even in a specific poem. It also allows for the search area to be defined in terms of the status of the annotation. Secondly, there is the Needle, where the querent can decide which variable to search for. Notice that variables can be added to the search to narrow it down.

Although the corpus can be used to search for various types of linguistic and/or metrical phenomena, it is particularly aimed at those interested in the interplay between metrics and syntax. Combining different metrical and syntactic factors makes it, for example, possible to search for all clause-initial subjects that alliterate, or all clause-initial subjects that do *not* alliterate. Moreover, questions about a possible correlation between the occurrences of prepositional stranding (postpositions) and type of meter can be easily tackled in Greinir skáldskapar.

[3] The English version of this part of Greinir skáldskapar is still under construction. Hence the information in Fig. 3 and 4 is in Icelandic.



● In all 9 collections in Icelandic ○ Only in *Greinir skáldskapar*

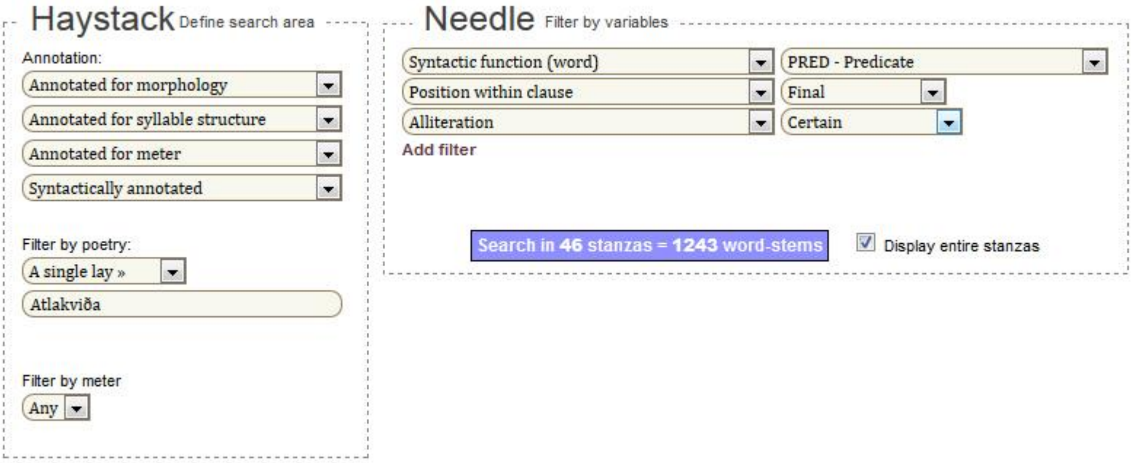
Annotation:
Annotated for morphology
Annotated for syllable structure
Annotated for meter
Syntactically annotated

Filter by poetry: Any

Filter by meter: Any

- - Variable - -

Figure 5: Greinir skáldskapar – the query system



Annotation:
Annotated for morphology
Annotated for syllable structure
Annotated for meter
Syntactically annotated

Filter by poetry: A single lay »
Atlakviða

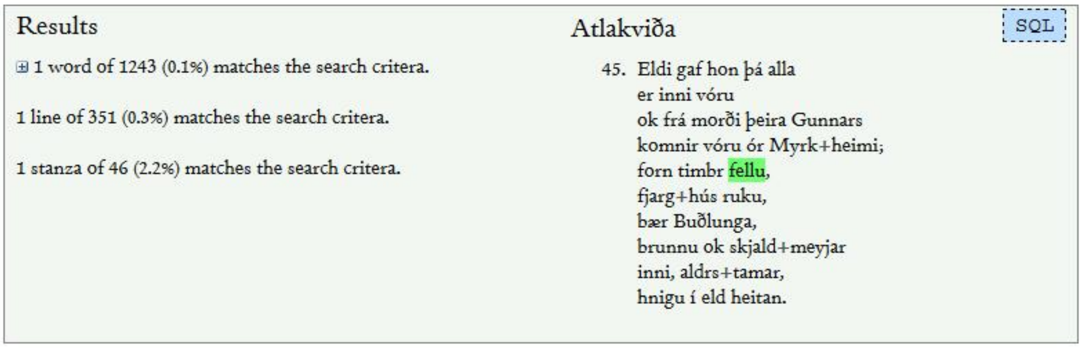
Filter by meter: Any

Syntactic function (word): PRED - Predicate
Position within clause: Final
Alliteration: Certain

Search in 46 stanzas = 1243 word-stems

☑ Display entire stanzas

Figure 6: Query example – verb-final position combined with alliteration of verb



Results

⊞ 1 word of 1243 (0.1%) matches the search critera.

1 line of 351 (0.3%) matches the search critera.

1 stanza of 46 (2.2%) matches the search critera.

Atlakviða

45. Eldi gaf hon þá alla
er inni vóru
ok frá morði þeira Gunnars
komnir vóru ór Myrk+heimi;
forn timbr fellu,
fjarg+hús ruku,
bær Buðlunga,
brunnu ok skjald+meyjar
inni, aldrs+tamar,
hnigu í eld heitan.

Figure 7: Results for verb-final position with alliteration (*Atlakviða*)

A very simple example of a possible query is shown in Figure 6. The query involves a clause-final verb that alliterates. As can be seen in the Haystack, the search is only conducted in *Atlakviða*. The variables searched for include a word with the syntactic function PRED, whose position is clause-final. In addition the verb must take part in alliteration. The results for the search in Figure 6 are given in Figure 7. As seen, only one word of a total of 1243 in *Atlakviða* meets these conditions, i.e. less than 1% of the poem.

6. Comparison with IcePaHC

Since only one other Icelandic treebank exists, i.e. the Icelandic Parsed Historical Corpus (IcePaHC; Wallenberg et al., 2011), a brief comparison of the two corpora in order. Crucially, these research tools have very different aims. While IcePaHC contains prose texts, Greinir skáldskapar focuses on poetic texts. A further comparison reveals that even though both corpora share the ultimate goal of facilitating diachronic research, they are in fact quite different with respect to their practical aspects. Ignoring for now the fact that these corpora are annotated differently (according to dependency vs. phrase structure), it is important to note that while one of them (IcePaHC) has a special query language that needs to be learned, the other one (Greinir skáldskapar) has a very user-friendly interface which requires little or no knowledge of how particular query-strings are designed. However, this user-friendly interface comes at a cost, as the researcher using Greinir skáldskapar has to rely on preprogrammed search possibilities (although scholars with access to the system can in fact add possibilities).

The relevant characteristics of the two Icelandic corpora can be summarized as follows:

(1) IcePaHC – annotated according to phrase structure rules.

- Pros: open access; 1,002,390 words (current version 0.9), downloadable (it is possible to download the raw files); results are easily transferrable to a text editor (notepad, word etc.), endless query-possibilities since the researcher designs his/her own queries.
- Cons: rather complicated query system (special query language); one needs some practice to read through the results due to the way they are represented.

(2) Greinir skáldskapar – annotated according to dependency grammar.

- Pros: open access; user-friendly interface; easily accessible online (no download needed); it is possible to choose which texts are searched; easily readable results; possibility to perform integrated search from numerous modules simultaneously (syntax, morphology, phonology, metrics).
- Cons: small (so far only the Eddic Poetry from Codex Regius and a few other texts), limited possibilities for making queries.

As has been mentioned, Greinir skáldskapar and IcePaHC make use of different syntactic frameworks (dependency vs. phrase structure). This raises the question whether these corpora can really be used together. In this connection, it is worth pointing out that there is already work in progress aiming at mapping phrase structure to dependency structure and vice versa, so perhaps the fact that Greinir skáldskapar makes use of dependency grammar while other corpora use phrase structure is not necessarily very relevant.

In conclusion, both corpora, IcePaHC and Greinir skáldskapar, have strengths and weaknesses, but perhaps the most important thing is that they supplement each other in numerous ways so searching for diachronic linguistic information in Icelandic has never been as easy and straightforward. In both corpora, however, when searching for syntactic structures, the researcher has to know how s/he can retrieve the desired information. Usually this means s/he has to have prior knowledge of how the text contained in the relevant corpus is annotated and what kind of queries have to be made in order to perform a successful search. This raises the obvious question of how much corpus-computer-programming knowledge we can assume the researcher needs to have in order to find what s/he is looking for. The answer to this question in regard to both of the Icelandic diachronic corpora is the same – to be able to make successful use of these corpora a considerable amount of such knowledge is required.

7. Conclusion

In this paper we have presented Greinir skáldskapar, a part-of-speech tagged, lemmatized and syntactically annotated corpus of historical Icelandic poetry (http://bragi.info/greinir/), which is also annotated for phonological and metrical factors. We have shown that the purpose of the corpus is to enable an integrated search for syntax (dependency grammar model), phonology and metrics. The database employs a preprogrammed query system, where the user chooses what to search for from an option window, and the current options include query possibilities concerning metrics, syllable structure, compound words, grammatical categories, clause structure, syntactic dominance and word search. As we exemplified, different factors can be combined to search for elements that fulfill more than one condition. Moreover, a search in a specific poem or a particular poetic genre is a possibility. A comparison of Greinir skáldskapar with a diachronic corpus of Icelandic prose (IcePaHC) reveals that even though both share the ultimate goal of facilitating research, they are in fact quite different with respect to their practical aspects. Finally, the strengths and the weaknesses of these corpora were contrasted and compared, with the aim of showing how these unique research tools complement each other.

8. References

Bernharðsson, H. and Gunnlaugsson, G.M. (2013). Medieval Manuscripts of the Eddic Poetry: An Electronic Edition and a Lemmatized Concordance. The Árni Magnússon Institute of Icelandic Studies, Reykjavík.

Haugen, O.E. and Øverland, F.Th. (2013). Guidelines for Morphological and Syntactic Annotation of Old Norwegian Texts. The Project Menotec, The University of Bergen. https://www.academia.edu/4875502/ Menotec_Guidelines_for_Syntactic_ Annotation

Jónsson, F. (1912-15). *Den norsk-islandske skjalde-digtning*. København/ Kristiania: Gyldendal.

Karlsson, B.; Árnason, K.; and Eythórsson, Th. (2012). Greinir skáldskapar – a tagged corpus of Old Norse-Icelandic poetry. Developed within the project Interfaces of Metrics, Phonology and Syntax, University of Iceland 2009-11 (PI: Eythórsson, Th.). http://bragi.info/greinir/

Þorgeirsson, H. (2013). *Hljóðkerfi og bragkerfi – Stoðhljóð, tónkvæði og önnur úrlausnarefni í íslenskri bragsögu ásamt útgáfu á Rímum af Ormari Fraðmarssyni*. (English summary: How poetic is phonology? Studies in Old and Middle Icelandic Poetry.) Doctoral dissertation, University of Iceland. Reykjavík: Hugvísindastofnun Háskóla Íslands.

Wallenberg, J.; Ingason, A.K.; Sigurðsson, E.F.; and Rögnvaldsson, E. (2011). Icelandic Parsed Historical Corpus (IcePaHC). Version 0.9. http://www.linguist.is/ icelandic_treebank

Section 2:
Historical Newspaper Archives

On the Role of Historical Newspapers in Disseminating Foreign Words in German

Oliver Pfefferkorn, Peter Fankhauser

IDS-Manneim
Germany
pfefferkorn@ids-mannheim.de, fankhauser@ids-mannheim.de

Abstract

Newspapers became extremely popular in Germany during the 18^{th} and 19^{th} century, and thus increasingly influential for modern German. However, due to the lack of digitized historical newspaper corpora for German, this influence could not be analyzed systematically. In this paper, we introduce the Mannheim Corpus of Digital Newspapers and Magazines, which in its current release comprises 21 newspapers and magazines from the 18^{th} and 19^{th} century. With over 4.1 Mio tokens in about 650 volumes it currently constitutes the largest historical corpus dedicated to newspapers in German. We briefly discuss the prospect of the corpus for analyzing the evolution of news as a genre in its own right and the influence of contextual parameters such as region and register on the language of news. We then focus on one historically influential aspect of newspapers – their role in disseminating foreign words in German. Our preliminary quantitative results indeed indicate that newspapers use foreign words significantly more frequently than other genres, in particular belle lettres.

Keywords: Historical Corpora, Newspapers, Language Variation

1. Introduction

Newspapers became extremely popular in Germany during the 18^{th} and 19^{th} century. In the 18^{th} century several types of newspapers, among them political newspapers, advertising supplements or intelligencers, and weekly magazines emerged, coinciding with the rise of a civil public. In contrast to other types of public and private writing, newspapers are characterized by actuality, periodicity, topical diversity, and public access (Wilke, 1999, p. 388). They reached a wider audience than any other text type and spread across all social classes.

Wilke (1999, p. 397) estimates that German newspapers reached about 3 Mio readers in the late 18^{th} century. Thus, the language of newspapers has been very influential for the development of German in the 18^{th} and 19^{th} century. Up to now this influence has not been analyzed systematically; rather analysis of the more recent history of German has focussed on few selected varieties, in particular (high) literature.

A historical reason for this marginalization of newspapers as an influential factor for (early) modern German is the rather widespread contempt for the medium by the educated middle class (Theobald, 2012). A more practical reason though is the unavailability of newspaper corpora. Newspapers typically have not been reprinted, and the original facsimiles are difficult to access. Therefore, research on language history is typically focussed on the analysis of individual newspapers and rather specific linguistic questions.

The remainder of this paper is organized as follows: In Section 2. we describe the Mannheim Corpus of Historical Newspapers and Magazines and in Section 3. we outline some of its potential uses for analyzing the influence of newspapers on modern German. In Section 4. we analyze the role of newspapers in popularizing foreign words, and in Section 5., we conclude and outline future work.

2. The Mannheim Corpus of Historical Newspapers and Magazines

The Mannheim Corpus of Historical Newspapers and Magazines (MKHZ) consists of 21 German newspapers and magazines from the 18^{th} and 19^{th} century. The currently publically available version[1] comprises about 650 individual volumes with over 4.1 Mio tokens on 4678 pages overall[2].

In addition to the original page scans, it is available in TUSTEP format (TUSTEP, 2013), acquired in a double keying procedure, and in TEI P5, which was generated semi-automatically from the TUSTEP version (Fankhauser et al., 2013b). On this basis, a human readable version in HTML aligned with the page scans, CMDI metadata (Broeder et al., 2011) for its long term archival in the IDS Repository (Fankhauser et al., 2013a), and an *IDS–XCES* version for import into the Corpus Search and Analysis System *COSMAS II* (Bodmer, 2005) were generated.

For each volume, basic metadata including title, publication date and place, and a broad classification into newspapers vs. magazines are available. The logical structure is very simple, individual volumes consist of paragraphs and tables, a division into individual articles is currently underway.

3. Digitized Newspapers for Linguistic Research

Digitized newspaper corpora enable a quantitative perspective on the role of news for the historical development of German. Here is a nonexhaustive list of potentially interesting questions to investigate:

[1] *MKHZ*: http://hdl.handle.net/10932/00-01B8-AE41-41A4-DC01-5

[2] For comparison, the historical newspaper corpus compiled for the GerManC corpus (Bennett et al., 2010) comprises about 300.000 tokens.

News as a Genre: How did the news genre differentiate itself from other genres, such as science or fiction?

Types of News: How did various types of news and newspapers evolve and differentiate themselves (e.g. news, announcement, background feature, reportage (Püschel, 1999))? And how do these types manifest themselves linguistically?

Homogenization: Whether and how were regional and dialectal characteristics reduced over time?

Advertisements: How did various types of advertisements evolve (e.g. product ads, job ads, real estate ads, private ads, personal ads, purchase ads, etc.)?

Spoken Language: Newspapers contain text types pertaining to some extent to spoken language, such as speeches, protocols, calls, or letters to the editor, which may help in analyzing historical spoken language.

4. Foreign Words in Newspapers

In this section we analyze one aspect of newspapers as a genre in more detail. Newspapers and magazines have never been shy of using foreign words. This can already be observed for the early German weekly magazines in the 17^{th} century (Gloning, 1996, pp. 164 – 179). Authors of dictionaries of foreign words in the 18^{th} and 19^{th} almost formulaically refer to "assisting their readers in comprehending newspapers" in their subtitles. This gives rise to the hypothesis that newspapers took over and popularized many foreign words from specialized subject domains. As a consequence many of these foreign words have been taken over into standard German.

Until now though, the historical role of newspapers in disseminating foreign words has not been analyzed systematically due to the lack of appropriate historical newspaper corpora. It is unclear which foreign words were used most frequently in newspapers, which foreign words in contemporary dictionaries of the 18^{th} and 19^{th} century were used at all, and what their percentages in comparison to other genres or registers were.

For a preliminary investigation of these questions, we have derived a list of foreign words from a German dictionary of foreign words, originally started in 1913 and continually revised since then[3]. The list comprises about 3700 *main* lemmata for the letters "A" to "Q". It is by no means exhaustive; it does not cover letters "R" to "Z", it only takes into account the main lemmata, and it does not cover foreign words that have become obsolete again. Nevertheless, it can be regarded as a representative sample for foreign words that have been considered part of standard German by lexicographers [4]

For comparing the historical role of newspapers in picking up foreign words as opposed to other registers, we use in addition the German Text Archive *DTA*[5] (Geyken et al., 2011), compiled as a representative cross section of texts from 1600 – 1920. *DTA* currently comprises about 90 Mio tokens, and is drawn from the following three broad genres: factual writing (*Gebrauchsliteratur*), belle lettres (fiction, drama, poetry etc.), and learned (scientific writing), but has only a relatively small coverage of historical newspapers.

From both corpora a list of all types together with their frequencies is extracted. To match inflected forms with the lemmata from the list of foreign words, they are lemmatized with a word based lemmatizer (Belica, 1994) for German, and normalized using a small, non-exhaustive set of heuristic rules. The resulting matches are not 100% accurate; ambiguous words such as *modern* (as a (foreign) adjective: *modern*, as a (native) verb: *to molder/rot*) hurt precision, incomplete normalization and lemmatization, and not taking into account derived lemmata hurt recall. However, these inaccuracies have largely the same effect on the various (sub)corpora; fairly low recall is the dominant factor, and thus the reported percentages of foreign words in German constitute lower bound estimates.

Table 1 lists the overall number of tokens (*#t*), the number of tokens which match foreign words (*#f*), and the according percentage of foreign words (*%f*) for *MKHZ* and *DTA* over time[6] The relatively low percentage of foreign words in the 17^{th} century (*DTA* only) is mainly due to two reasons: *DFWB* does not include foreign words that have become obsolete again, and normalization and lemmatization are less accurate for this period. However, in the subsequent periods, which are covered by both corpora, the percentage of foreign words in *DTA* is consistently smaller than in *MKHZ*[7]. This indicates that indeed newspapers have historically contributed more strongly to disseminating foreign words into general language than other genres.

Table 2 analyzes the percentage of foreign words by genre. *MKHZ* is broadly divided into newspapers vs. (weekly) magazines, which often contain text pertaining more to belle lettres, such as serial novels. Indeed newspapers proper have a significantly higher percentage of foreign words than magazines. This is mirrored in *DTA* where foreign words in factual writing are also significantly more frequent than in belle lettres. As is to be expected scientific writing (*learned*) has the highest frequency of foreign words in *DTA*. Note that newspapers in *MKHZ* also have a higher percentage than scientific writing in *DTA*, but this may again well be due to the longer time span covered by *DTA*.

For a more detailed perspective, we analyze the dissemination of foreign words during the 18^{th} and 19^{th} century along three examples. In addition to *MKHZ* and *DTA*, we also use the historical corpus *DGB01*[8], which comprises

[3] Deutsches Fremdwörterbuch (DFWB). http://www1.ids-mannheim.de/lexik/fremdwort/

[4] Main lemmata in *DFWB* comprise loan words as well as some words derived by combining loan stems, prefixes, and suffixes with other words (derived lemmata).

[5] Deutsches Text Archiv. http://www.deutschestextarchiv.de. Version: Nov. 6, 2013, downloaded Feb. 11, 2014.

[6] The overall number of tokens is smaller for both corpora, because only alphabetic words occurring within paragraphs – no titles, tables, etc. – have been taken into account.

[7] All reported differences are significant according to a χ^2 test with a p-value well below 0.1%.

[8] Deutsche Bibliothek/Deutsche Literatur von Lessing bis Kafka (Digital Library/German Literature from Lessing to Kafka)

		MKHZ	DTA
1600 – 1700	#t	–	8176835
	#f	–	86664
	%f	–	1.06
1700 – 1800	#t	202654	14743225
	#f	5071	211507
	%f	2.50	1.43
1800 – 1850	#t	1548983	18565922
	#f	31168	348225
	%f	2.01	1.88
1850 – 1920	#t	2161855	28872818
	#f	51965	652814
	%f	2.61	2.26
Sum	#t	3913492	70358800
	#f	88222	1299210
	%f	2.25	1.85

Table 1: Foreign Words along Time

		MKHZ	DTA
newspapers / factual writing	#t	2293010	7724113
	#f	59924	140250
	%f	2.61	1.82
magazines / belle lettres	#t	1620482	16928838
	#f	28298	216876
	%f	1.75	1.28
learned	#t	–	45705849
	#f	–	942084
	%f	–	2.06

Table 2: Foreign Words by Genre

about 30 Mio tokens.

The noun *Agentur* (*agency*) was defined in dictionaries of the late 18th century as a derivation from *agent* meaning *office, capacity, assignment by an agent, mediator, representative*[9]. The revised *DFWB* (Strauß et al., 1995, p. 192) reports its first record in 1847 with the meaning *branch office, office of an agent*[10]. Since the beginning of the 19th century *Agentur* occurs regularly, but only since about 1870 it also occurs frequently. *DGB01* has only 3 occurrences, whereas *MKHZ* lists 149 occurrences between 1846 and 1877. In addition, *MKHZ* lists 30 occurrences of derived composites, such as *Generalagentur, Generalzeitungsagentur, Gesellschaftsagentur, Hauptagentur, Nachrichtenagentur, Patentagentur, Spezialagentur*. *DTA*, has only 32 occurrences of *Agentur* between 1855 and 1900. This strongly suggests that mainly newspapers have contributed to establishing *Agentur* in standard German.

The first edition of the *DFWB* reports 1779 as the first record of the adjective *provisorisch* (*provisional, temporary*) in German (Schulz and Basler, 1942, p. 716). Until the mid 19th century, it occurs almost exclusively in scientific contexts, in particular legal and philosophical writing, and only rarely in literary publications, as also evidenced by *DTA*. Only since the mid 19th century it occurs more frequently outside of scientific writing, and indeed primarily in newspapers and magazines (367 occurrences in MKHZ between 1847 and 1905 with a peak between 1848 and 1850). It is typically used in political contexts, such as in *der provisorische Zustand (the provisional state of affairs)*, *die provisorische Regierung/Zentralgewalt (the provisional government/central power)*, or *provisorische Anordnungen/Gesetze (provisional orders/laws)*. In comparison, the literature corpus *DGB01* only records 53 occurrences in the period of 1793 and 1923. In this case, newspapers adopt the scientific terminology, adapt it to more domains, and thereby disseminate it.

As a last example, the adjective *reaktionär* (*reactionary*) was borrowed from French in the eighteen-thirties. Until the middle of the 18th century it was used sparsely in political writing (10 occurrences in *DTA* from 1833 to 1849). Only since 1855 it started occurring more frequently in more domains. This increase is paralleled in *MKHZ*, where its frequency reaches a peak between 1848 and 1852 with 51 occurrences, with 73 occurrences overall. This again suggests that the language of newspapers may have served as a mediator, disseminating a specialized foreign word into standard German. In comparison, *DBG01* contains only 26 occurrences between 1848 and 1905.

5. Conclusions and Future Work

We have presented the Mannheim Corpus of Historical Newspapers and Magazines, and on this basis investigated the role of newspapers for popularizing foreign words in German. Currently we are working on transforming another 2500 pages from TUSTEP to TEI P5 to improve its coverage. Moreover, we cooperate with the Berlin Brandenburg Akademie der Wissenschaften to integrate the corpus into *DTA*, which will enable us to use the more advanced techniques to normalization employed by the *DTA* for our empirical analysis.

Lexical change is not confined to the introduction of foreign words. In the course of the lexical and semantic change in German during the early 19th century many words from middle and early modern German – in particular in poetic literature – became obsolete or changed their meaning (Beutin, 1972), leading to the final norm of modern German also in the lexical domain. So far the reasons for this change have not been described systematically, e.g., which were the driving social forces, or in which genres and registers did this change occur first. Like with respect to foreign words, newspapers as a fairly new medium could have played an important role in this change. To help tracking possible semantic innovation in newspapers, we plan to systematically analyze and compare the local contexts of words in various registers and over time.

Finally, we also plan to compare the historical newspaper corpus with contemporary newspapers to get a better understanding of the evolution of news as a genre, looking at effects of conventionalization and diversification w.r.t. other genres.

[9] Amt, Funktion, Auftrag eines Bevollmächtigten, Vermittlers, Vertreters

[10] Geschäftsstelle, Büro eines Agenten

Acknowledgements

The curation of MKHZ has been partially supported by DFG (Deutsche Forschungsgemeinschaft; area: Wissenschaftliche Literaturversorgungs- und Informationssysteme, project: Zentrum für germanistische Forschungsprimärdaten).

6. References

Cyril Belica. 1994. WP2 – Lemmatizer, Final Report. Technical report, Institut für Deutsche Sprache, July.

Paul Bennett, Martin Durrell, Silke Scheible, and Richard J. Whitt. 2010. Annotating a historical corpus of german: A case study. In *Proceedings of the LREC 2010 workshop on Language Resources and Language Technology Standards*, pages 64–68, Valletta, Malta, May.

Wolfgang Beutin. 1972. *Das Weiterleben alter Wortbedeutungen in der neueren deutschen Literatur bis gegen 1800.* Lüdke, Hamburg.

Franck Bodmer. 2005. COSMAS II. Recherchieren in den Korpora des IDS. *Sprachreport*, 3:2–5.

Daan Broeder, Oliver Schonefeld, Thorsten Trippel, Dieter Van Uytvanck, and Andreas Witt. 2011. A pragmatic approach to XML interoperability - the Component Metadata Infrastructure (CMDI). In *Proceedings of Balisage : The Markup Conference 2011*, volume 7. Balisage Series of Markup Technologies.

Peter Fankhauser, Norman Fiedler, and Andreas Witt. 2013a. Forschungsdatenmanagement in den Geisteswissenschaften am Beispiel der germanistischen Linguistik. *Zeitschrift ZfBB, Zeitschrift für Bibliothekswesen und Bibliographie*, 60, December.

Peter Fankhauser, Oliver Pfefferkorn, and Andreas Witt. 2013b. From TUSTEP to TEI in Baby Steps. In Fabio Cotti and Arianna Ciula, editors, *Abstracts of the TEI Conference and Members Meeting*, pages 34–38, Rome, October. DIGILAB Sapienza University & TEI Consortium.

Alexander Geyken, Susanne Haaf, Bryan Jurish, Matthias Schulz, Jakob Steinmann, Christian Thomas, and Frank Wiegand. 2011. Das Deutsche Textarchiv: Vom historischen Korpus zum aktiven Archiv. In Silke Schomburg, Claus Leggewie, Henning Lobin, and Cornelius Puschmann, editors, *Digitale Wissenschaft. Stand und Entwicklung digital vernetzter Forschung in Deutschland, 2010, 20./21.September*, pages 157–161. 2. ergänzte Fassung, hbz.

Thomas Gloning. 1996. Bestandsaufnahme zum Untersuchungsbereich ”Wortschatz”. In Gerd Fritz and Erich Straßner, editors, *Die Sprache der ersten deutschen Wochenzeitungen im 17. Jahrhundert*, pages 164 – 195. deGruyter, Tübingen.

Ulrich Püschel. 1999. Präsentationsformen, Texttypen und kommunikative Leistungen der Sprache in Zeitungen und Zeitschriften. In Joachim-Felix Leonhard, Hans-Werner Ludwig, Dietrich Schwarze, and Erich Straßner, editors, *Medienwissenschaft. Ein Handbuch zur Entwicklung der Medien und Kommunikationsformen*, volume 1, pages 865 – 868. deGruyter, Berlin, New York.

Hans Schulz and Otto Basler. 1942. *Deutsches Fremdwörterbuch. B2: L–P.* Walter de Gruyter, Berlin.

Gerhard Strauß, Elke Donalies, Heidrun Kämper-Jensen, Isolde Nortmeyer, Joachim Schildt, Rosemarie Schnerrer, and Oda Vietze. 1995. *Deutsches Fremdwörterbuch. Bd. 1, a-Präfix Antike. Völlig neubearbeitet im Institut für Deutsche Sprache*, volume XVII. de Gruyter, Berlin/New York, 2 edition.

Tina Theobald. 2012. ”Dieses unselige Zeitungsdeutsch” Reflexion über die Presse und ihren sprachlichen Einfluss im 19. Jahrhundert. *Sprachreport*, (3):12 – 21.

TUSTEP. 2013. Handbuch und Referenz (electronic version). Technical report, Universität Tübingen; Zentrum für Datenverarbeitung.

Jürgen Wilke. 1999. Die Zeitung. In Ernst Fischer, Wilhelm Haefs, and York-Gothart Mix, editors, *Von Almanach bis Zeitung. Ein Handbuch der Medien in Deutschland 1700 - 1800*, pages 388 – 402. C.H. Beck, München.

Digital libraries of historical Icelandic newspapers, periodicals, magazines and old printed books

Örn Hrafnkelsson, Jökull Sævarsson
National and University Library of Iceland
Reykjavík, Iceland
orn@landsbokasafn.is, jokull@landsbokasafn.is

Abstract

The National and University Library of Iceland has since its opening in 1994 actively pursued the digitalization of its national collections and made them freely available on the internet. Two projects, the digitization of historical newspapers, magazines and periodicals (Timarit.is), and the digitization of historical Icelandic printed books (Baekur.is), open up new methods for the academic world in its research. The metadata and the text of the material are available to our users.

Keywords: digital library, digital humanities, open access

1. Introduction

This paper describes two open access digital libraries that are maintained and developed by the National and University Library of Iceland – NULI[1]. The first one consists of retroactive digitalization of historical Icelandic newspapers, magazines and periodicals[2] and the second of old Icelandic printed books[3]. The intention is to give access to historical printed material which is out of copyright or the library has a written agreement from the publishers or copyright holders with permission to digitize and distribute the work online.

Work on the historical newspaper, periodicals and magazines from 18th and 19th century project started in 1999. The site was opened to the public in the September 2002 with 71 titles and around 40 thousand pages. Today it consists of 870 titles with about 4.5 million pages. The text is OCR-read and made searchable with free text search. The digital library of old printed books was opened to the public in December 2010. Today it consists of 1.000 titles with around 300 thousand pages.

The National Library of Iceland (established 1818) and the Library of the University of Iceland (established 1940) were amalgamated in 1994 as the National and University Library of Iceland – NULI. When the changes occurred, an opportunity to transform operations and services opened. It was done by rationalizing and modernizing the functions of the two libraries in a new building and developing new strategies and priorities (Sigurðsson, 2000). Since then the mission of the library has been to serve users in a new way. Access has been given to both primary and secondary sources within the collections, sources which sometimes are unique or the only exemplars known. A short description of NULI's digital activities for the last years can be read in Sverrisdóttir (2012).

The paper is structured as follows. In Section 2, we describe briefly the idea and the purpose of the newspaper, magazine and periodical digital library. In Section 3, we describe briefly the idea and the purpose of the old Icelandic printed books project. The availability and use of the material in both projects is described in Section 4. In Section 5 there is discussion about how the library views the digitalization projects as a part of its conservation policy. Finally, there are few closing words in Section 6.

2. Digital library of historical newspapers

The idea behind the project started after staff of the National Library had been sitting in on project meetings with representatives from other Nordic national libraries about digitizing historical newspapers from microfilms and giving access to them on library websites. The name of the project was TIDEN and it was active from 1998 to 2001 (Bremer-Laamanen, 2006). After some consideration in NULI it was decided to start another project and digitize the Icelandic material directly from the originals. That was mostly due to the bad condition of the Icelandic microfilm collection – the films were old and fragile, they did not fulfill accepted standards and measurement of sharpness of microfilms to be digitized. Probably more importantly, only a small part of the collection had been microfilmed. In the first project of this kind in the library (started in 1996) which involved digitalization of its historical map collection, photographic slides of the maps were sent to the National Library of Norway, where they were converted to digital format. But by transferring the images from one medium to another we experienced loss of image quality. Therefore it seemed a reasonable choice to digitize directly from the originals. Another aspect to take into consideration is that the library had at that time recently bought a new digital camera for digitizing manuscripts as part of the SAGANET project which goal was to digitize Icelandic Sagas and related material that is kept in The National Library, The Árni Magnússon Institute in Iceland and in the Fiske Collection at Cornell University (Hallgrímsson, 2006).

The national libraries of the Faroe Islands and Greenland where involved in the historical newspaper project from the start and still are (Hrafnkelsson, 2001).

2.1 Selection of material

It was decided from the start to digitize all printed Icelandic material up to 1920, both printed in Iceland and abroad, in Icelandic or other languages, if it could in any

[1] http://landsbokasafn.is
[2] http://timarit.is
[3] http://baekur.is

way be related to Iceland or Icelanders. The selection was based on a periodical index compiled by Böðvar Kvaran and Einar Sigurðsson (1991) for the period from 1773 to 1973. 1920 was chosen as the end-year it was thought of as a safe year regarding copyright.

The first Icelandic periodical was *Islandske Maaneds-Tidender* 1773–1776 with the exception of printed proceedings of the parliament, first published in 1696. When nearly finished with material published in up to 1920 it was decided to extend the time span to the year 1940. Additionally to this many titles published in the later half 20th and 21st century have also been digitized and made available. This was made possible by signing contracts with individual publishers or copyright holders. Major stepping stones were contracts with publishers of the biggest newspaper, who sponsored the digitization of their own material, and Alþingi, the Icelandic Parliament which sponsored the digitalization of all the main historical newspapers of the 20th century. Therefor we have all major Icelandic newspapers available online free of charge. Some of them are accessible up till today and others with a three year delay outside of the library but accessible inside it.

The first approach of the project was to digitize all the newspapers of the 19th century and magazines and periodicals published in the 18th and 19th century. The main argument for this selection of material was that the originals could only be accessed in a few research libraries across Iceland and microfilms in even fewer places. This phase of the project included material published in Canada in Icelandic and by Icelanders.

According to working plans it is estimated that the total number of titles published between 1773 to 1940 is around 1.600. Currently 840 titles of Icelandic origin have been digitized.

In the coming years the plan is to start digitizing Icelandic peer-reviewed journals where the publisher gives permission to do so, continue with the material up to 1940 and to digitize local newspapers and magazines published across the island.

Period	Number of titles digitized	Number of pages digitized
1773–1800	5	6,761
1801–1850	20	15,010
1851–1900	136	115,903
1901–1950	598	873,922
1951–2000	249	2,786,158
2001–	74	732,269

Table 1: Breakdown of the material available on Timarit.is.

2.2 OCR

It was decided from the start to OCR-read and enable free text search for all the material. In the beginning of the project the OCR-reading did not take place at the same time as digitization; there was always some backlog of material to process and the results of the reading were not satisfactory. The delay was due to how the production lines of the project where structured at that time. This is no longer the case; as soon as an item has been digitized and images approved for display it is OCR-read. This is an automatic process. The Russian program ABBYY FineReader has from the start been used for OCR-reading. It was primarily selected because it recognized Icelandic letters like *þ* and *ð*. In the very beginning of the project training-logs where made for improving results but as the material to process is so huge–many pages–it was decided to just use the program as it comes of the shelf. Improvements in the FineReader software greatly aided this move.

It is worth mentioning that only prints with latin letters are OCR-read; no OCR-reading is done with material printed with gothic letters as the FineReader software has been unable to cope with them. The latest version of FineReader has added support for such texts and we will be addressing this in the near future.

2.3 Free text search

Free text search of the material has from year 2005 been available as a part of the project. In the last update of the whole site in 2008 several new search features were implemented. Firstly, users can now search in Timarit.is by headwords (lemmas) and get results with the inflectional words and therefore get better results in their research. To implement this complex search feature the Database of Modern Icelandic Inflection of The Árni Magnússon Institute for Icelandic Studies was used (Kristín Bjarnadóttir, 2012). Secondly, when displaying free text search results users can narrow the results based on facets such as titles and from what year or decade the results come from.

2.4 Usage of the material

The text of the scanned material is, as said earlier, extracted by using OCR-technology and is searchable both on the site and in Google. When it was opened for Google, and other search-engines the number of users and usage more than doubled in a very short time.

The text search, especially in the newspapers, has been very well received by scholars and the public alike. This search has opened up the material for users and instead of having to browse through many volumes of text when they are trying to find information about a particular subject or an individual they can now let the search engine look for key words.

The site is also frequently used for citations in Wikipedia. Timarit.is is by far the most popular collection of NULI (digital or otherwise) with over 15 thousand unique visitors each week. It is one of the thirty most used sites in Iceland (Modernus, 2012).

3. Digital library of old Icelandic printed books

The library policy on digitalization projects has always been to digitize whole collections. The digital library of old Icelandic printed books is no exception from that rule. The scope of the project is to digitize all printed books from the start of printing in Iceland in 1534 to 1844, when the only printing press in the country was moved from Viðey to Reykjavík. The first book published in Icelandic is a translation of the New Testament printed in Roskilde, Denmark, in 1540. A few years earlier printing had started in Iceland at the episcopal seat Hólar. But nothing survives from that time except for two folios from a breviary (Breviarium Holense, 1534). Many Icelandic books were also printed in other countries during this period especially in Copenhagen, Denmark, because Iceland was for a long time a Danish dependency. In all there are about 1,750 known titles from the hand-press printing period up to 1844. This includes titles that are only known from external sources.

Information about these books is taken from an unpublished descriptive catalogue of Icelandic bibliography up to 1844 which the library has been working on for the last few years. In this bibliography we have detailed cataloguing of every publication from this period by Icelandic authors, books by foreign authors printed in Icelandic and books by foreign authors in foreign languages if they have been printed in Iceland.

Each book is digitized as a whole object. Not just the pages with printed text but also fly-leafs, all edges and binding.

The aim in the first phase of the book project is to digitize all titles or copies that are owned by the library, later on the library will contact owners of other copies and discuss if they are willing to digitize them for us. Of the about 1,750 titles printed before 1844 NULI has in its collection nearly 1,390. Some libraries that have extensive holdings of old Icelandic material have already digitized their collections like The Royal Library in Denmark. Discussions are ongoing about including the Icelandic material in the database in the future to complete it.

Nearly all titles up to the middle of the 18th century that are held by NULI have now been digitized. Later this year titles that where published from 1750 to 1844 will be made available. In some instances more than one copy of each book has been digitized. This applies mostly to the oldest books where we are often working with incomplete copies with pages missing. A few books also have printing variants, different title-pages, fonts etc.

Period	Number of titles digitized	Number of pages digitized
1530–1600	29	11,424
1601–1650	55	16,910
1651–1700	123	31,290
1701–1750	144	42,570
1751–1800	353	74,522
1801–1850	160	44,512

Table 1: Breakdown of the material available on Baekur.is

Cataloguing information for the books is now extracted from Gegnir.is–the Icelandic national catalogue of library information. In the future this information will be augmented by more detailed cataloguing from the Icelandic bibliography mentioned earlier. That list is now being encoded using TEI P5[4].

As mentioned in Section 2.2 there is at the moment no OCR-work has been done on the material that is printed in gothic. Next year we expect to update our OCR software to a version capable of reading gothic fonts. At that point we will retroactively process all books printed with a gothic font. OCR text from books with Latin fonts are already searchable on Baekur.is

4. Availability and use of material

It was decided from the start of both projects, and others by NULI, to have the material open and free for use. Users are made aware that some of the material is protected by copyright-law and that they must use it according to law and seek permission for all reusing. All the metadata is open and available, structural cataloguing and detailed cataloguing.

The OAI/PMH – Open Archives Initiative / Protocol for Metadata Harvesting – methods are used to enable centralized search services to index both digital libraries irrespective of cataloguing methods, and to enable the library to exchange metadata with its cooperation partners.

The corpus of the text of the OCR-read material has not yet been made available, but it has been taken under consideration, especially as it is known to be of great value for linguistic studies.

5. Conservation

The library views the digitization projects as a part of its conservation policy. Old books, periodicals and especially newspapers can be hard to handle. They are often fragile and can be sensitive to external influences. Things like air, heat, moisture, light, vermin, acid, mold and fungi can cause them to deteriorate. Not to mention people who can cause extensive damage by careless treatment.

NULI wanted to protect its collections of rare items by putting them in permanent storage and using the originals as little as possible. Detailed cataloging of them can never replace the items themselves. When the user is looking for books and periodicals about a particular subject he would nearly always want to have access to them rather than descriptions in catalogues. Therefore the best preservation policy is to make digital copies and allow the user access to them instead of the originals. That will make the collections much more accessible and make it easier for researchers to study them while keeping the originals safe from harm.

6. Conclusion

It can be said without doubt that the newspaper project has had great impact on research in the humanities in Iceland. Researchers now have access to more material right from their desks than ever before and by doing free text searches they are able to dig into the material in ways that were never before possible. This will only improve as

[4] http://www.tei-c.org/

more material will is digitized, and the puzzle completed with all printed Icelandic newspapers, journals and magazines online. The same can also be said about the old books project. All Icelandic books from the hand-press period, owned by NULI, will be available before the end of this year, and when OCR-reading of gothic print improves and the text will be searchable. That will doubtlessly change the way how research is done.

7. Acknowledgements

The newspaper project received in the start a grant from NORDINFO and was sponsored by several parties in each of the participating countries, such as the Icelandic Research Council (RANNÍS). The Alþingi–Icelandic parliament–has twice funded special digitalization tasks within the project.

8. References

Bjarnadóttir, Kristín. (2012). The Database of Modern Icelandic Inflection. *LREC 2012 Proceedings*: Proceedings of "Language Technology for Normalization of Less-Resourced Languages", SaLTMiL 8 -- AfLaT 2012.

Bremer-Laamanen, M. (2006). A Nordic Digital Newspaper Library. In Hartmut Walravens (Ed.), *International newspaper librarianship for the 21st century*. München: K.G. Saur, pp. 251--256.

Hallgrímsson, Þorsteinn. (2006). Sagnanet – SagaNet. *Nordisk tidskrift för bok- och bibliotekshistoria* 6, pp. 235--245.

Hrafnkelsson, Örn. (2001). The VESTNORD project: digitizing newspapers and magazines from the 18th and 19th centuries. In *Nordiskt Forum for forsknings-bibliotekschefer*, pp. 131--138. Helsingfors: Nordinfo.

Kvaran, Böðvar, Sigurðsson, Einar. (1991). Íslensk tímarit í 200 ár : skrá um íslensk blöð og tímarit frá upphafi til 1973 = 200 years Icelandic periodicals : a bibliography of Icelandic periodicals, newspapers, and other serial publications 1773-1973. Reykjavík: [s.n.].

Modernus. 2012. "Coordinated webmeasure." Accessed September 4. http://veflistinn.is/

Sigurðsson, Einar. 2000. The next ten years in national libraries: the National and University Library of Iceland. *Alexandria* 12(2): pp. 134--135.

Sverrisdóttir, Ingibjörg Steinunn, Sigurðsson, Kristinn, Hrafnkelsson, Örn (2012). Access and Curation of Digital Cultural Heritage in the National and University Library of Iceland. *Microform & Digitization Review* 41(3-4), pp. 97--102.

Historical Newspapers & Journals for the DTA

Susanne Haaf, Matthias Schulz
Deutsches Textarchiv (DTA) at the Berlin-Brandenburg Academy of Sciences and Humanities (BBAW)
Jägerstr. 22/23, Berlin, Germany
E-mail: haaf@bbaw.de, mschulz@bbaw.de

Abstract

In this paper we present work in progress on the digitization of historical German newspapers in the context of the DFG-funded project Deutsches Textarchiv (DTA) at the Berlin-Brandenburg Academy of Sciences and Humanities. Currently, the DTA core corpus consists of a selection of 1,300 works (ca. 419,000 pages; ca. 97 million tokens) which is balanced with regard to time of creation, genre and theme. It contains mostly monographic works from different disciplines and genres as well as a selection of scientific articles. Though it includes a balanced selection of relevant functional literature, newspaper texts have not been in the scope of text digitization for the DTA, yet. However, the text type newspaper constitutes a significant representative of widespread functional literature and is thus an important source for the usage and development of the German colloquial language. Thus, in the course of the module DTA Extensions (DTAE) we are currently working on the integration of historical newspapers which were digitized by other, external projects. As examples for the integration of historical newspapers into the DTA we here present four different cooperation projects currently maintained by the DTA.

Keywords: Historical Newspapers, Historical Corpora, XML/TEI, Encoding Formats, Homogeneous Text Annotation

1. Introduction

In this paper we present work in progress on the digitization of historical German newspapers in the context of the DFG-funded project Deutsches Textarchiv (DTA) at the Berlin-Brandenburg Academy of Sciences and Humanities.[1]

The goal of the DTA is to provide the basis for a reference corpus which reflects the development of the historical New High German language (1600–1900). For this purpose, since 2007 the DTA has been building up a core corpus of currently 1,300 historical German works (with another 200 works to come until the end of 2014) dating back to the 17th to 19th century.

The DTA core corpus contains fictional and non-fictional printed texts, the latter including scientific as well as functional literature. In order to document the state of the German language at different points in time and for different discourses, we focus on the first or, if this is not possible, on early editions of the respective works. The text selection for the DTA core corpus is based on a bibliography, which has been developed at the beginning of the project according to the guiding principle to create a corpus which is balanced with regard to time of creation, text type, and thematic scope.

At present (April 2014), the DTA core corpus comprises ca. 419,000 pages (97 million tokens) of mostly monographic works from different disciplines and genres as well as a selection of scientific articles. Though it includes a balanced selection of relevant functional literature, newspaper texts have not been in the scope of text digitization for the DTA, yet. However, the text type *newspaper* constitutes an important source for the usage and development of the German colloquial language. Thus, we are currently working on the integration of historical newspapers which were digitized by other, external projects. This work is performed in the context of the module DTA Extensions (DTAE), in the course of which external texts are converted into the standardized input format of the DTA, the DTA ›Base Format‹ (DTABf), and are then further processed by the DTA tools and made available within the DTA corpus as a whole.[2] This way, genuinely heterogeneous resources are provided in a homogeneous format, enabling users to explore them collectively by similar methods and with reliable and reproducible research results.

Currently, the integration of newspapers into the DTA comprises resources from several different projects. These historical sources were digitized according to individual project guidelines and text digitization formats. In this paper, we introduce four of these resources and explain the respective workflows which were set up for the gradual integration of these resources into the DTA. We also show how the TEI/P5 based DTABf was cautiously adjusted in order to meet specific structuring necessities of newspaper texts while remaining firm and unambiguous concerning the annotation of historical texts across the various DTA subcorpora.

2. Historical Newspapers for the DTA

As examples for the integration of historical newspapers into the DTA we here present four different cooperation projects currently maintained by the DTA:

- Institut für Deutsche Sprache (Institute for the German Language, IDS) Mannheim: Digitization of selected issues of different historical newspapers (i. e. *Mannheimer Korpus Historischer Zeitungen und Zeitschriften*);
- Marx-Engels-Gesamtausgabe (The Complete Works of Marx and Engels; MEGA) at the Berlin-Branden-

[1] Cf. the project website http://www.deutschestextarchiv.de/.

[2] The DTA Extensions corpus currently comprises ca. 1,000 volumes with 100 million tokens. For more information about the module DTAE cf. http://www.deutschestextarchiv.de/dtae/.

burg Academy of Sciences and Humanities (BBAW) Berlin: Digitization of the newspaper *Neue Rheinische Zeitung*;
- University of Paderborn with the DTA as associated partner: Digitization of parts of the *Staats- und Gelehrte Zeitung des Hamburgischen unpartheyischen Correspondenten*;
- joint DFG-project of the SuUB Bremen and the BBAW: Digitization of the newspaper *Die Grenzboten*.

These resources were digitized using different digitization methods (manual transcription as well as OCR) and were encoded in different primary formats (TUSTEP, XML/TEI, XML, IDS-XCES). They were primarily digitized without or with (partial) involvement of the DTA. Therefore, different workflows are necessary for the conversion of these resources into the DTABf in order to make them interoperable and interchangeable with one another and with the remaining resources provided within the DTA corpora.

After the conversion process all newspapers will first be provided within the DTA quality assurance platform DTAQ,[3] and, after a correction phase, will be published under a Creative Commons license on the DTA website. Each document is provided with extensive metadata according to a homogeneous metadata format.[4] All resources will be available for download in different text and metadata formats, but can also be explored on the DTA plattforms by usage of the powerful linguistic search engine DDC with regard to the structural, linguistic[5] and metadata information they contain.

3. Projects

In the following chapter we present four different newspaper resources currently being attended to by the DTA and we describe the individual workflows which were set up for the integration of these resources into the DTA infrastructure.

3.1 From TUSTEP to DTABf: Mannheimer Korpus Historischer Zeitungen und Zeitschriften

The *Mannheimer Korpus Historischer Zeitungen und Zeitschriften* (Mannheim Corpus of Historical Newspapers; MKHZ) which has been built up at the Institut für Deutsche Sprache (IDS) Mannheim consists of 652 complete issues of 21 different newspapers and magazines, which were published in the 18th, 19th, and early 20th century. The selection was based on the goal to document the German newspaper language of the respective centuries.[6]

The selected issues were digitized manually using the double keying method and were annotated according to the TUSTEP format. These TUSTEP text sources were then further processed at the IDS according to different scenarios for their subsequent use,[7] including their conversion into a reduced version of the DTABf. The resulting DTABf texts are currently being further corrected and structured semi-automatically, and are being integrated into the DTA corpora successively.

The most significant adjustment necessary is the addition of article structures: In the original text sources the main text container elements were divisions corresponding to the columns on a newspaper page. The DTABf, however, only considers the markup of column breaks as milestones whereas text divisions are defined semantically, and thus in the case of newspapers contain articles and article sections. These article structures have hence to be applied to the text ex post. In most cases, it is possible to support this manual work on article structuring by automatically interpreting certain typographic specifics, which were already annotated in the text, considering their semantic meaning. For instance, considering that article headings in newspapers are often typed as centered text, it is possible to provide the editor with suggestions about where a new article might be beginning. Nevertheless, there remains a considerable amount of manual (editorial) work, including the evaluation of the proposed suggestions on the basis of the facsimile and text content as well as the attribution of article types according to the DTABf classification for newspaper articles.

Apart from this there are further post-processing steps undertaken towards a deeper semantic annotation than the one gained by automatic conversion so far. These steps include for example the DTABf conformant re-structuring of the newspaper's heading including its title, number, date and place of publication, etc., the annotation of author names as given at the beginning or end of articles, or the structuring of poems.

Figure 1 shows one page of the already completed issue 5 of the journal "Der allerneuesten Europäischen Welt- und Staats-Geschichte", 2nd part, as it now appears within the quality assurance platform DTAQ.

3.2 From XML to DTABf: Neue Rheinische Zeitung

The *Neue Rheinische Zeitung* (NRhZ) was published in 301 issues with a total amount of 1,718 pages between June 1st 1848 and May 19th 1849, i. e. during the time of the bourgeois revolution in Germany. The newspaper was published in Cologne, the editors being Karl Marx and Friedrich Engels.

This resource has been digitized is its entirety as a by-product of the project "Marx-Engels-Gesamtausgabe" (MEGA) at the BBAW. Text digitization here was twofold: For the most part, the issues were transcribed based on digital images by native German speakers using the double-keying method. The transcribers also performed the annotation of text structures using an individual XML format. All articles written by Karl Marx and Friedrich Engels, however, were transcribed by staff members of the MEGA project in the course of their editorial work and by usage of the TUSTEP format. While the former texts by now have been converted and post-processed

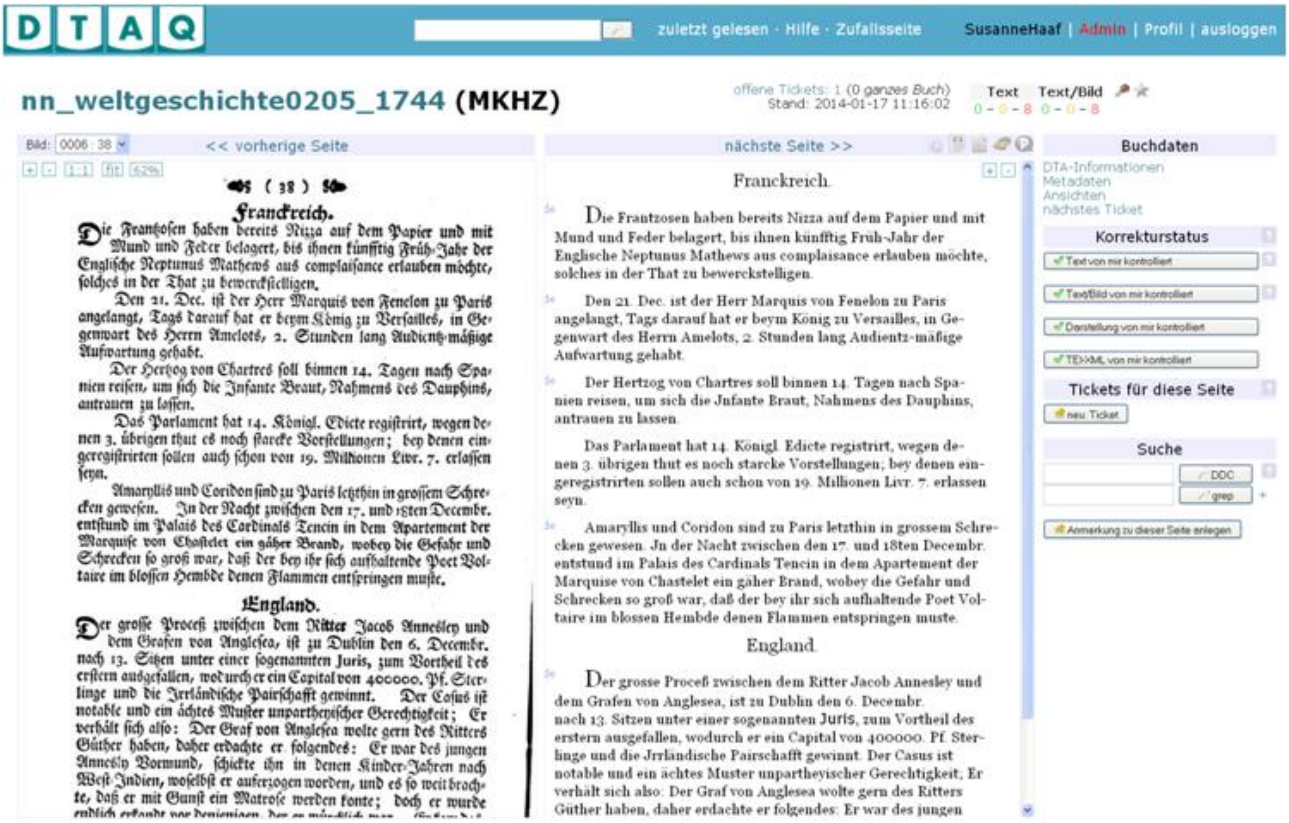
[3] http://www.deutschestextarchiv.de/dtaq;
cf. http://www.deutschestextarchiv.de/dtaq/about.

[4] Cf. http://www.deutschestextarchiv.de/doku/basisformat_header.

[5] Each DTA text is enriched with automatically gained linguistic information such as canonical vs. contemporary form, part-of-speech tag, and lemma. The linguistic information is applied as stand-off markup. For more information about linguistic analyses carried out at the DTA cf. http://www.deutschestextarchiv.de/doku/software.

[6] Cf. Fankhauser et al. 2013.

[7] Ibid.


DTAQ
zuletzt gelesen · Hilfe · Zufallsseite
SusanneHaaf | Admin | Profil | ausloggen
nn_weltgeschichte0205_1744 (MKHZ)
offene Tickets: 1 (0 ganzes Buch)
Stand: 2014-01-17 11:16:02
Text
Text/Bild
0 - 0 - 8
0 - 0 - 8
Bild: 0006 : 38
<< vorherige Seite
nächste Seite >>
1:1
fit
62%
Buchdaten
DTA-Informationen
Metadaten
Ansichten
nächstes Ticket
Korrekturstatus
Text von mir kontrolliert
Text/Bild von mir kontrolliert
Darstellung von mir kontrolliert
TEI-XML von mir kontrolliert
Tickets für diese Seite
neu Ticket
Suche
DDC
grep
Anmerkung zu dieser Seite anlegen

Franckreich.

Die Frantzosen haben bereits Nizza auf dem Papier und mit Mund und Feder belagert, bis ihnen künfftig Früh-Jahr der Englische Neptunus Mathews aus complaisance erlauben möchte, solches in der That zu bewerckstelligen.

Den 21. Dec. ist der Herr Marquis von Fenelon zu Paris angelangt, Tags darauf hat er beym König zu Versailles, in Gegenwart des Herrn Amelots, 2. Stunden lang Audientz-mäßige Aufwartung gehabt.

Der Hertzog von Chartres soll binnen 14. Tagen nach Spanien reisen, um sich die Jnfante Braut, Nahmens des Dauphins, antrauen zu lassen.

Das Parlament hat 14. Königl. Edicte registrirt, wegen denen 3. übrigen thut es noch starcke Vorstellungen; bey denen eingeregistrirten sollen auch schon von 19. Millionen Livr. 7. erlassen seyn.

Amaryllis und Coridon sind zu Paris letzthin in grossem Schrecken gewesen. Jn der Nacht zwischen den 17. und 18ten Decembr. entstund im Palais des Cardinals Tencin in dem Apartement der Marquise von Chastelet ein gäher Brand, wobey die Gefahr und Schrecken so groß war, daß der bey ihr sich aufhaltende Poet Voltaire im blossen Hembde denen Flammen entspringen muste.

England.

Der grosse Proceß zwischen dem Ritter Jacob Annesley und dem Grafen von Anglesea, ist zu Dublin den 6. Decembr. nach 13. Sitzen unter einer sogenannten Juris, zum Vortheil des erstern ausgefallen, wodurch er ein Capital von 400000. Pf. Sterlinge und die Jrrländische Pairschafft gewinnt. Der Casus ist notable und ein ächtes Muster unpartheyischer Gerechtigkeit; Er verhält sich also: Der Graf von Anglesea wolte gern des Ritters Güther haben, daher erdachte er folgendes: Er war des jungen

Figure 1: Der allerneuesten Europäischen Welt- und Staats-Geschichte II. Theil. No. V, 3. Woche, Erfurt (Thüringen), 13. Januar 1744, p. 38. In: Deutsches Textarchiv Qualitätssicherung, <http://www.deutschestextarchiv.de/dtaq/book/view/30541?p=6>.

according to the DTABf guidelines, the integration of the latter texts into the thus existing DTABf conformant NRhZ sub-corpus is currently being conducted. Those texts which will also be a part of the MEGA edition are provided with additional bibliographic references to the MEGA publication.

So, the predominant part of the NRhZ text corpus was primarily encoded using an individual XML format. The resulting XML texts were converted (semi-)automatically into the DTABf. Structures which had not been annotated in the course of text transcription but are required or recommended by the DTABf since they play a significant role for the semantics of a text, were added manually ex post.

This work included for instance the linking of discontinuous parts of articles, which were separated within the source documents either by new articles on the same page (case 1) or because they were continued/begun in a following/previous issue (case 2). Case 1 occurs, if a new article or section, respectively, is inserted at the bottom of a page, thus interrupting a running article which is then generally continued on the next page. Since text transcription according to the DTA guidelines is carried out page-wise and (wherever possible) under preservation of the layout in the text source, the interrupting article was transcribed at the place of its occurrence. Of course, for further corpus analyses it is necessary to connect the parts of such discontinuous articles. Case 2 occurs, if an article, e. g. within the feuilleton, was continued within the following newspaper issue(s). Connecting the discontinuous parts of articles helps users to research the respective articles as a whole. Therefore, all discontinuous articles within the NRhZ corpus were linked by usage of the TEI attributes `@xml:id`, `@prev` and `@next`. This rather sophisticated task could not be done automatically, but had to be conducted manually with regard to the content as well as the respective facsimiles.

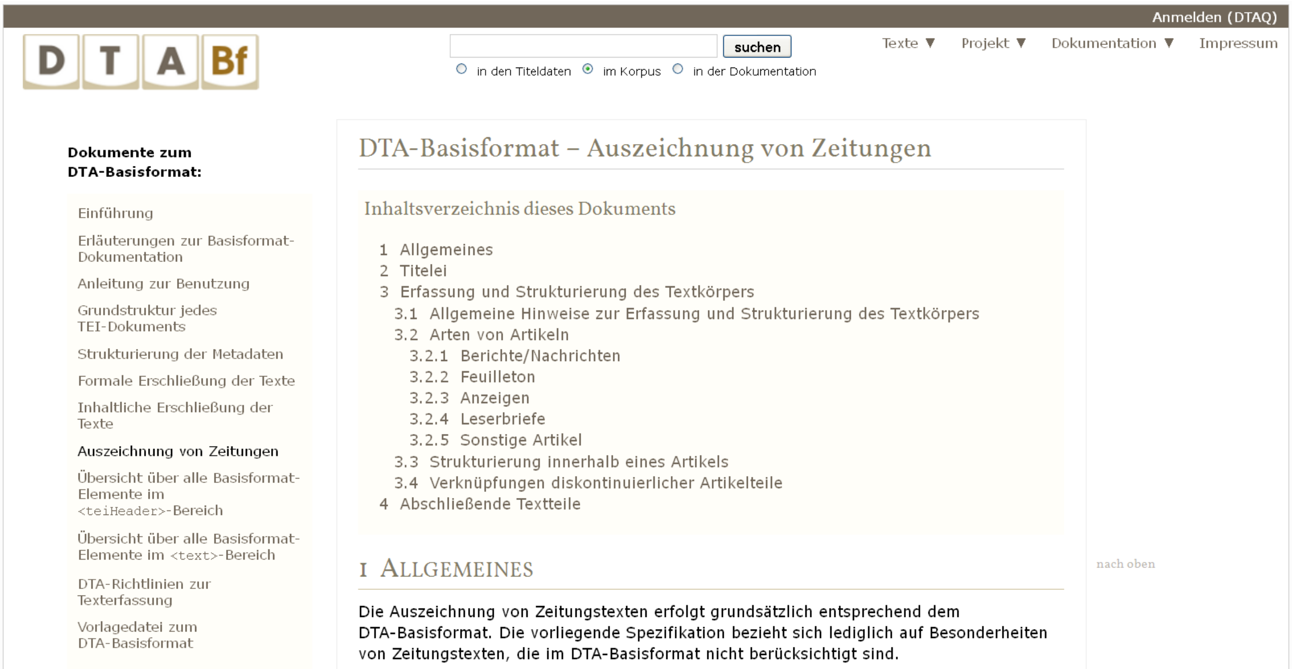
Furthermore, some shallow annotations within the source transcriptions had to be revised and concretized in order to meet the requirements of the DTABf. This work included for instance the tagging of poems, as well as (within reason) the further structuring of announcements which show quite manifold layouts. Moreover, similar to the MKHZ, the tagging of bibliographic information as given within the heading at the beginning of each paper as well as information about the authors of articles was included in the subsequent text structuring procedure.

3.3 Digitization According to the DTA Workflow: *Hamburgischer Correspondent*

In the course of a project at the University of Paderborn 320 selected issues of the *Staats- und Gelehrte Zeitung des Hamburgischen unpartheyischen Correspondenten* and its predecessors published between 1712 and 1851 are being digitized.[8] The *Hamburgischer Correspondent* newspaper was the first daily newspaper in Hamburg with supraregional reach.

This digitization project was already planned in cooperation with the DTA; thus, text digitization is genuinely conducted according to the DTA workflow.

[8] Cf. http://kw.uni-paderborn.de/institute-einrichtungen/institut-fuer-germanistik-und-vergleichende-literaturwissenschaft/germanistik/personal/schuster/projekte/.



Dokumente zum DTA-Basisformat:

- Einführung
- Erläuterungen zur Basisformat-Dokumentation
- Anleitung zur Benutzung
- Grundstruktur jedes TEI-Dokuments
- Strukturierung der Metadaten
- Formale Erschließung der Texte
- Inhaltliche Erschließung der Texte
- Auszeichnung von Zeitungen
- Übersicht über alle Basisformat-Elemente im `<teiHeader>`-Bereich
- Übersicht über alle Basisformat-Elemente im `<text>`-Bereich
- DTA-Richtlinien zur Texterfassung
- Vorlagedatei zum DTA-Basisformat

DTA-Basisformat – Auszeichnung von Zeitungen

Inhaltsverzeichnis dieses Dokuments

1 ALLGEMEINES

Die Auszeichnung von Zeitungstexten erfolgt grundsätzlich entsprechend dem DTA-Basisformat. Die vorliegende Spezifikation bezieht sich lediglich auf Besonderheiten von Zeitungstexten, die im DTA-Basisformat nicht berücksichtigt sind.

Figure 2: DTABf Specification for the Structuring of Newspapers <http://www.deutschestextarchiv.de/doku/basisformat_zeitungen>

This means, the source images are being prepared for transcriptions by the project staff and are then transcribed and annotated by non-native speakers using the double-keying method. The resulting texts are structured according to a simplified DTABf based XML format and therefore may, for the most part, be converted automatically into the DTABf. Only little manual post-processing of special cases is necessary.

3.4 From Abby-XML to DTABf: *Die Grenzboten*

The journal *Die Grenzboten* appeared weekly between 1841 and 1922 and comprises in 270 volumes with ca. 180,000 pages in total. It has been digitized in its entirety by usage of OCR technology by the SUB Bremen. In the course of a joint DFG project of the SUB Bremen and the BBAW,[9] the OCR output is currently being corrected and re-structured using (semi-)automatic methods. Text structuring is undertaken by the DTA according to the DTABf. For this task the DTA Zoning Tool[10] has been adjusted in order to support semi-automatic structuring. The corrected journal issues are successively being integrated into the DTA infrastructure.[11]

4. DTA Guidelines for the Annotation of Historical Newspapers

Text annotation for the DTA corpora is carried out according to the DTA ›Base Format‹ (DTABf), a strict subset of the TEI P5 tag set.[12] The DTABf has been designed as an annotation schema providing tagging solutions for a wide range of structural phenomena while avoiding ambiguities concerning the handling of similar phenomena. This way, we ensure consistent tagging over the large variety of historical texts the DTA comprises.[13] Though the DTABf has been created upon and applied to the large text basis of the entire DTA core corpus, it remains a “living” format, which has to be adjusted occasionally to new phenomena occurring in new texts which are integrated into the DTA. Such adjustments to the DTABf are carried out very carefully in order to avoid the evocation of ambiguities.[14]

For the structuring of newspaper texts the DTABf turned out to be quite sufficient. Nevertheless, in order to allow for structural analyses of the text material on a deeper level, we extended the DTABf tagset in a way that structures that are specific for the text type newspaper might be encoded as such. Adjustments affected for instance the field of division types, where we added a selection of newspaper and journal specific categories, namely:

- the section of political news: `<div type="jPoliticalNews">`
- the section of financial and economical news: `<div type="jFinancialNews">`
- the feuilleton: `<div type="jFeuilleton">`
- the section of announcements: `<div type="jAnnouncements">`
- the section of weather reports: `<div type="jWeatherReports">`
- an article: `<div type="jArticle">`
- an announcement: `<div type="jAn">`.

Furthermore, the <titlePage> element was extended by the attribute-value pair type=”heading”, indicating that title information about a newspaper issue is given within a heading on the first page of an issue, as opposed to books, which usually have a separate title page.

In addition to these adjustments concerning the DTABf schema, we complemented the DTABf documentation with specific guidelines for the annotation of newspaper

[9] Cf. http://gepris.dfg.de/gepris/projekt/196492153.
[10] Cf. http://www.deutschestextarchiv.de/doku/software#ZOT.
[11] For more information on this project cf. Thomas 2013.
[12] http://www.deutschestextarchiv.de/doku/basisformat.
[13] Cf. Geyken et al. 2012.
[14] Cf. Haaf/Geyken 2013.

issues, explaining the structuring of newspapers according to the DTABf in general as well as the handling of special cases.[15] This documentation is still work in progress, since with the integration of new newspaper texts into the DTA we are likely to come across new phenomena which will then be explained within the guidelines as well.

The screenshot in figure 2 illustrates the DTABf documentation for the annotation of newspapers and journals.

5. Conclusion and Further Prospects

In this paper we presented current efforts to extend the historical corpus of the Deutsches Textarchiv project by historical newspapers as an important representative of widespread functional literature. Based on the presentation of four different cooperation projects, we exemplified how genuinely heterogeneous text resources are being homogenized in terms of format, scope and depth of text annotation, transcription quality and metadata substance, so that they become interoperable and interchangeable.

The work presented here is still in progress, i.e. the respective newspapers are currently processed and published successsively on the DTA platforms. In addition, apart from the projects we introduced in this paper, there are further newspaper and journal resources currently being integrated into the DTA, e.g. Dingler's Polytechnisches Journal.

Thus, the DTA sub-corpus of historical newspapers of the 17th to 19th century will be augmented significantly during the next months, complementing the DTA core corpus by a solid amount of data for this important text type.

6. References

Geyken, A. & Haaf, S. & Wiegand, F. (2012). The DTA 'base format': A TEI-Subset for the Compilation of Interoperable Corpora. In 11th Conference on Natural Language Processing (KONVENS) – Empirical Methods in Natural Language Processing, Proceedings of the Conference. Edited by Jeremy Jancsary. Wien (= Schriftenreihe der Österreichischen Gesellschaft für Artificial Intelligence 5), pp. 383-390. http://www.oegai.at/konvens2012/proceedings/57_geyken12w/57_geyken12w.pdf

Fankhauser, P. & Pfefferkorn, O. & Witt, A. (2013). From TUSTEP to TEI in Baby Steps. In The Linked TEI: Text Encoding in the Web. TEI Conference and Members Meeting 2013. Book of Abstracts. Rome, pp. 38-43. http://digilab2.let.uniroma1.it/teiconf2013/program/papers/abstracts-paper#C121.

Haaf, S. & Geyken, A. (2013). The Lifecycle of the DTA Base Format (DTABf). In The Linked TEI: Text Encoding in the Web. TEI Conference and Members Meeting 2013. Book of Abstracts. Rome, pp. 54-62. http://digilab2.let.uniroma1.it/teiconf2013/program/papers/abstracts-paper#C137.

Thomas, Ch. (2013). Möglichkeiten der Nutzung des Grenzboten mit Werkzeugen des Deutschen Textarchivs (DTA). Talk at the presentation of the digitized journal ‚Die Grenzboten', May 31st 2013, Staats- und Universitätsbibliothek Bremen. http://www.deutschestextarchiv.de/files/DTA-fuer-Grenzboten_20130531.pdf.

6.1 Links

Deutsches Textarchiv DTA: http://www.deutschestextarchiv.de/.

DTA ›Base Format‹ DTABf: http://www.deutschestextarchiv.de/doku/basisformat.

DTABf Specification for the Structuring of Newspapers: http://www.deutschestextarchiv.de/doku/basisformat_zeitungen

DTABf Specification for Metadata Recording: http://www.deutschestextarchiv.de/doku/basisformat_header.

DTA Extensions DTAE: http://www.deutschestextarchiv.de/dtae/.

DTA Quality Assurance Platform DTAQ: http://www.deutschestextarchiv.de/dtaq; http://www.deutschestextarchiv.de/dtaq/about.

DTA Software Description: http://www.deutschestextarchiv.de/doku/software.

Project *Die Grenzboten - Digitalisierung, Erschließung und Volltexterkennung einer der herausragenden deutschen Zeitschriften des 19. und 20. Jahrhunderts*: http://gepris.dfg.de/gepris/projekt/196492153.

Project *Mannheimer Korpus Historischer Zeitungen und Zeitschriften*: http://repos.ids-mannheim.de/fedora/objects/clarin-ids:mkhz1.00000/datastreams/CMDI/content.

Project *Volltextdigitalisierung der Staats- und Gelehrte[n] Zeitung des Hamburgischen Unpartheyischen Correspondenten und ihrer Vorläufer (1712-1851)*: http://kw.uni-paderborn.de/institute-einrichtungen/institut-fuer-germanistik-und-vergleichende-literaturwissenschaft/germanistik/personal/schuster/projekte/.

[15] This documentation is available under the address: http://www.deutschestextarchiv.de/doku/basisformat_zeitungen.

Section 3:

Tools for analysis of historical documents

The Journal *Fjölnir* for Everyone: The Post-Processing of Historical OCR Texts

Jón Friðrik Daðason, Kristín Bjarnadóttir, Kristján Rúnarsson
The Árni Magnússon Institute for Icelandic Studies
University of Iceland
E-mail: jfd1@hi.is, kristinb@hi.is, krr1@hi.is

Abstract

The journal *Fjölnir* is a much beloved and romanticized 19th century Icelandic journal, published in 1835-1847, which is accessible in digitized form in the digital libraries of the National and University Library of Iceland. In the 19th century, Icelandic spelling was not standardized, and the *Fjölnir* texts were used for spelling experimentation. The spelling is therefore very varied. In the project described in this paper, the aim was making the text of *Fjölnir* accessible on the Web, both in the original spelling, and in modern (standardized) spelling, in a version suitable both for scholars and the general public. The modern version serves two purposes. It makes the text more readable for the general public, and it allows the use of NLP tools made for Modern Icelandic. The post-processing of the OCR texts described in this paper was done with the aid of an interactive spellchecker, based on a noisy channel model. The spellchecker achieved a correction accuracy of up to 71.7% when applied on OCR text, and 84.6% when used to normalize the 19th century text to modern spelling.

Keywords: OCR post-correction, historical texts, spelling normalization

1. Introduction

The topic of this paper is a description of a pilot project making historical Icelandic texts accessible to various groups of users. The texts are from the early 19th century journal *Fjölnir*, which will be made available in the original historical spelling, and in modern Icelandic spelling, with links to digitized copies of the originals.[1] The aim is serving both the general public and researchers, such as historians, linguists, etc., by using NLP tools developed for modern Icelandic, while also providing access to the original text. By adding PoS tagging and lemmatization, both the general public and scholars will be able to access the data in an efficient way. For scholars, the modern version will clearly be a secondary one, i.e., a layer to facilitate analysis, but for the general public the modern version removes the irritation of unfamiliar spelling, sometimes found to be prohibitively annoying. As the morphology of Icelandic is quite rich, and the ambiguity of word forms is extensive, PoS tagging and lemmatization are of great importance, even in the most elementary search (Bjarnadóttir, 2012).

Software for the post-correction of modern OCR text was adapted for this project (Daðason, 2012). The result is a web-based spell checking application based on a noisy channel model, which can be used to achieve a true copy of the original spelling of historical texts, and to produce a parallel text with modern spelling. The software is adapted and used with different lexicons and training data for each task.[2] The project described here is also an attempt at creating the infrastructure for an archive of historical Icelandic texts, where the texts are made accessible to various groups of users, as tailor-made resources are not practicable in a tiny language community.

The organization of the paper is as follows: In chapter 2 the choice of the text for the project is described. Chapter 3 contains comments on Icelandic spelling and language cohesion. Chapter 4 contains the body of the paper, describing the process of post-correction, methodology, the noisy channel model, the error model and the language model. The evaluation of the process is in chapter 5. Chapter 6 shows examples from the Web production. Conclusion and thoughts on the future of the project are to be found in chapter 7.

2. The Journal *Fjölnir*[3]

The Icelandic journal *Fjölnir* was published in 1835–1847, and the instigators were four young Icelandic intellectuals in Copenhagen. The topics of the journal were varied, ranging from articles on politics, history, natural sciences (e.g., ornithology, geology, astronomy, and ichthyology), to articles on language and spelling reforms. Book reviews are an important part of *Fjölnir*, and short stories and poetry are also included, both original Icelandic works and translations. Some of the literary texts are among the most exquisite works of the period, known by most Icelanders. The journal appeared in 9 issues, yearly, with intervals, as shown in Table 1.

Fjölnir was chosen for the project described in this paper partly because of its immediate appeal to the Icelanders, as the journal was both very influential in the Icelandic struggle for independence in the 19th century, and also a

[1] The OCR process itself is not a part of the *Fjölnir* project. The *Fjölnir* texts are a part of the digital library of Icelandic newspaper and journals, *Tímarit.is*, produced and maintained by the National and University Library of Iceland (Hrafnkelsson & Sævarsson, 2014). The OCR post-correction described in this paper is also used in the creation of a corpus of early Modern Icelandic (Ásta Svavarsdóttir et al., 2014).

[2] The spellchecker, named Skrambi, is in fact used for other purposes also, i.e., in context sensitive spellchecking for Modern Icelandic.

[3] *Fjölnir* is one of the names of the Norse god Óðinn (Odin).

cornerstone in the evolution of the romantic period in Icelandic literature. An additional reason for choosing this text is that the spelling of the journal poses interesting problems in itself. At the time of the publication of *Fjölnir*, Icelandic spelling was not standardized, and one of the aims of the four original authors of the journal was establishing very drastic spelling reforms and standardization. These proved to be too drastic to be acceptable to the public, and they were in fact only used partly in the first two issues of the journal. The result is that the spelling of *Fjölnir* is extremely varied, and therefore a real challenge in the OCR post-correction process.

Year	*Pages*	*Words*	*Characters*
1835	180	41,951	243,713
1836	108	31,968	185,994
1837	114	34,272	202,851
1838	92	26,186	155,445
1839	186	59,484	343,139
1843	88	15,974	95,381
1844	140	42,646	248,671
1845	84	20,824	121,975
1847	96	22,867	131,365
Total	1,088	296,172	1,728,534

Table 1: Figures for the journal *Fjölnir*.

3. The Cohesion of Icelandic

The Icelandic language community is very small, with approximately 320 thousand speakers, and limited financial resources. It is therefore imperative to be able to use NLP tools made for modern Icelandic for older Icelandic texts (Svavarsdóttir et al., 2014). Developing NLP tools for each period is too costly to be feasible, even for the periods for which there are sufficient texts for such an undertaking to be remotely possible. As Icelandic spelling was not standardized until modern times, variation has to be taken into account anyway, and the method adopted in this project entails using the modern language to anchor all variants of word forms to lemmas in the Database of Modern Icelandic Inflection (Bjarnadóttir, 2012). This is feasible because the cohesion of Icelandic word forms through the history of the language is sufficiently stable to make the modern forms predictable. In fact, experiments with 15th century texts have shown that approximately 40% of the word forms there were identical to forms in the modern language, after the older texts have been transcribed to a modern character set, without a change of spelling.[4] Because of this, spellchecking methods can be used, and a translation system is not needed.

This does not imply that there have not been linguistic changes through the centuries of Icelandic language history. Part of the motivation of undertaking this project is precisely making the texts available for research on that topic. However, both the rules of word formation and inflection are stable enough and predictable enough for this method to work, as is the vocabulary.

4. The Post-Correction Process

The OCR process has introduced a large number of errors to the text from *Fjölnir*. In this work, we will focus on correcting word errors, to the exclusion of zoning errors where the OCR software has failed to correctly recognize the layout of the text, resulting in text appearing out of order. The zoning errors were corrected manually in this project.

As the same kinds of character recognition errors tend to occur within a given document, a noisy channel model is a good fit, as it can efficiently model the probability of a particular error occurring.

4.1. Previous Work

Tong and Evans (1996) present a method for the correction of OCR errors using a noisy channel model approach combined with a bigram language model. They report an error reduction rate of 60.2% when the method is evaluated on digitized newspaper texts in modern English.

Volk et al. (2011) compare various strategies for reducing OCR errors in a multilingual corpus of digitized 19th century texts. These strategies include enlarging the modern lexicon of the OCR software with words from the targeted time period, applying predefined character substitution rules as well as applying a merging algorithm between the outputs of multiple OCR tools. They combine all correction methods in a single pipeline and find that only the merging algorithm has a significant positive contribution to the overall quality of the text, and that turning it off results in up to a 20% increase in uncorrected OCR errors.

Jurish (2010) generates candidates for the normalization of historical word forms using a variety of methods, including the application of hand-crafted transformational rules and phonetic conflation. The likeliest candidate is chosen using a HMM (Hidden Markov Model) based on a corpus of contemporary German. An F-Score of 99.4% is achieved when this method is applied on a corpus of historical German dating from 1780 to 1880.

Oravecz et al. (2010) normalize historical Hungarian word forms using a noisy channel model combined with a morphological analyzer and a decision tree. The error model is trained on a parallel corpus of manually normalized historical texts. The normalizer achieves a precision of 73.3% when evaluated on Old Hungarian texts.

Bollman et al. (2011) describe a rule-based approach to normalization, where transformational rules are automatically derived from a word-aligned parallel corpus of historical and modern texts. This method increases the ratio of tokens with correct modern spelling from 64.7% to 83.8% when applied on a version of the Bible in historical German. Limiting normalization candidates to word forms which appear in the Bible further improves the ratio to 91.0%.

[4] These experiments were carried out in trials of our normalizer. The texts are a part of the Parsed Historical Icelandic Corpus, IcePaHC: http://www.linguist.is/icelandic_treebank/ (Rögnvaldsson et al., 2012).

Pettersson et al. (2012) normalize a selection of historical Swedish texts using a small number of hand-crafted transformational rules, raising the average number of tokens with modern spelling from 65.2% to 73.0%. Applying contemporary NLP tools on the normalized text was found to yield improved results for a variety of tasks, including verb and complement extraction.

4.2. Methodology

Sufficient language resources for the creation of a lexicon with a reasonable coverage of 19th century Icelandic are available, but the lack of historical corpora precludes the use of statistical language models (beyond unigram models). The problem is compounded by the morphological richness of the language. Also, while Volk et al. (2011) achieved some success by improving the OCR process itself and by utilizing the output of multiple OCR tools, the work on *Fjölnir* is limited to the post-correction of the OCR text.

The OCR process may be likened to transmitting a text string through a noisy channel. The channel may introduce errors to the text by replacing certain characters with others which are similar in appearance. The errors which can occur in a digitized document depend on a number of different factors, such as the OCR software used, the font(s) used, the condition of the paper, and the quality of the scanned image. The probability of a given error, such as the likelihood that the letter *l* could be replaced with an *i*, can vary considerably from one digitized document to another, based on these (and other) factors. The word *ljós* 'light' might therefore consistently be replaced with the nonword *ijós* in one document, yet always be recognized correctly in another. However, even between different documents, it is always unlikely that characters with dissimilar shapes (such as *i* and *s*) be confused.

4.3. Noisy Channel Model

The noisy channel model approach to spelling correction combines an error model and a language model in order to estimate the probability that a misspelled (noisy) string *s* should in fact be the string *w*. The noisy channel model probability is estimated by multiplying the probabilities from the error model and the language model.

4.4. Error Model

The error model estimates the probability that a certain transformation can occur to a string which has been transmitted through the noisy channel, and is trained using pairs of strings prior to and after the transmission.

Kernighan et al. (1990) derive the probability of specific edit operations (the deletion, insertion and substitution of single characters and the transposition of two adjacent characters) from each string pair. The error model probability that *ijós* should be corrected to *ljós* would be calculated as

$$P(ijós|ljós) = \frac{sub(i,l)}{count(l)}$$

where $sub(i,l)$ is the number of times where the letter *i* was replaced with an *l*, and $count(l)$ is the number of times *l* appeared in the training corpus.

This method is improved upon by Brill and Moore (2000), whose model can deal with multiple distinct errors within the same string, while also modeling multiple character edit operations, such as *ph* → *f* in *physical* or *ante* → *anti* in *antechamber*. According to their error model, the probability that the nonword *sern* should be corrected to *sem* 'which' could be calculated as

$$P(sern|sem) = P(s|s) * P(e|e) * P(rn|m)$$

where $P(rn|m)$ is computed as

$$P(rn|m) = \frac{count(m \rightarrow rn)}{count(m)}$$

and $count(m \rightarrow rn)$ is the number of times the letter *m* was replaced with *rn* in the training corpus and $count(m)$ is the number of times *m* occurred in the correct strings. This approach will be followed in this work.

4.5. Language Model

The language model is an n-gram model that returns the probability of a given word. It can be constructed from a lexicon of correctly spelled word forms along with their frequencies (i.e., a unigram model), or derived from some large text corpus. For the purpose of OCR correction, a unigram language model derived from the Database of Modern Icelandic Inflection (DMII; Bjarnadóttir, 2012), which contains approximately 5.8 million Icelandic word forms, along with word frequencies from the 500 million word Web corpus Íslenskur orðasjóður (Hallsteinsdóttir, 2007) is used. Additionally, historical word forms and their word frequencies from the Written Language Archive (WLA), the main historical lexicographic archive at the Árni Magnússon Institute for Icelandic Studies, are used.[5]

4.6. Unsupervised Training of the Error Model

A drawback to the noisy channel model approach is the need for a training corpus for the error model. As mentioned before, OCR error probabilities can vary considerably between different documents, and therefore a single generalized error model will probably not be a good fit for all circumstances.

In this work, we propose a method for the unsupervised training of the error model. Initially, misspelled words are corrected using only the language model probability (i.e., the word frequency of the candidates). Any word form which is not known to the language model is considered to be a misspelling, and is corrected. The error model is then trained using these corrections. With the error model

[5] DMII: http://bin.arnastofnun.is/
Íslenskur orðasjóður: http://wortschatz.uni-leipzig.de/ws_isl/
WLA: http://www.arnastofnun.is/page/ritmalssafn

in place, the misspellings are corrected again using the full noisy channel model probability. The error model is then retrained using the improved corrections. This process is repeated several times. This is essentially an application of the expectation-maximization algorithm (Dempster et al., 1977).

4.7. Candidate Generation

Candidates are generated by the use of a Levenshtein automaton (Schulz & Mihov, 2002), which returns all words $W = \{w_1, w_2, \ldots, w_n\}$ in a lexicon that are within *n* edit operations of a given string s. In the first training iteration, the edit operations are limited to single character deletions, insertions and substitutions. In the following iterations, multiple character edit operations (e.g., $rn \rightarrow m$) are also allowed, and are derived from corrections made in the previous iteration. The partition for $P(s|w_i)$ is determined by the edit operations with which the candidate was generated, though it is quite possible for the same candidate to be generated in multiple different ways, in which case all partitions will be ranked.

First, the spellchecker will attempt to generate a list of candidates which are a single edit operation away from the misspelling. If no such words are found, it will attempt to find all words which are within two edit operations of the error. If no candidates can be generated in this manner, the spellchecker will not offer any suggestions to the user, though the word in question will still be underlined as an error.

4.8. Spelling Normalization

The use of predefined transformational rules has been successful when applied to the task of normalizing historical texts (Jurish, 2010; Bollman et al., 2011; Pettersson et al., 2012). However, as the spelling in *Fjölnir* is extremely varied, different rule sets might be needed for each issue. While such rule sets might yield good results when applied to the specific task of normalizing *Fjölnir*, their suitability for normalizing other historical texts, including ones from other time periods, is not as certain. A more general approach is therefore desirable.

As is the case with OCR post-correction, the task of spelling normalization can be viewed as a spellchecking problem. In this sense, historical variants of modern words are considered to be spelling errors that must be corrected to their modern forms. As with OCR texts, the probability of a given transformation can vary wildly between documents.

The noisy channel model is used to normalize *Fjölnir* using the same methods as for the OCR post-correction process, but replacing the historical language model (and lexicon) with a modern one. Here, the language model is derived from the Database of Modern Icelandic Inflection combined with word frequencies from Íslenskur orðasjóður, and the same training method as described above is used to adapt to the characteristics of individual documents.

5. Evaluation

The methods described in this work are evaluated by applying them to the 8th issue of *Fjölnir*, and comparing the results to the already corrected and normalized versions of the text which were manually reviewed for errors. The evaluation extends only to tokens containing at least one alphabetical character. The OCR text was reformatted prior to evaluation in order to eliminate all zoning errors (i.e., instances where the OCR software failed to output the text in the correct order). No other changes were made to the original text.

5.1. OCR Post-Correction

The 8th issue of *Fjölnir* contains a total of 18,714 alphabetical tokens, of which 2,591 were misrecognized during the OCR process (resulting in a word accuracy of 86.2%). The evaluation results can be seen in Table 2.

	Iter. 1	Iter. 2	Iter. 3	Iter. 4
N=1	38.1%	51.6%	52.9%	52.9%
N=5	49.4%	58.1%	57.8%	58.0%

Table 2: Correction suggestion accuracy for OCR errors.

The table above shows the portion of errors where the correct word is the top suggestion (N=1) or among the top five suggestions (N=5), through four iterations of the training algorithm. The correction accuracy of the spellchecker increases substantially after the first iteration, and remains more or less unchanged after the third. The correct word is among the top five suggestions 58% of the time. A review of the remaining errors shows that a considerable portion has been severely misrecognized by the OCR software, containing too many character errors for the correct (or even any) candidate to be generated (e.g., *töflunum* 'the tables' → *töfiiiiiuiu* and *varðveiti* 'preserve' → *oartwttt*). Further investigation reveals that this is very common for words in Fraktur (Gothic font), which appear with some frequency in this issue.[6] As the spellchecker can only handle two distinct edit operations within a single misspelled word before giving up (even though each operation can correct multiple character errors), it is unable to make a suggestion for the majority of these errors. Repeating the evaluation with words in Fraktur removed from the text yields the following results:

	Iter. 1	Iter. 2	Iter. 3	Iter. 4
N=1	47.9%	65.0%	66.4%	66.6%
N=5	62.0%	72.0%	72.1%	71.7%

Table 3: Correction suggestion accuracy for OCR errors, excluding words in Fraktur.

[6] Words or phrases in Fraktur are quite often interspersed with Roman fonts in *Fjölnir*. This can be seen in Figure 2, where the title of a book in a book review appears in Fraktur, as do direct quotes. The body of the text of *Fjölnir* is in Roman fonts.

As expected, when words in Fraktur are excluded from the evaluation (which raises the word accuracy to 90.0%), the accuracy of the suggestions improves considerably.

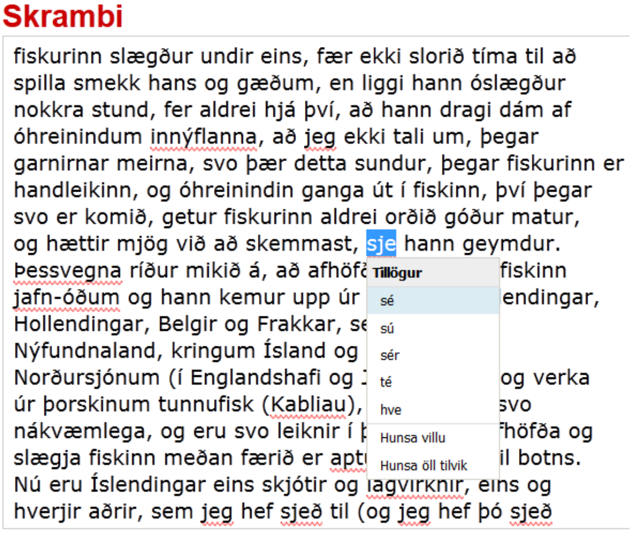
5.2. Spelling Normalization

Applying the noisy channel error model to the corrected version of the text to normalize the spelling to the modern form yields the following results:

	Iter. 1	Iter. 2	Iter. 3	Iter. 4
N=1	35.7%	68.9%	73.4%	73.6%
N=5	48.6%	84.6%	84.6%	84.6%

Table 4: Accuracy of suggestions for the normalization of historical words forms.

The correct modern form is among the top five suggestions in 84.6% of cases. The majority of the remaining errors are real-word errors (most notably where *en* ‘but’ has been written as *enn* ‘still’). These results show that the noisy channel model is well suited to normalizing historical Icelandic text. Figure 1 contains an example of the interactive normalization.



fiskurinn slægður undir eins, fær ekki slorið tíma til að spilla smekk hans og gæðum, en liggi hann óslægður nokkra stund, fer aldrei hjá því, að hann dragi dám af óhreinindum innýflanna, að jeg ekki tali um, þegar garnirnar meirna, svo þær detta sundur, þegar fiskurinn er handleikinn, og óhreinindin ganga út í fiskinn, því þegar svo er komið, getur fiskurinn aldrei orðið góður matur, og hættir mjög við að skemmast, sje hann geymdur. Þessvegna ríður mikið á, að afhöfð fiskinn jafn-óðum og hann kemur upp úr endingar, Hollendingar, Belgir og Frakkar, se Nýfundnaland, kringum Ísland og Norðursjónum (í Englandshafi og og verka úr þorskinum tunnufisk (Kabliau), svo nákvæmlega, og eru svo leiknir í fhöfða og slægja fiskinn meðan færið er apt il botns. Nú eru Íslendingar eins skjótir og lagvirknir, eins og hverjir aðrir, sem jeg hef sjeð til (og jeg hef þó sjeð

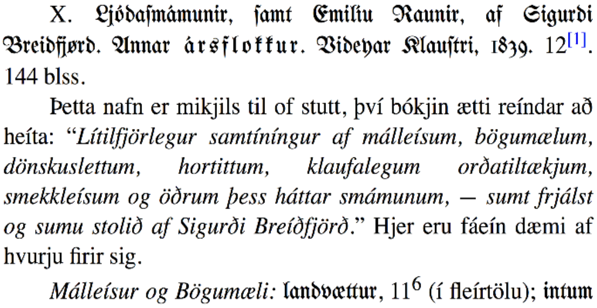
Fig. 1: The interactive spellchecker.

6. Web Production

The goal of the project was to publish *Fjölnir* on the Web, with free access for all users. Following the correction of the OCR text, the final form of the text (in its original spelling) was achieved through manual post-processing and formatting, preserving in the HTML and CSS markup the original layout in a standardized manner: italic, bold, and stretched text, superscript, font face (Fraktur vs. Roman) and size changes, block capitals, headers and subheaders, footnotes, tables, centred poem blocks with left-aligned text, etc. Graphics were drawn up in SVG and MathML was used for fractions and mathematical formulae. Figure 3 contains an example of the original spelling as presented on the website (*Fjölnir* 1843, p. 62); for comparison Figure 2 contains the same text from the original OCR file from Tímarit.is.

XI. £ji5bafmámunir, famt Gnnilíu Sfauntr, af ©ícutrbi 83reíbfj0rb. 2ínnar drðfloffur. S3iber,ar £Iaujtri, 1839. 121. 144 blss. Jietta nafn er niikjils til of stutt, því bókjin ætti reíndar að heífa: "látilffdrlegur smntiningur af málleísum, bögumœlum, dö-nskuslettum, hortittum, klaufalegum orða- tiltækj'um, smekkleisum og öðrum þess húttar smámunum, — sumt frjálst og sumu stolið af Siguroi Breíðfj'úrð." Hjcr eru fáei'n dæmi af hvurju firir sig. Málleísur og Bögumœli. lanbttcettur, ll6 (í fleírtölu); intum

Fig. 2: Example of OCR text, from Tímarit.is.



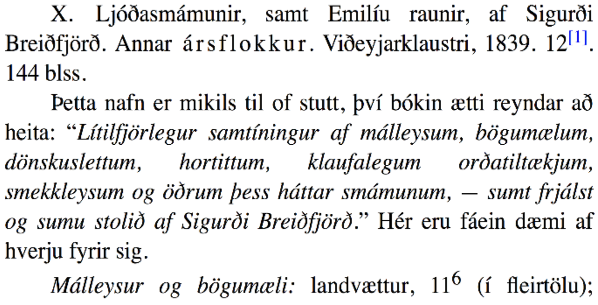
X. Ljóðaſmámunir, ſamt Emilíu Raunir, af Sigurði Breiðfjørð. Annar árſflokkur. Viðeyar Klauſtri, 1839. 12[1]. 144 blss.

Þetta nafn er mikjils til of stutt, því bókjin ætti reíndar að heíta: “*Lítilfjörlegur samtíníngur af málleísum, bögumælum, dönskuslettum, hortittum, klaufalegum orðatiltækjum, smekkleísum og öðrum þess háttar smámunum, – sumt frjálst og sumu stolið af Sigurði Breíðfjörð*.” Hjer eru fáeín dæmi af hvurju firir sig.

Málleísur og Bögumæli: landvættur, 11^{6} (í fleírtölu); intum

Fig. 3: Example from the website, original spelling.[7]

The next step is to create the modern spelling layer, which will inherit the formatting already present in the original spelling layer. The Web interface will allow the user to easily switch between the layers. Figure 4 shows the same text as in Figures 2 and 3, in the modern spelling.



X. Ljóðasmámunir, samt Emilíu raunir, af Sigurði Breiðfjörð. Annar ársflokkur. Viðeyjarklaustri, 1839. 12[1]. 144 blss.

Þetta nafn er mikils til of stutt, því bókin ætti reyndar að heita: “*Lítilfjörlegur samtíningur af málleysum, bögumælum, dönskuslettum, hortittum, klaufalegum orðatiltækjum, smekkleysum og öðrum þess háttar smámunum, – sumt frjálst og sumu stolið af Sigurði Breiðfjörð*.” Hér eru fáein dæmi af hverju fyrir sig.

Málleysur og bögumæli: landvættur, 11^{6} (í fleirtölu);

Fig. 4: Example from the website, modern spelling.

[7] Translation: Poetic Trivia, with the Lament of Emilia, by Sigurður Breiðfjörð. Second annual part. Viðey Monastery, 1839. 12[1]. 144 pps. This title is much too short, because the book should be called “A trivial hotchpotch of blunders, solecisms, Danishisms, waffle, clumsy phrasing, bad taste, and other trivialities, some freely available and some stolen by Sigurður Breiðfjörð.” Here are some examples of each of those. Blunders and solecisms. *landvættur* 11-6 (in the plural); *intum*

Further along, a unified file format incorporating any number of named text layers is envisioned, from which HTML-files in each version may be generated. An example of the information contained in such a unified file may be seen in Table 5.

OCR	*Post-corr.*	*Modern*	*Lemma*	*Tag*
Hjcr	Hjer	Hér	hér	aa
eru	eru	eru	vera	sfg3fn
fáei´n	fáeín	fáein	fáeinir	fohfn
dæmi	dæmi	dæmi	dæmi	sþghfn
af	af	af	af	aþ
hvurju	hvurju	hverju	hver	fsheþ
firir	firir	fyrir	fyrir	ao
sig	sig	sig	sig	fphfo

Table 5: Example of the 3 layers of *Fjölnir* 1838, with lemmas and tags.[8]

The *Fjölnir* website will be accessible from the website of the Árni Magnússon Institute for Icelandic Studies, http://arnastofnun.is/.

7. Conclusion

The two versions of the texts of the *Fjölnir* project are due to be made accessible online in the spring of 2014. A unified file format incorporating the text layers, with annotation, are a part of larger prospective project at the Árni Magnússon Institute for Icelandic Studies.

The noisy channel model proved to be successful, even for a very error-prone OCR text, but using a more complex language model, such as a bi- or tri-gram model would likely improve the correction accuracy for both OCR post-correction and normalization. For better results, context-sensitive error correction is needed for real-word errors. Additional updates to the spellchecker are planned, such as dynamically updating the error model probabilities as the user makes corrections.

While the tool described in this work is interactive, it could easily be converted into a fully automated spellchecker for the correction (as well normalization) of large-scale digitization efforts.

8. Acknowledgements

The project described in this paper has been run in tandem with the creation of a corpus of early Modern Icelandic (Svavarsdóttir et al., 2014), with close cooperation of participants in both projects. We would like to thank Ásta Svavarsdóttir, Sigrún Helgadóttir, and all the participants in the corpus building. Thanks are also due to Örn Hrafnkelsson and Kristinn Sigurðsson at the National and University Library of Iceland for access to the OCR texts of *Fjölnir*. The Icelandic Student Innovation Fund financed the work of Jón Friðrik Daðason and Kristján Rúnarsson on the *Fjölnir* project in the summer of 2012, and the creation of the software was partly financed by a grant from the same fund to Jón Friðrik Daðason in 2011.

9. References

Bjarnadóttir, K. (2012). The Database of Modern Icelandic Inflection. In *Proceedings of the workshop Language Technology for Normalization of Less-Resourced Languages, SaLTMiL 8 - AfLaT, LREC 2012*, pp. 13-18, Istanbul, Turkey.

Bollmann, M., Petran, F., & Dipper, S. (2011). Rule-based normalization of historical texts. In *Proceedings of the International Workshop on Language Technologies for Digital Humanities and Cultural Heritage* (pp. 34-42).

Brill, E., & Moore, R. C., (2000). An Improved Error Model for Noisy Channel Spelling Correction. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics,* Hong Kong.

Daðason, J. F. (2012). *Post-Correction of Icelandic OCR Text.* MS thesis at the University of Iceland, http://hdl.handle.net/1946/12085.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal statistical Society*, 39(1), 1-38.

Hallsteinsdóttir, E., Eckart, T., Biemann, C., Quasthoff, U., & Richter, M. (2007). Íslenskur Orðasjóður - Building a Large Icelandic Corpus. In *Proceedings of NODALIDA-07*, Tartu, Estonia

Hrafnkelsson, Ö., & Sævarsson, J. (2014). Digital libraries of historical Icelandic newspapers, periodicals, magazines and old printed books. In *Proceedings of the workshop Language resources and technologies for processing and linking historical documents and archives-Deploying Linked Open Data in Cultural Heritage, LRT4HDA, LREC 2014,* Reykjavík.

Jurish, B. (2010). More than Words: Using Token Context to Improve Canonicalization of Historical German. *JLCL*, 25(1), 23-39.

Kernighan, M. D., Church, K. W., & Gale, W. A. (1990). A spelling correction program based on a noisy channel model. In *Proceedings of the 13th conference on Computational linguistics-Volume 2* (pp. 205-210). Association for Computational Linguistics.

Loftsson, H. (2008). Tagging Icelandic Text: A linguistic rule-based approach. *Nordic Journal of Linguistics* 31(1):47-72.

Oravecz, C., Sass, B., & Simon, E. (2010). Semi-automatic normalization of Old Hungarian codices. In *Proceedings of the ECAI Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)* (pp. 55-59).

Rögnvaldsson, E., Ingason, A. K., Sigurðsson, E. F., & Wallenberg, J. (2012). The Icelandic Parsed Historical Corpus (IcePaHC). In *Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012*, Istanbul, Turkey.

[8] The text was tagged using IceNLP (Loftsson, 2008), http://nlp.cs.ru.is/icenlp/?lang=en.

Pettersson, E., Megyesi, B., & Nivre, J. (2012). Rule-Based Normalisation of Historical Text–a Diachronic Study. In *Empirical Methods in Natural Language Processing: Proceedings of the Conference on Natural Language Processing* (pp. 333-341).

Schulz, K. U., & Mihov, S. (2002). Fast string correction with Levenshtein automata. *International Journal on Document Analysis and Recognition*, *5*(1), 67-85.

Svavarsdóttir, Á., Helgadóttir, S., & Kvaran, G. (2014). Language resources for early Modern Icelandic. In *Proceedings of the workshop Language resources and technologies for processing and linking historical documents and archives-Deploying Linked Open Data in Cultural Heritage, LRT4HDA, LREC 2014*, Reykjavík.

Tong, X., & Evans, D. A. (1996). A statistical approach to automatic OCR error correction in context. In *Proceedings of the fourth workshop on very large corpora* (pp. 88-100).

Volk, M., Furrer, L., & Sennrich, R. (2011). Strategies for reducing and correcting OCR errors. In *Language Technology for Cultural Heritage* (pp. 3-22). Springer Berlin Heidelberg.

IceTagging the "Golden Codex". Using language tools developed for Modern Icelandic on a corpus of Old Norse manuscripts

Ludger Zeevaert
Stofnun Árna Magnússonar í íslenskum fræðum
Árnagarði við Suðurgötu, 101 Reykjavík, Iceland
E-mail: ludger@zeevaert.de

Abstract

Compared to corpora of texts from modern languages that can be based on texts already existing in electronic form or on electronic versions produced mechanically with OCR, the compilation of a corpus of older Icelandic manuscript texts is a rather laborious and time-consuming task because the texts have to be transcribed manually from the manuscripts. Whereas the current state of the art does not allow for a substantial reduction of the workload needed for the transcriptions, the preparation of the corpus for practical tasks offers promising possibilities for automatic or semiautomatic procedures. The following article describes an attempt to use language tools developed for a corpus of Modern Icelandic texts on an Old Icelandic corpus built on manuscript transcriptions in order to enrich the corpus with information useful for linguistic research.

POS-tagging of versions of the manuscript texts transferred to Modern Icelandic spelling with taggers developed for Modern Icelandic delivered absolutely satisfactory results, but only after adapting the procedure of manuscript transcription especially to the needs of the utilised language tools. With a reverse approach, i.e. an adaptation of the tools to the demands of medieval Icelandic manuscript corpora, it might be possible to extend their usability in the construction of a multi-purpose corpus and to profit for their further development from work done in more traditional approaches to textual science.

Keywords: Old Icelandic, POS-tagging, Njáls saga

1. Introduction

The application of language tools to historical corpora and especially to less resourced cultural heritage languages such as Old Norse is one of the main topics of interest for the LRT4HDA-workshop in Reykjavík in 2014. The following article describes an attempt to design a corpus of 17th century Icelandic manuscripts that seeks to reconcile the demands on a corpus mainly designed for philological use with the applicability of language tools for linguistic analyses. The article gives a short overview of the background of the philological and linguistic research on the Early Modern transmission of the Old Norse *Njáls saga* in the frame of the project *Variance of* Njáls saga, describes the architecture of the corpus that was originally designed for a philological approach (edition, stemmatological questions etc.) to 14th century manuscripts, proposes a method to circumvent the problems that older Icelandic texts in their original spelling pose for automatic tagging and addresses issues in connection with the compatibility of language tools and linguistic research originating in a philological context.

2. The *Gullskinna*-corpus

Gullskinna (app. 'The Golden Codex') is the name of a medieval parchment manuscript. The book itself is lost, but its existence can be derived from variants in the margin of a 17th-century copy of *Njáls saga* (AM 134 fol., cf. fn. 2) that according to the scribe stem from a source named *Gullskinna*. In the frame of the project *Variance of* Njáls saga[1], Margrét Eggertsdóttir (Stofnun Árna Magnússonar í íslenskum fræðum, Reykjavík) together with Alaric Hall (University of Leeds) and Ludger Zeevaert (Stofnun Árna Magnússonar í íslenskum fræðum, Reykjavík) is currently investigating the textual relationships of the 64 manuscripts containing *Njáls saga*, an Old Norse prose narrative composed around 1280, with special emphasis on 17th-century paper manuscripts. In a study based on a limited sample of the text (chapter 86 of the saga), 17 manuscripts (in addition to the 6 already known)[2] could be identified as deriving from the lost *Gullskinna*-manuscript.

This research is based on normalised transcriptions of the manuscripts using a transcription method originally developed to examine linguistic variance in different manuscripts of this Icelandic family saga from a linguistic, philological and literary perspective.

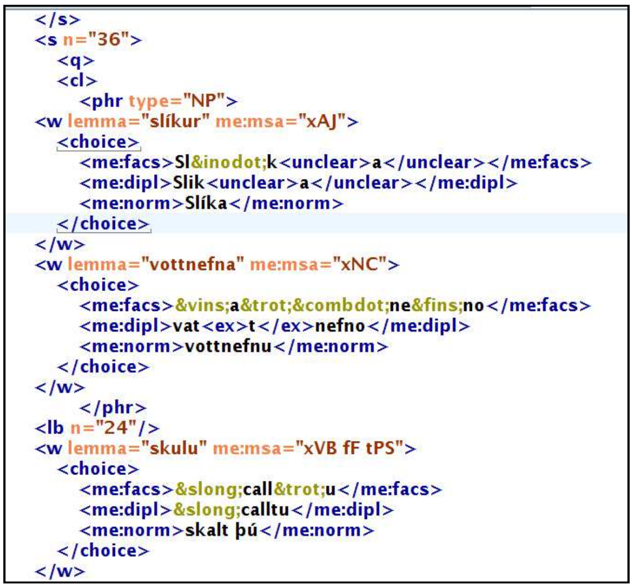
The majority of the manuscripts of *Njáls saga* are fragmentary. The distribution of the text preserved in the different manuscript fragments makes it necessary to include 7 chapters from different parts of the saga to be able to clarify the relationships between the different manuscripts and present them in a stemma[3] including all textual witnesses (i.e. all surviving manuscripts). Given the rather time consuming nature of this work it seems reasonable to prepare the transcriptions in a way that

[1] http://www.arnastofnun.is/page/breytileiki_njalu (visited 20.1.2014).

[2] On the basis of the variants found in the manuscript AM 134 fol. ("Hofsbók", Árnsasafn, Reykjavík), Jón Þorkelsson (1889: 719ff.) identified three 17th-century copies of *Njáls saga* belonging to the *Gullskinna*-class, and Már Jónsson (1996) was able to add three more copies deriving from *Gullskinna*, likewise from the 17th or early 18th century.

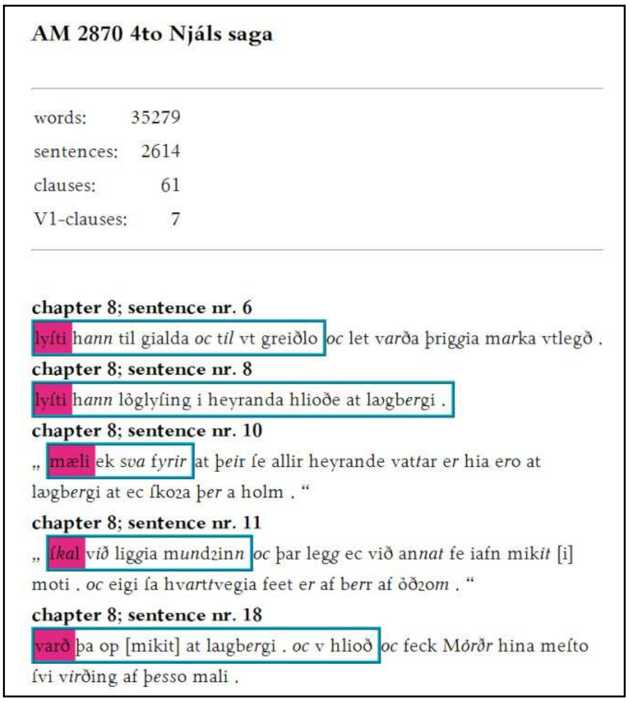
[3] The relation between manuscripts is usually presented in the form of a stemma, a branching diagram comparable to a phylogenetic tree. While the earlier text transmission, i.e. the parchment codices, has been investigated quite exhaustively, this does not at all account for the younger paper manuscripts. One of the main goals of the *Njáls-saga* project is the construction of a complete stemma including all existing manuscripts.

makes the corpus useable for different scientific tasks. One part of the project *Variance of* Njáls Saga is concerned with research on linguistic variation in the earliest manuscripts of the text from the 14th century. It thus appeared desirable to prepare the *Gullskinna*-corpus in a way that among other things would allow for a comparison of the findings from the 14th-century texts with the manuscripts from the 17th century.



```
</s>
<s n="36">
  <q>
  <cl>
    <phr type="NP">
<w lemma="slíkur" me:msa="xAJ">
  <choice>
    <me:facs>Sl&inodot;k<unclear>a</unclear></me:facs>
    <me:dipl>Slik<unclear>a</unclear></me:dipl>
    <me:norm>Slíka</me:norm>
  </choice>
</w>
<w lemma="vottnefna" me:msa="xNC">
  <choice>
    <me:facs>&vins;a&trot;&combdot;ne&fins;no</me:facs>
    <me:dipl>vat<ex>t</ex>nefno</me:dipl>
    <me:norm>vottnefnu</me:norm>
  </choice>
</w>
    </phr>
<lb n="24"/>
<w lemma="skulu" me:msa="xVB fF tPS">
  <choice>
    <me:facs>&slong;call&trot;u</me:facs>
    <me:dipl>&slong;calltu</me:dipl>
    <me:norm>skalt þú</me:norm>
  </choice>
</w>
```

Figure 1: Transcription on three levels with segmentation, syntactic tagging, tagging of direct speech and POS-tagging



AM 2870 4to Njáls saga

words:	35279
sentences:	2614
clauses:	61
V1-clauses:	7

chapter 8; sentence nr. 6

lyſti *hann* til gialda *oc til* vt greiðlo *oc* let *var*ða þriggia m*ar*ka vtlegð .

chapter 8; sentence nr. 8

lyſti *hann* lôglyſing i heyranda hlioðe at lꜹgb*er*gi .

chapter 8; sentence nr. 10

„ mæli ek *sva* f*yrir* at þ*eir* ſe allir heyrande vat*t*ar *er* hia ero at lꜹgb*er*gi at ec ſkoꝛa þ*er* a holm . “

chapter 8; sentence nr. 11

„ ſkal við liggia *mund*ꝛinn *oc* þar legg ec við an*nat* fe iafn miki*t* [i] moti . *oc* eigi ſa hv*art*t*vegia feet *er* af b*er*r af ôðꝛom . “

chapter 8; sentence nr. 18

varð þa op [mikit] at laugb*er*gi . *oc* v hlioð *oc* feck Mǫrðr hina meſto ſvi v*ir*ðing af þesso mali .

Figure 2: Examples for narrative inversion in *GKS 2870 4to* ("*Gráskinna*", Árnasafn, Reykjavík), html-display with XSLT-style sheet

For the 14th-century corpus, the texts were transcribed on a diplomatic level[4] from photographs of the manuscripts or the manuscripts themselves. XML-tags for the segmentation/layout of the texts (words, columns, pages, chapter headings etc.) were inserted, and a normalised transcription, a syntactic segmentation (sentences, clauses) and a partial part-of-speech tagging were added by hand (for a more detailed description cf. Zeevaert in prep.).

Both segmentation and POS-tagging were successfully used to identify linguistic variation between the 14th-century manuscripts of *Njáls saga* that consists amongst others of differences in word order, the use of different constructions for syntactic subordination and differences in the use of tenses (cf. Zeevaert 2013).

3. A modified approach to the corpus architecture

From a traditional philological attitude which aims at making the contents of manuscripts accessible to the research community and wants to investigate the relationship of different manuscripts of the same text the approach described in the previous paragraphs appears to be logical. In a first step transcriptions of the manuscripts have to be made to facilitate the access to the different versions of the text in the single manuscripts. In a second step those transcriptions can be modified for further research that largely consists of a comparison of the different manuscripts. By providing additional information, the usability of the transcriptions can be extended. Syntactical and morphological mark-up would e.g. allow for research on linguistic phenomena like historical developments of the inflexional system or word order.

However, it seemed not to be a realistic option to manually insert this additional mark-up during the funding period of 18 months. For obvious reasons the development of language tools for historical varieties of Icelandic is commercially not very promising, and financial resources for technical solutions especially designed for the requirements of our project were not at hand.

We thus tried to develop simple procedures with maximal efficiency that would allow for an economic use of the limited resources of our project by adapting existing methods and tools and at the same time keep the transcriptions open for the wide range of different approaches and research questions of manuscript-related research. This includes on the one hand the observance of common philological standards of accuracy of transcription (which are e.g. necessary to produce editions, for research on the abbreviation system of single manuscripts, to detect scribal errors that give hints about the relationship between exemplars and copies etc.) and on the other hand the use of language tools that were developed for Modern Icelandic but have proven to deliver good results for older texts. The use of those tools requires a normalisation of orthography and an adjustment of (inflexional) morphology to the Modern Icelandic standard.

[4] In the literature, the term 'diplomatic edition/transcription' is used in a rather broad sense for representations of historical texts (manuscripts or early printings) that follow the text witness closely, in contrast to normalised editions/transcriptions that unitise spelling and grammar on the basis of standard grammars and dictionaries (cf. e.g. Haugen 2002: 540). A more sophisticated system of different degrees of accurateness in rendering the spelling of manuscript witnesses is described by Gunnlaugsson (2003).

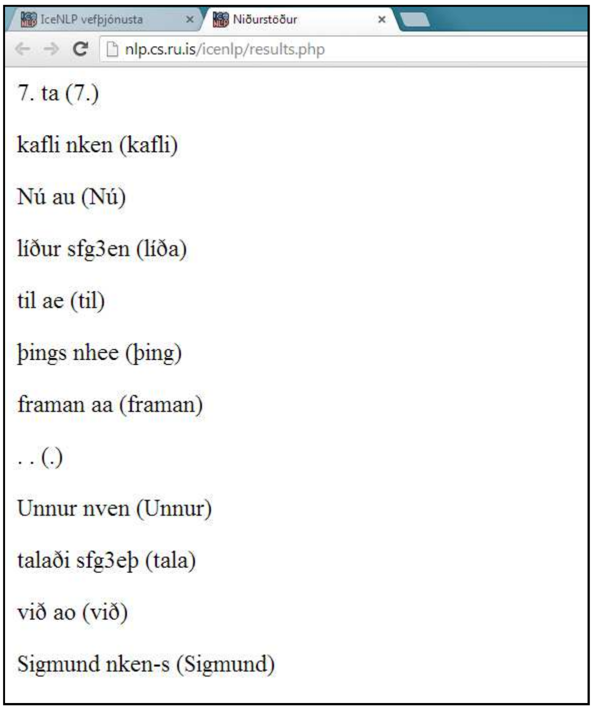
Our basic idea was to develop a step-by-step routine to produce a representation of the text of a single manuscript that would be able to satisfy as many potential users interested in Old Icelandic texts as possible. The starting point of this modified procedure was a text file containing the complete transcription of one of the most important text-witnesses of *Njáls saga*, *Reykjabók* (*AM 468 4to*, Copenhagen, Den Arnamagnæanske Samling). The file contains Sveinn Yngvi Egilsson's (Ed. 2003) edition of the text which for two reasons is especially suitable for our work: The text is reliable because it is (apart from the spelling) an accurate transcription of the manuscript text and the text is highly usable for automatic handling because the spelling is normalised to modern Icelandic orthography. As is customary in modernised editions of Old Icelandic texts, certain traits of Old Icelandic spelling and inflection are kept in Egilsson's edition. This applies e.g. to the form of the personal pronoun *eg* 'I', an archaising compromise between the Modern Icelandic form *ég* and the Old Icelandic form *ek*. Originally the decision was made in the project to keep the archaic traits of the edition on the normalised level of the transcriptions. However, linguistic analyses that include morphology have to be based on the diplomatic level anyway. A complete conversion of the normalised level to the Modern Icelandic standard would thus not have any influence on linguistic research but exhibit advantages for searches in the text.[5]

Egilsson's text was used to (by and large automatically) produce a TEI-XML-document[6] containing basic mark up (tags for words and punctuation marks, sequentially numbered sentences) that could serve as a template for transcriptions of the different manuscripts. In the next step the text of the XML-document was rearranged and corrected according to the manuscript, and for a philologically satisfactory representation of the manuscript texts it was also necessary to add more accurate levels of transcription (diplomatic/strictly diplomatic) and information about the segmentation of the manuscript (pages, columns, chapter headings etc.) to the XML-file. Finally syntactical and grammatical information that was necessary to compare the use of certain stylistic features in different manuscripts was tagged manually: clause boundaries and finite verbs to determine the use of narrative inversion, direct speech and tense of verbs to find instances of historical present tense etc.

For the *Gullskinna*-corpus we decided to go one step further in the rearrangement of the workflow and to try to provide an XML-file as a template for the transcription that already contained the necessary grammatical mark-up instead of adding it manually to every single transcription afterwards. Adding grammatical annotation manually is a rather time consuming and error-prone task. Together with our colleagues from the Department of Lexicography we therefore decided to explore the possibilities for using existing tools and solutions for this task. As a first step, a text-file of chapter 7 of Sveinn Yngvi Egilsson's edition of the saga (cf. above) was analysed using the web interface of *IceNLP* (http://nlp.cs.ru.is/icenlp/, visited 16.01.2014). *IceNLP* is a processing toolkit for Modern Icelandic (cf. Loftsson & Rögnvaldsson, 2007b), and Sveinn Yngvi's edition contains the text of a 14th century manuscript transformed to modern Icelandic spelling. The web interface was chosen because, in contrast to other options, it did not require certain ways of formatting the text and was easily accessible.

4. POS-tagging

The tagging output produced by *the IceNLP* web interface is a list where each word token of the text is placed in one line together with a morphosyntactic annotation (a description of the tag set is given e.g. by Helgadóttir, 2007: 106f.) and a lemma. The original order of words in the text is kept intact.



```
7. ta (7.)

kafli nken (kafli)

Nú au (Nú)

liður sfg3en (líða)

til ae (til)

þings nhee (þing)

framan aa (framan)

. . (.)

Unnur nven (Unnur)

talaði sfg3eþ (tala)

við ao (við)

Sigmund nken-s (Sigmund)
```

Figure 3: *IceTagger*-output

A manual verification showed that both correct lemma and word class were recognised for 97.9% of the 2304 word tokens in the sample. If only the prose part of the text is considered,[7] the accuracy goes up to 98.7%.[8] A

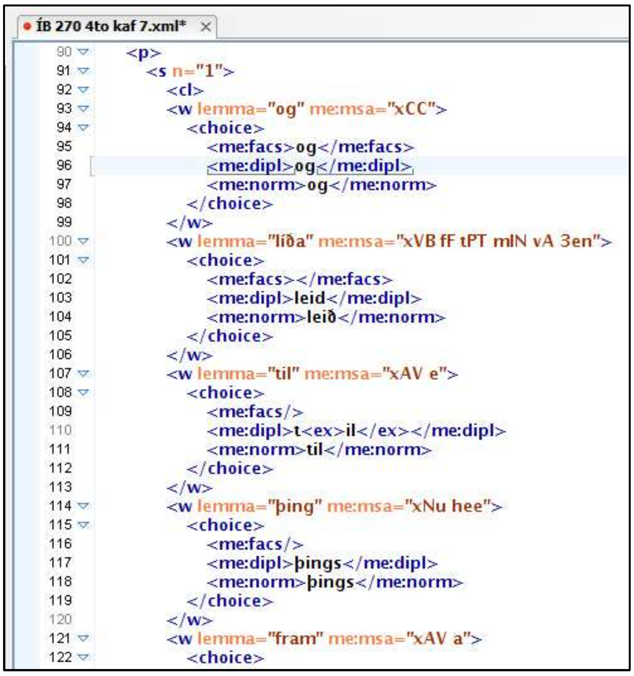
[5] For a discussion of standards of transcription of Old Norse texts cf. Zeevaert (2013) with references to further reading.

[6] TEI is a common standard for the encoding of texts in the digital humanities (cf. http://www.tei-c.org/index.xml, visited 22.01.2014). The encoding used in the *Njáls-saga* project is mainly based on the Menota-guidelines (a modified version of TEI-XML adapted to medieval Scandinavian texts, http://www.menota.org/, visited 22.01.2014). The Menota-tag-set was supplemented with tags for syntactic segmentation defined in the TEI-standard.

[7] This seems to be appropriate in this case because the skaldic stanzas contained in the saga are not part of the linguistic analyses in the project.

[8] In comparison, *IceTagger* assigned the correct word class and the correct lemma to 38% of the tokens in a non normalised-diplomatic transcription of the corresponding text (chapter 7) from a 14th-century manuscript (*AM 162 B fol. beta*, Árnasafn, Reykjavík). Results from earlier attempts to tag normalised Old Icelandic texts with language tools designed for Modern Icelandic were 88% (Rögnvaldsson & Helgadóttir 2011: 68) and between

precise determination of the accuracy of the complete tag-set was not performed as it was considered too time consuming given the fact that the major part of the annotated grammatical features are not relevant for the analyses in our project. The taggers used in the *IceNLP* package are trained and developed on modern Icelandic.



```
<p>
  <s n="1">
    <cl>
    <w lemma="og" me:msa="xCC">
      <choice>
        <me:facs>og</me:facs>
        <me:dipl>og</me:dipl>
        <me:norm>og</me:norm>
      </choice>
    </w>
    <w lemma="líða" me:msa="xVB fF tPT mIN vA 3en">
      <choice>
        <me:facs></me:facs>
        <me:dipl>leid</me:dipl>
        <me:norm>leið</me:norm>
      </choice>
    </w>
    <w lemma="til" me:msa="xAV e">
      <choice>
        <me:facs/>
        <me:dipl>t<ex>il</ex></me:dipl>
        <me:norm>til</me:norm>
      </choice>
    </w>
    <w lemma="þing" me:msa="xNu hee">
      <choice>
        <me:facs/>
        <me:dipl>þings</me:dipl>
        <me:norm>þings</me:norm>
      </choice>
    </w>
    <w lemma="fram" me:msa="xAV a">
      <choice>
```

Figure 4: Transformation of the *IceTagger*-output to Menota-TEI-XML

One of the main differences between Old and Modern Icelandic lies in the vocabulary which means that a number of words from the text were not represented in the Modern Icelandic lexicon used by the tagger. However, this did not seriously affect the recognition of word classes. Less than 5% of the word tokens were unknown to the tagger (mostly compound words and names, but also single nouns connected to legal practice), and 85% of them were tagged correctly, although in ca. 19% of those cases a correct lemma was not assigned. The results mirror the fact that on the one hand syncretism in the inflexional system is less, but irregularity in the formation of word stems more prominent in Icelandic than in comparable inflexional languages like German. In the skaldic stanzas that use a special archaic poetic vocabulary (cf. fn. 7) the amount of unknown words was substantially higher. [9]

By using the data-driven tagger *TnT* trained on a combined corpus of Modern and Old Icelandic texts the recognition rate could be improved substantially: Sigrún Helgadóttir (Stofnun Árna Magnússonar í íslenskum fræðum, Reykjavík) was able to reach an accuracy of 99.8% for the POS-tagging (i.e. word-class identification, a lemmatisation was not made) using the same text sample (chapter 7 of Sveinn Yngvi Egilsson's *Njáls saga* edition). The approach used by her is described in detail by Rögnvaldsson & Helgadóttir (2011): The *TnT*-tagger (cf. Brants, 2000) was trained on the corpus of the Icelandic Frequency Dictionary (Pind et al., 1991) and *Saga-Gold*, a manually corrected corpus with 95,000 words from Old Iceland prose texts. With this method Rögnvaldsson & Helgadóttir reached an accuracy of 92.7% for all features in the morphosyntactic tagging for a corpus of ca. 1.65 million words.[10]

Despite the slightly poorer results we decided to continue to work with the web application for two reasons: The technical requirements for the improved procedure were not at hand in our project, and the output did not contain the lemmata of the analysed words. For the work with non-normalised texts (as already mentioned, a diplomatic transcription of the manuscript was to be added at a later stage) a lemmatised text is of great value as it allows to retrieve all word forms of a lemma regardless of their spelling or grammatical form.[11]

The form of the list generated by the *IceTagger* web application allows for a rather uncomplicated transformation to the TEI-XML-format defined for the Menota-corpus (c.f. fn 6). A transformation to this format is essential for the project because the analysis and the comparison of the different text witnesses are performed with style sheets that build on XPath expressions and a system of numbered semantic entities (sentences) that allows a retrieval of corresponding chunks of text in all manuscripts. A prerequisite for this approach is that borders of clauses and constituents/phrases can be identified and words are tagged for certain grammatical features.

5. Parsing

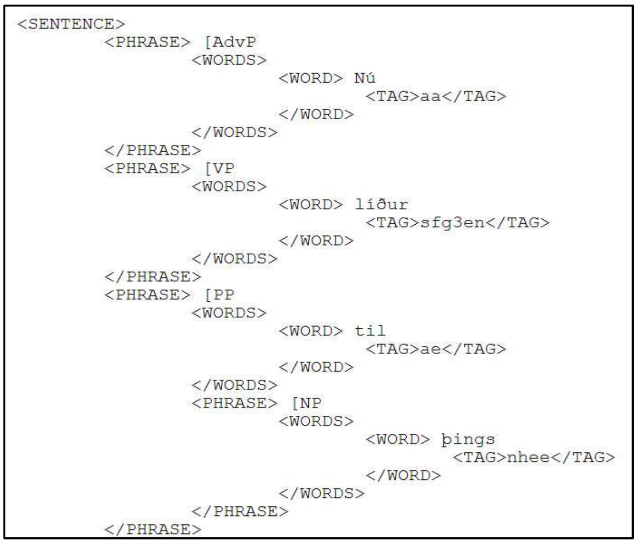
Part of the *IceNLP* web application is the parser *IceParser* that marks constituents by adding brackets and labels to an input text. With our sample text it produced results that were satisfactory with respect to accuracy. However, the application of the output from the parser in the frame of our approach met some, mainly technical, difficulties. The output of *IceParser* which was primarily developed for tasks in connection with the processing of Modern Icelandic texts, e.g. grammar checking, (Loftsson & Rögnvaldsson 2007: 131) follows the usual practice in treebank annotation, i.e. a display of constituents in a hierarchical structure. The output generated by *IceParser* consists of a linear sequence of constituents in which the hierarchic structure is represented with a nesting of phrase tags.

83.55% and 89.29% (Loftsson 2013).

[9] Frequent scribal errors in the medieval copies of the skaldic stanzas show that this vocabulary was difficult to comprehend already in the 14th century.

[10] Loftsson (2013) corrected the *Saga-Gold* corpus further and trained three taggers on the corrected *Saga-Gold*. By combining those taggers he was able to reach an accuracy of 92.32 for all features in the morphosyntactic tagging and 97.55% for word class. The fact that the combined tagger does not perform better on the same corpus that was used for the training seems to rely mostly on lexical ambiguities present in Old Icelandic but not in Modern Icelandic. One of the frequent error sources described by Loftsson, *er* 'is' (1/3SG.PRS.IND of *vera* 'be') vs. *er* 'which' etc. (relative particle), was one of the main error sources in our experiment, too.

[11] A lemma-based word list would e.g. be useful to detect variation in the spelling of certain words that could be connected to a change of the exemplar or the scribe or to determine the age of a manuscript.



```
<SENTENCE>
        <PHRASE> [AdvP
                <WORDS>
                        <WORD> Nú
                                <TAG>aa</TAG>
                        </WORD>
                </WORDS>
        </PHRASE>
        <PHRASE> [VP
                <WORDS>
                        <WORD> líður
                                <TAG>sfg3en</TAG>
                        </WORD>
                </WORDS>
        </PHRASE>
        <PHRASE> [PP
                <WORDS>
                        <WORD> til
                                <TAG>ae</TAG>
                        </WORD>
                </WORDS>
                <PHRASE> [NP
                        <WORDS>
                                <WORD> þings
                                        <TAG>nhee</TAG>
                                </WORD>
                        </WORDS>
                </PHRASE>
        </PHRASE>
```

Figure 5: *IceParser*-output

Syntactic tagging in the *Njáls saga* project is on the one hand used to detect different word order patterns between different manuscripts of the saga. This is done on the basis of an annotation of immediate constituents that by and large correspond to the phrase tagging generated by *IceParser.* For an analysis of word order patterns for which variation can be shown in our manuscripts, e.g. the order of head and modifier in the NP, the output from *IceParser* would be a good starting point. More important, however, is a segmentation into syntactic units, i.e. clauses, which are used as a point of reference for word order to distinguish e.g. between verb-second as the unmarked word order in declarative sentences and the stylistically marked verb-first word order, which, usually under the designation *narrative inversion* plays an important role as a typical feature of the saga-style. Clause segmentation is also necessary for the differentiation of different indirect speech constructions (accusative with infinitive vs. conjunctional clause). For both verb-position and indirect speech constructions a certain variation can be found in our corpus.

Calculating costs and benefits of full parsing vs. shallow parsing, Loftsson & Rögnvaldsson (2007: 128) come to the conclusion that a complete analysis of sentences is not necessary for the main tasks of a parser for modern Icelandic. Rögnvaldsson & Helgadóttir (2011: 70ff.) were able to show that POS-tagging without clause annotation can be used successfully to identify certain syntactical constructions in a corpus of Old Icelandic texts. A segmentation into clauses was therefore not implemented. However, this method is highly dependent on constructions that involve closed word classes that can be lexically identified (negations, prepositions etc.) which limits completeness and precision of the results (cf. Rögnvaldsson & Helgadóttir 2011: 72).

This is a much more serious problem for historical corpora that have to be based on the surviving texts and therefore contain a much smaller and less representative corpus of the language than modern language corpora. Research on stylistic phenomena very often involves syntactic constructions (e.g. infinitive constructions replacing subordinate clauses in indirect speech) that are not very frequent in the texts and others (e.g. narrative inversion) that are not identifiable with a search for certain lexical items alone. A syntactical annotation that covers both clauses and phrases in addition to the POS-tagging is thus of clear advantage because more accurate and exhaustive results can be reached. A determination of the order of constituents which plays an important role for research on stylistic differences between manuscripts can only be done with relation to their position in the clause (with the XPath expression "//s[.//cl[*[1][contains (@me:msa,'fF')]]]" examples for narrative inversion can be found in a POS-tagged XML-document, but only if also the beginnings of clauses are tagged, e.g. with the TEI-tag <cl>).

IceTagger is able to generate XML-files (c.f. Figure 5) and it is generally possible to manually add a clause-segmentation to this XML-output. What in our case speaks against the usage of the *IceParser* output is that it does not contain the lemmatisation produced by *IceTagger*. An easy access to different forms of a word via a common lemma is of considerable advantage for certain research questions that involve the comparison of different spellings of one word in one manuscript or between manuscripts (e.g. for the identification of scribes, the dating of manuscripts, the identification of exemplars and copies etc.), and to us a manual phrase tagging in addition to the anyway necessary manual clause tagging seemed to be less laborious than the manual lemmatising of the texts. We therefore opted for the method that delivered an output that appeared to be most suitable for the tasks and the technical capabilities of our project, the web interface of the hybrid tagger *HMM+Ice+HMM*.

However, from our point of view a further development of *IceParser* in the direction of a more TEI-compatible output that includes both lemmata and a phrase-segmentation would be desirable.

It was also necessary to convert the tag set used by *IceTagger* to the Menota-tag set for morphosyntactic annotation which is less ambiguous and due to its structure more suited for searches with XPath expressions than the tag set produced by the *IceTagger* that was originally developed for the Icelandic frequency dictionary (Pind et al., 1991). The XML-file could then be modified according to the text of a single manuscript in an XML-editor.

6. Discussion

The application of *IceTagger,* a language tool designed for Modern Icelandic, on Icelandic texts written between the 14th and 18th century went remarkably well in the case of the *Njáls saga* project. It has to be pointed out, though, that in our case the preconditions were remarkably favourable:

- A normalised electronic transcription based on one manuscript of *Njáls saga* was at our disposal.
- The differences between the manuscripts of the saga are, in comparison to other Icelandic prose texts, let alone historical texts from other languages, rather small.
- Linguistic change from Old to Modern Icelandic in the domain of morphology, syntax and lexicon has to be considered as rather insignificant in comparison to most other languages, and the comparably more prominent phonological change is by and large not rendered in the Modern Icelandic orthography.

It remains to be seen whether this method of assembling a corpus can be extended to other Old Icelandic texts or to texts from different languages. Orthographic standardisation built on dictionaries and grammars is a rather recent phenomenon in language history, and it is obvious that an automatic tagging of texts from historical linguistic varieties always profits from some sort of normalisation of the text which does not only level out graphic variation inside or between individual texts but also smaller historical changes during one historical linguistic epoch.

The comparably large differences between historical variants of languages like German or English make it a rather unrealistic option to base such a normalisation on the modern language variety, as the following example from chapter 11 of the Old English *Beowulf* illustrates:

> *Ða com of more under mist hleoþum grendel gongan godes yrre bær mynte se man scaða manna cynnes sumne besyrwan insele þam hean wod under wolcnum to þæs þe he win reced gold sele gumena gearwost witte.* (Zupitza Ed. 1959, p. 34)
>
> Then, from the moor, Grendel came moving under the misty hills. God's curse rested on him. The foe of men intended to ensnare some human being in the high hall. He advanced under the clouds to a place where he could easily recognize the wine hall of men, decorated with gold. (Heatt Trans. 1988, p. 20)

For Icelandic, on the other hand, a normalisation based on Old Icelandic is not necessarily a preferable option. The manuscripts of *Njáls saga* stem from between 1300 and 1875, and for a majority of the manuscripts the classical Old Icelandic norm used for dictionaries and editions is more distant than the modern Icelandic orthography. The following example gives a short extract from the medieval parchment manuscript AM 486 4to (*Reykjabók*, Copenhagen, Den Arnamagnæanske Samling), and the late 17th century paper manuscript Lbs 3505 4to (Reykjavík, Landsbókasafn Íslands) together with the text from AM 468 4to normalised to Modern Icelandic spelling:

> *hann keyrði þa hest sinn ok ridr mikit ok er hann metir kol mælti atli til hans. Gengr vel klyfia bandit segir Atli. þat mun þik skipta ongv mann fylan segir kolr ok ongan þann er þaðan er.* (AM 468 4to, ca 1300)
>
> *Hann keyrde þá hest sinn og rydur miked, og er hann mæter Kol mællte Atla til hanns, geingur vel Klifiaburdurenn sagde Atle, þad mun þig skipta aungvu Mannfylann þyn segir Kolur, og aungvann þann er frá þier er.* (Lbs 3505 4to, 1698)
>
> *Hann keyrði þá hest sinn og ríður mikið. Og er hann mætir Kol mælti Atli til hans: „Gengur vel klyfjabandið?“ segir Atli. „Það mun þig skipta engu, mannfýlan,“ segir Kolur, „og engan þann er þaðan er.“* (AM 468 4to in modern orthography)
>
> He spurred his horse and rode hard. When he came to Kol he said, 'Is your pack-horse work going well?' 'That's no business of yours, you scum,' said Kol, 'or of anybody from your place.' (Cook Trans. 2001, p. 62)

At least for the framework of our project, considering the current general conditions for research on Icelandic manuscripts which require an efficient use of the limited financial and human resources and thus an approach oriented on multiple usability, it seems to be more realistic and sensible to opt for the adaptation of existing technology to research questions in connection with historical linguistic data and not for the development of technical solutions specifically developed for medieval texts with only a limited application area.

Research on Old Icelandic texts can definitely benefit from technologies developed in the framework of the NLP-community. The results from our pilot study using language tools developed for Modern Icelandic to analyse Old Icelandic manuscript texts exceeded our expectations.

It goes without saying that a linguistically annotated corpus of Old Icelandic texts would be of great use for research on medieval Icelandic texts. However, a precondition for the use of such tools with Old Icelandic corpora is a compatibility with the common standards in the field of Old Norse manuscript studies. At the moment, TEI-XML is the most widespread format for philologically ambitious electronic representations of medieval Icelandic texts, annotation tools would thus have to be able to handle this format. Due to the high workload of manual transcription, corpus projects dealing with manuscripts usually involve the participation of a larger number of people over a longer time period. An easy manageability of the tools and a compatibility with standard operating systems and hardware would therefore be an advantage.

On the other hand we are convinced that our experience from work with historical texts can also be utilised to develop language tools and resources further and to broaden their usability. A manually corrected lemmatised sample corpus of older Icelandic texts in their original spelling and containing a syntactic segmentation and a morphosyntactic annotation would not only be a useful tool for compiling larger corpora of earlier Icelandic texts, but also for (historical) linguistic studies.

Research done in the project *Variance of* Njáls saga shows that contemporaneous manuscripts of one text often exhibit differences on linguistic levels that are usually assumed to be connected to historical developments of the language (e.g. word order or the use of certain grammatical constructions, cf. Zeevaert 2013). Grammatical differences between manuscripts and deviations from standard forms are often not rendered in normalised editions. In addition to this, normalised editions in some instances treat grammatical deviations in certain constructions not as variants, but as errors that have to be corrected. Both may distort our picture of historical linguistic developments. A corpus that gives access to the original spelling and wording in a manuscript without abandoning the advantages of a normalised spelling would be a useful tool to avoid this.

I don't know whether the NLP-community is willing to judge our experiment of adapting an existing language tool designed for a contemporary language to its historical variant as successful. Eventually we did not tag an Old Icelandic text with a tagger built for Modern Icelandic but rather tagged an Old Icelandic text transferred to Modern Icelandic and converted the text back to Old Icelandic only after the tagging was finished. For us it is definitely an advantage to be able to use the limited work force of our project more efficiently and to focus more on the scientific goals of our project. We would be delighted if in addition to this we might be able to contribute to the construction of a corpus that would serve as a useful tool for all scientists interested in Old Icelandic texts.

7. Acknowledgements

The project *Variance of* Njáls saga *(Breytileki Njáls sögu,* principal investigator Svanhildur Óskarsdóttir) is funded by the Icelandic Centre for Research (Rannsóknarmiðstöð Íslands, Rannís). Thanks to Sigrún Helgadóttir (Stofnun Árna Magnússonar í íslenskum fræðum, orðfræðisvið) for her generous support in connection with all kinds of questions concerning the tagging of Icelandic texts and Kristín Bjarnadóttir (Stofnun Árna Magnússonar í íslenskum fræðum, orðfræðisvið) and Eiríkur Rögnvaldsson (Háskóli Íslands, Íslensku- og menningardeild) for their valuable help with this article.

8. References

Brants, T. (2000). TnT – A statistical part-of-speech tagger. In *6th Applied Natural Language Processing Conference. Proceedings of the conference.* [Stroudsburg]: Association for Computational Linguistics, pp. 242--231.

Egilsson, S. Y. (Ed.) (2003). *Brennu-Njáls saga. Texti Reykjabókar.* Reykjavík: Bjartur.

Cook, R. (Trans.) (2001). *Njal's saga.* London etc.: Penguin.

Gunnlaugsson, G. M. (2003). Stafrétt eða samræmt? Um fræðilegar útgáfur og notendur þeirra. *Gripla,* 14, pp. 197--235.

Haugen, O. E. (2002). Nordic language history and philology: Editing earlier texts. In O. Bandle et al. (Eds.), *The Nordic languages. An international handbook of the history of the North Germanic languages. Volume 1* (Handbooks of Linguistics and Communication Science, 22.1). Berlin, New York: Walter de Gruyter, pp. 535--543.

Helgadóttir, S. (2007). Mörkun íslensks texta. *Orð og tunga,* 9, pp. 75--107.

Hieatt, C. (Trans.) (1988). *Beowulf and other English poems.* Toronto etc.: Bantam.

Jónsson, M. (1996). Var þar mokað af miklum usla. Fyrsta atrenna að Gullskinnugerð Njálu. In *Þorlákstíðir sungnar Ásdísi Egilsdóttur fimmtugri 26. október 1996.* Reykjavík: Menningar- og minningarsjóður Mette Magnussen, pp. 52--55.

Loftsson, H. (2013). Tagging the past. Experiments using the Saga Corpus. In *NODALIDA 2013. Proceedings of the 19th Nordic Conference of Computational Linguistics. May 22-24, 2013. Oslo, Norway* (Linköping electronic conference proceedings, 85). Linköping: Linköping University Electronic Press, pp. 89--104.

Loftsson, H., Rögnvaldsson, E. (2007a). IceNLP: A natural language processing toolkit for Icelandic. In *Interspeech 2007. 8th Annual Conference of the International Speech Communication Association. Antwerp, Belgium, August 27-31, 2007. Volume 1.* N.P.: ISCA, pp. 717--720.

Loftsson, H., Rögnvaldsson, E. (2007b). IceParser: An incremental finite-state parser for Icelandic. In J. Nivre et al. (Eds.), *Proceedings of the 16th Nordic Conference of Computational Linguistics NODALIDA-2007.* Tartu: University of Tartu, pp. 128--135.

Pind, J., Magnússon, F., Briem, S. (1991). *Íslensk orðtíðnibók.* Reykjavík: Orðabók Háskólans.

Rögnvaldsson, E., Helgadóttir, S. (2011). Morphosyntactic tagging of Old Icelandic texts and its use in studying syntactic variation and change. In C. Sporleder, A. van den Bosch, K. Zervanou (Eds.), *Language Technology for cultural heritage. Selected papers from the LaTeCH workshop series.* Heidelberg, Dordrecht, London, New York: Springer, pp. 63--76.

Þorkelsson, J. (1889). Om håndskrifterne af Njála. In *Njála udgivet efter gamle håndskrifter af det Kongelige nordiske oldskrift-selskab. Andet bind* (Íslendinga sögur udgivne efter gamle haandskrifter, 4). København: Gyldendalske boghandel, pp. 647--787.

Zeevaert, L. (2013). Axes, halberds, bows or stones? Tools to get to grips with linguistic variation in the manuscripts of Njál's saga. Manuscript submitted for publication.

Zeevaert, L. (in prep.). Mörkum Njálu! An annotated corpus to analyse and explain grammatical divergences between 13th-century manuscripts of *Njáls saga.* To appear in the proceedings of LREC 2014.

Zupitza, J. (Ed.) (1959): *Beowulf.* London etc.: Oxford University Press.

Computational Analysis of Historical Documents: An Application to Italian War Bulletins in World War I and II

Federico Boschetti[*], Andrea Cimino[*], Felice Dell'Orletta[*], Gianluca E. Lebani[†], Lucia Passaro[†], Paolo Picchi[*], Giulia Venturi[*], Simonetta Montemagni[*], Alessandro Lenci[†]

[*]Istituto di Linguistica Computazionale "Antonio Zampolli", CNR- Pisa (Italy)
[†]CoLing Lab, Dipartimento di Filologia, Letteratura e Linguistica, University of Pisa (Italy)
E-mail: federico.boschetti@yahoo.com, andrea.cimino@ilc.cnr.it, felice.dellorletta@ilc.cnr.it,
gianluca.lebani@for.unipi.it, lucia.passaro@for.unipi.it, paolo.picchi@ilc.cnr.it, giulia.venturi@ilc.cnr.it,
simonetta.montemagni@ilc.cnr.it, alessandro.lenci@ling.unipi.it

Abstract

World War (WW) I and II represent crucial landmarks in the history on mankind: They have affected the destiny of whole generations and their consequences are still alive throughout Europe. In this paper we present an ongoing project to carry out a computational analysis of Italian war bulletins in WWI and WWII, by applying state-of-the-art tools for NLP and Information Extraction. The annotated texts and extracted information will be explored with a dedicated Web interface, allowing for multidimensional access and exploration of historical events through space and time.

Keywords: Digital History, War Bulletins, NLP, World War I, World WAR II, Information Extraction

1. Introduction

World War (WW) I and II represent crucial landmarks in the history of mankind: They have affected the destiny of whole generations and their consequences are still alive worldwide. Unfortunately, the knowledge of these events is progressively fading away, especially among young generations. The first centenary of the beginning of the WWI raises the moral issue of how to preserve the historical memory of these events, making them accessible to a larger audience, not limited to scholars and experts. Methods and tools for Natural Language Processing (NLP) can play an important role to achieve this goal, by providing new way to access historical documents and events (cf. Ide and Woolner 2004, Cybulska and Vossen 2011).

In this paper we present an ongoing project to carry out a computational analysis of Italian war bulletins in WWI and WWII, by applying state-of-the-art tools for NLP and Information Extraction. The project relies on the collaboration between computational linguists and historians, in particular Prof. Nicola Labanca (University of Siena), one of the major experts on Italian military history during the WWs. This project has several elements of originality and challenge. To the best of our knowledge, this is the first computational analysis of this kind of historical texts. Moreover, WWI Italian war bulletins have never been digitalized before. The type of language (Italian of first half of the 20th century) and domain (military) require an intense effort of adaptation of existing NLP tools. Bulletins are annotated automatically with different types of information, such as simple and multi-word terms, named entities, events, their participants, time and georeferenced locations. The annotated texts and extracted information will be explored with a dedicated Web interface, allowing for multidimensional access and explorations of historical events through space and time.

The paper is organized as follows. In the next section, we provide an overview of the project, as well as a motivation of the text choice. In section 3, we describe some of the current works, mainly focusing on the digitalization of the WWI bulletins, the adaptation of existing NLP tools, the first experiments for term and event extraction carried out on WWII bulletins. In section 4 we present the next steps for the project implementation and some future plans.

2. Project overview

2.1 War bulletins and NLP

War bulletins (WBs) were issued by the Italian *Comando Supremo* "Supreme Headquarters" during WWI and WWII as the official daily report about the military operations of the Italian armed forces. WBs were published on major newspapers, and during WWII they were also radio broadcasted. WBs provide a dynamic picture of the unfolding of war events, from the official perspective of the Italian Government. They allow us to follow the complex series of events in the two WWs, respectively from the 24th May 1915 to the 11th November 1918, and then from the 10th June 1940 to the 8th September 1943 (date of the armistice between Italy and the Allies, and the dissolution of the Italian army). It is important to remark that WBs do not provide a faithful and objective picture of the war. Events can be missing or misrepresented for military or propaganda reasons. For instance there is a systematic overestimation of enemy losses and Italian achievements, and conversely the underestimation of Italian defeats and losses.

We have focused our research on the computational analysis of WBs because they represent a particularly interesting source not only to reconstruct the military history of the two WWs, but also to study the propaganda strategies of the Italian Government, as well as the way the enemy was depicted by official sources. The collection of WBs for WWI was first published in 1923, and the one for WWII was published in 1970. The former has never been digitalized, while an html version of the latter is freely available on the Web.[1]

The reason for working simultaneously on WWI and WWII is twofold. First of all, nowadays historians commonly assume that, despite their several differences, WWI and WWII should not be regarded as two separated events, but rather as two episodes of a single 30-years European war. Secondly, the comparison between the bulletins of the two WWs is extremely interesting under many respects. From an historical point of view, we can observe the radical change in warfare between the two conflicts: for instance, WWI was mainly a static trench war, while WWII was a movement war fought on as different fronts as Greece, North Africa, Atlantic Ocean, etc. Some weapons, like gases, represent the hallmark of WWI, but were not used in WWII, which was instead dominated by tanks and aviation. These types of information easily emerge from the analysis of WBs. Moreover, during WWI Italy had a liberal, aristocratic government, while in WWII it was ruled by the dictatorial fascist regime. This difference has important consequences on the language, propaganda style, etc. to be found in WBs.

2.2 Work program

Our project of computational analysis of WBs include the following phases:

1. **text digitalization of WWI bulletins**. This phase is currently ongoing, and described in section 3.1.
2. **NLP annotation of WB**. The corpus of bulletins will be POS-tagged, lemmatized and dependency parsed with existing tools for Italian NLP. Waiting for the complete digitalization of WWI bulletins, we have started processing the WWII ones. This part of the project also involves an important work on NLP tools adaption to the target domain and genre, as illustrated in section 3.2.
3. **Statistical analysis and information extraction.** This is the core and most challenging part of the project. Its goal is to index texts with a large amount of linguistic and semantic information, to highlight significant text features as well as to identify the most prominent historical events. This part includes:
    a. *statistical profiling* – each bulletin will be assigned textual statistics like number of word tokens, type-token ratio, readability scores, etc. This information is also extremely useful to characterize the development of historical events. For instance, shorter bulletins correspond to periods of less intense military operations, successful operations are described in greater details, while defeats are usually reported in a more sketchy form, etc.;
    b. *term extraction* – simple and multi-word terms will be automatically extracted from texts, to provide users more advanced search keys. This activity is carried out by adapting existing term extraction tools (cf. section 3.3.);
    c. *named entity recognition* – proper names will be identified and classified into general (e.g., person, location, etc.) and domain-specific (e.g., ship, military unit, airplane, etc.) semantic categories (cf. section 3.4);
    d. *event extraction* – using standard Information Extraction techniques (cf. in particular Ide and Woolner (2004), Cybulska and Vossen (2011) for previous research on event extraction from historical texts) we will identify instances of major event types (e.g., *bombing*, *sinking*, *battles*, etc.) their participants, places and times. Event timestamps will be derived from the bulletin date, and will be used to reconstruct the event timeline.
4. **data linking** – various types of extracted data will be linked to external sources. First of all, location names will be georeferenced. This is particularly challenging given the spelling variations of many location names (e.g., Arabic ones), as well as the changes that some of them have undergone through time (e.g. various WWI locations were Italian at that time, and have now become Slovenian, etc.). Moreover, we plan to provide links between other extracted data external sources (e.g., Wikipedia pages describing weapon types or warships, etc.).
5. **browse and search interface** – a key aspect of the project is the development of an advanced and user-friendly interface to explore the texts. Our target users are not limited to scholars but also crucially include students. We intend to provide multidimensional access keys to WBs, which will be queried for simple words, multi-word terms, semantic classes, event types, locations, and persons, etc. The tools will also include user-friendly visualization modules such as "word clouds", event timeline, even projection on dynamical geographical maps, etc. The interface will allow experts as well as non-expert users to follow historical events in space and time, thereby gaining a new view of the parallel development of war actions across multiple fronts.

3. Ongoing work

3.1 Text digitalization of WWI bulletins

We are currently digitalizing the WBs of WWI published in *I Bollettini della Guerra 1915-1918*, preface by Benito Mussolini, Milano, Alpes, 1923 (pages VIII +

[1] http://www.alieuomini.it/pagine/dettaglio/bollettini_di_guerra

596). The exemplar in our possession is preserved in a good state, due to the high quality of paper and printing. The book has been accurately unbound, in order to acquire images from loose pages. This technique drastically reduce scanning artifacts, avoiding the necessity to digitally straighten and deskew page images. The same conditions of brightness and contrast are assured not only to each page of the book, but also to each area of the page, increasing the accuracy of the Optical Character Recognition (OCR).

OCR has been performed using the open source application Tesseract, in bundle with the Italian training set.[2] The accuracy and the F-score have been calculated on a random sample of 10 pages out of 604. In order to calculate accuracy and F-score, the texts of the sample have been manually corrected and aligned to the OCR output applying the Needleman-Wunsch algorithm, in order to identify the number of exact matches, the number of substitutions (e.g., "m" instead of "n"), the number of omissions (i.e., characters neglected by OCR) and, finally, the number of insertions (i.e., artifacts added by the OCR engine, such as an "i" at the end of the line).

Accuracy is defined as the ratio between the matches and the sum of the matches (m) with all the other phenomena, substitutions (s), insertions (i) and deletions (d). Precision (P) is defined as $m/(m+s+i)$, Recall (R) as $m/(m+s+d)$, and F-score as $2PR/(P+R)$. The accuracy on the test sample is 97.87% and the F-score is 98.68%. OCR performances will be improved by using three different OCR engines: the aforementioned Tesseract accompanied by OCRopus[3] and Gamera.[4] The results will be aligned and a voting system will be applied, according to the methods described in Lund-Ringger (2009) and in Boschetti et al. (2009). Finally, manual corrections will be performed.

3.2 Text processing and NLP tools adaptation

Parallely to the digitalization of WWI bulletins, we are carrying out experiments for the automatic linguistic annotation of WWII bulletins with NLP tools. The bulletins were automatically downloaded and cleaned of HTML tags and boilerplates. The resulting corpus was automatically POS tagged with the Part-Of-Speech tagger described in Dell'Orletta (2009) and dependency-parsed with the DeSR parser (Attardi et al., 2009) using Support Vector Machines as learning algorithm. They represent state-of-the-art tools for Italian NLP. In particular, the POS tagger achieves a performance of 96.34% and DeSR, trained on the ISST–TANL treebank (consisting of articles from newspapers and periodicals), achieves a performance of 83.38% and 87.71% in terms of Labeled Attachment Scores (LAS) and Unlabeled Attachment Scores (UAS) respectively when tested on texts of the same type.

However, since Gildea (2001) it is widely acknowledged that statistical NLP tools have a drop of accuracy when tested against corpora differing from the typology of texts on which they were trained. This is also the case with WBs: they contain lexical and syntactic structures characterising the Italian of the past century and they contain domain-specific lexicon. Sentences are typically shorter than in newspapers, but on the other hand they are often quite elliptic and full of omissions due to the telegraphic style. They also contain lots of old-fashioned syntactic constructions that may hamper linguistic annotation. The percentage of lexical items contained in the WWII corpus and in the training of DeSR parser (74%) is much lower than in the corpus of contemporary newspaper articles used as test set (about 90%). We expect this trend to be even in WWI bulletins, since Italian of early 20th century was very different from contemporary one. In fact, standard Italian was still very much under formation, due to the recent political unification of the country just 50 years before the Great War. Assuming that new lexicon introduces new syntactic constructions, we can assume that the parser tested on Bulletins can have a quite high drop of accuracy with respect to the accuracy achieved on the reference test set.

In order to overcome this problem, in the last few years several methods and techniques have been developed to adapt current NLP systems to new kinds of texts. They can be broadly divided in two main typologies: Self-training (McClosky et al., 2006) and Active Learning (Thompson et al., 1999). To adapt NLP tools to WBs, we are using the self-training approach to domain adaptation described in (Dell'Orletta et al., 2013), based on ULISSE (Dell'Orletta et al., 2011). ULISSE is an unsupervised linguistically-driven algorithm to select reliable parses from the output of dependency annotated texts. Each dependency tree is assigned a score quantifying its reliability based on a wide range of linguistic features. After collecting statistics about selected features from a corpus of automatically parsed sentences, for each newly parsed sentence ULISSE computes a reliability score using the previously extracted feature statistics. From top ranked parses according to their reliability score, different pools of parses are selected for training. The new training set extends the original one with the new selected parses including lexical and syntactic characteristics specific to the target domain, in this case the bulletins. We expect that the NLP tools trained on this new training set can improve their performance when tested on the target domain.

3.3 Term extraction

Single–word terms, e.g. *velivolo* (aircraft), and multi–word terms (complex terms), e.g. *velivolo da ricognizione marittima* (*maritime reconnaissance aircraft*) are the first types of information we intend to extract from WBs. They will be used for text indexing and querying. We are currently applying to WWII bulletins two methods for automatic term extraction from

[2] https://code.google.com/p/tesseract-ocr/downloads/list
[3] https://code.google.com/p/ocropus
[4] http://gamera.informatik.hsnr.de

Italian texts, T2K^2 (Dell'Orletta et al., 2014) that follows the methodology described in (Bonin et al. 2010) and EXTra (Passaro et al. 2014), both combining NLP techniques, linguistic and statistical filters.

Term extraction with T2K^2 is articulated in three main steps. In the first step, the POS-tagged and lemmatized text is searched for on the basis of linguistic filters aimed at identifying a) nouns, expressing candidate single terms and b) POS patterns covering the main morphosyntactic patterns expressing candidate complex terms: e.g., noun + adjective (e.g., *velivoli britannici*, British aircraft), noun + preposition + noun (e.g. *velivolo d'assalto, Aircraft of Assault*), etc. In the second step, the candidate terms are ranked according to their C-NC Value, a statistical filter described in (Frantzi et al. 1999) and (Vintar 2004). C-value is a method for term extraction which aims to improve the extraction of nested terms. The method produces a list of candidate terms that are ordered by their termhood. Then, the NC-value incorporates context information to the C-value method, improving term extraction. In the last step, a contrastive method is applied against the list of ranked terms using a contrastive function *CSmw* newly introduced in (Bonin et al. 2010). This function is applied to the top list of the terms resulting from the statistical filtering step. This procedure is oriented to a) prune common words from the list of domain-relevant terms and b) rank the extracted terms with respect their domain relevance. This method is based on the comparison of the distribution of terms across corpora of different domains. We also used T2K^2 to extract relevant domain-specific verbs representing instances of some of the major events. Table 1 reports a sample of the top-ranked domain verbs extracted with T2K^2 using contemporary newspaper collections as reference corpora in the contrastive method.

mitragliare	"to machine-gun"
spezzonare	"to bomb with incendiary devices"
bombardare	"to bomb"
abbattere	"to shoot down"
silurare	"to torpedo"
incendiare	"to set on fire"
affondare	"to sink"
attaccare	"to attack"

Table 1 – Sample of domain verbs extracted with T2K^2

Term extraction with EXTra is carried out in a very similar way, except for two major differences. First of all, instead of using flat POS-sequences, EXTra identifies candidate multi-word terms with structured patterns that take into account the internal syntactic structure of term phrases. For instance, the term *bomba di grosso calibro* "heavy bomb" is identified as an instance of the pattern *[noun, preposition [adjective, noun]]*, while the term *apprestamenti difensivi del nemico* "enemy defensive works" is identified as an instance of the pattern *[noun, adjective, preposition [noun]]*. Pattern structure is then used to guide the process of statistical term weighting. Terms are weighted using a new measure that recursively applies standard association measures (e.g., Pointwise Mutual Information, Local Mutual Information, Log-Likelihood Ratio, etc.) to the internal structure of complex terms. The intuition is that the degree of termhood of a candidate pattern depends not only the statistical association between its parts, but also on whether these parts are also terms. The EXTra term weighting algorithm works as follows:

i) *base step* - we measure the association strength σ of each candidate 2-word term <w$_1$,w$_2$>, and we then select the set of terms T={t$_1$,…,t$_n$} whose score σ is above an empirically fixed threshold;

ii) *recursive step* - we measure the association strength σ of any *n*-word candidate term <c$_1$,c$_2$>, where either c$_1$, or c$_2$ or both belong to T:

$$\sigma(<c_1,c_2>)*S(c_1)*S(c_2)$$

where S(c$_i$) =1, if c$_i$ is a word, while *S(c$_i$) = (log$_2$ σ(c$_i$))/k* if c$_i$ ∈T. The parameter *k* controls the length of complex terms: The smaller the *k*, the higher weight is assigned to longer terms. The candidate terms whose score σ is above an empirically fixed threshold are then added to T. The recursive step ii) is repeated for any extracted pattern, so that multi-word terms of any length are assigned a weight. Table 2 reports a sample of the top ranked complex terms extracted with EXTra from the WWII bulletins, using Local Mutual Information (as the association score σ (Evert 2008).

term	**LMI**
fronte greco	927.30
tenente di vascello	699.04
lieve danno	659.14
aereo nemico	623.10
capitano di corvetta	593.89
artiglieria contraerea	548.13
bomba di grosso calibro	500.12
velivolo nemico	496.32
bollettino odierno	456.01
caccia germanico	441.91
obiettivo militare	423.78
campo di aviazione	422.07
vasto incendio	416.86
caccia tedesco	413.63
piroscafo di medio tonnellaggio	366.60

Table 2 – Sample of terms extracted with EXTra

Extracted domain-specific entities are then organized into fragments of taxonomical chains, grouping entities which share the semantic head (e.g., *fronte cirenaico, fronte egiziano, fronte greco, fronte tunisino, fronti dello scacchiere, fronti* terrestri share the semantic head *fronte* "front") or the modifiers defining their scope (eg. a*ereo britannico, apparecchio britannico, aviazione britannica, caccia britannico, incursione aerea britannica, velivolo* *britannico* share the modifier *britannico* “British”).

3.4 Named Entity Recognition

WBs report military events and therefore are full of proper names of places, persons and organizations (e.g. military formations). Therefore, NER plays a crucial role to obtain a semantic access to the content of these texts. The named entities are identified and classified using ItaliaNLP NER (described in Dell'Orletta et al., 2014). This module is a classifier based on Support Vector Machines using LIBSVM (Chang and Lin, 2001) that assigns a named entity tag to a token or a sequence of tokens. ItaliaNLP NER relies on 5 kinds of features:

- *orthographic features*: e.g., capitalized letters, presence of non–alphabetical characters, etc.;
- *linguistic features*: the lemma, POS, prefix and suffix of the analyzed token;
- *dictionary look-up features*: check if the analyzed token is part of an entity name occurring in People, Organization and Geo-political gazetteers;
- *contextual features*: these features refer to orthographic, linguistic and dictionary look–up features of the context words of the analyzed tokens;
- *non local features*: in the case of identical tokens , they take into account previous label assignments to predict the label for the current token (Ratinov and Roth, 2009).

ItaliaNLP NER is trained on I-CAB (Italian Content Annotation Treebank) (Magnini et al., 2006), the dataset used in the NER Task at EVALITA 2009 (Speranza, 2009) including four standard named entity tags, i.e. Person, Organization, Location and Geopolitical classes. The NE tagger accuracy is in line with the state of the art when compared with the systems that participated to the EVALITA shared task.(F-Measure: ~80%).

For the analysis of the Italian, we are currently adapting the NER under various respects. First of all, we plan to extend the range of semantic classes covered by the NER, for instance identify names of airplanes (e.g., *Gloster Gladiator*), ships (e.g., *Valiant*), and military organizations (e.g., *Divisione Ariete* “Ariete Division”) which frequently occur in this type of texts. Moreover, we intend to adapt the NER to the domain of WBs. To this purpose we plan to exploit the rich analytical index accompanying WWII bulletins. This index contains the names of places, persons and military formations mentioned in the WBs, with information about its semantic class. Using this index, we are automatically identifying named entities in the bulletins, thereby producing a fully annotated version of the WWII corpus with NE classes. This corpus will be used to train the NER before its application to the WWI bulletins (which instead lack this type of semantic indexing).

4. Conclusions and future plans

The short-term agenda of our project includes: i.) completing the digitalization and manual correction of the WWI bulletins; ii.) developing the module for event extraction and location georeference of WWII texts; iii.) applying the NLP and Information Extraction modules to WWI texts; iii.) designing and implementing the Web search interface. The output of text processing will consist of the two bulletin corpora annotated with XML and RDF metadata indexing texts with various types of linguistic and semantic information.

This project has also a great possibilities for future, long-term developments. First of all, we plan to enrich the linking of WBs to other types of external data. As we said above, WBs provide a very biased view of military events, because of propaganda or military reasons. It is therefore interesting to perform a kind of cross-text event co-reference to link the events extracted from the WBs to other historical sources reporting information about the same historical fact. Secondly, we intend to extend our project to cover WBs issued by other countries involved in WWI and WWII. This will again allow us to gain information about the way the same events are reported by different fighting countries (e.g., the battle of El Alamein described by the Axis Powers or the Allies). Computational linguistic methods have surely a great potential for applications on historical text. We believe that a project like ours can contribute to prove the possibilities offered by NLP to find new ways to study our past and to learn history.

Acknowledgements

We gratefully thank Prof. Nicola Labanca for his constant advice and support on the military history of World War I and II.

References

Attardi, G., Dell’Orletta, F., Simi, M., Turian, J. (2009). Accurate Dependency Parsing with a Stacked Multilayer Perceptron. In *Proceedings of EVALITA 2009*, Reggio Emilia, Italy.

Bonin, F., Dell’Orletta, F., Montemagni, M., Venturi, G. (2010). A Contrastive Approach to Multi-word Extraction from Domain-specific Corpora. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10),* Valletta, Malta, pp. 19-21.

Boschetti, F., Romanello, M., Babeu, A., Bamman, D. and Crane, G. (2009). Improving OCR accuracy for classical critical editions. In *Proceedings of the 13th European conference on Research and advanced technology for digital libraries (ECDL'09),* Berlin, Heidelberg: Springer-Verlag, pp. 156-167.

Chang C-C., Lin, C-J. (2001). LIBSVM: a library for Support Vector Machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

Cybulska, A., and Vossen, P. (2011). Historical Event Extraction from Text. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities,* Portland, OR, USA, pp. 39-43.

Dell’Orletta, F. (2009). Ensemble system for Part-of-Speech tagging. In *Proceedings of EVALITA*

2009, Reggio Emilia, Italy.

Dell'Orletta, F., Venturi, G., Cimino, A., Montemagni, S. (2014). T2K^2: a System for Automatically Extracting and Organizing Knowledge from Texts. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14),* Reykjavik (Iceland) (forthcoming).

Dell'Orletta, F., Venturi, G., Montemagni, S. (2011). ULISSE: an unsupervised algorithm for detecting reliable dependency parses. In *Proceedings of CoNLL 2011*, Portland, Oregon, pp. 115-124.

Dell'Orletta, F., Venturi, G., Montemagni, S. (2013). Unsupervised Linguistically-Driven Reliable Dependency Parses Detection and Self-Training for Adaptation to the Biomedical Domain. In *Proceedings of the 2013 Workshop on Biomedical Natural Language Processing (BioNLP 2013),* Sofia, Bulgaria, pp. 45-53.

Evert, S. (2008). Corpora and collocations. In A. Lüdeling & M. Kytö (eds.), *Corpus Linguistics. An International Handbook*, article 58. Mouton de Gruyter, Berlin.

Frantzi, K., Ananiadou, S. (1999). The C–value / NC Value domain independent method for multi-word term extraction. *Journal of Natural Language Processing*, 6(3), pp. 145-179.

Gildea, D. (2001). Corpus Variation and Parser Performance. In *Proceedings of EMNLP 2001*, Pittsburgh, PA, pp. 167-202.

Ide, N., and Woolner, D. (2004). Exploiting Semantic Web Technologies for Intelligent Access to Historical Documents. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisboa, Portugal, pp. 2177-2180.

Lund, W. B., and Ringger, E. K. (2009). Improving optical character recognition through efficient multiple system alignment. In *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries (JCDL '09)*. ACM, New York, NY, USA, 231-240.

Magnini, B., Pianta, E., Girardi, C., Negri, M., Romano, L., Speranza, M., Bartalesi Lenzi, V., Sprugnoli, R. (2006). I-CAB: the Italian Content Annotation Bank. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06).*

McClosky, D., Charniak, E., Johnson, M. (2006). Reranking and self-training for parser adaptation. In *Proceedings of ACL 2006*, Sydney, Australia, pp. 337-344.

Passaro, L., Lebani, G. and Lenci A. (2014), Extracting terms with EXTra. submitted.

Ratinov, L., Roth, D. (2009). Design challenges and misconceptions in named entity recognition, In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning,* pp. 147-155.

Thompson, C. A., Califf, M. E., Mooney, R. J. (1999). Active Learning for Natural Language Parsing and Information Extraction. In *Proceedings of the Sixteenth International Conference on Machine Learning (ICML'99),* San Francisco, CA, USA, Morgan Kaufmann Publishers Inc., pp. 406-414.

Vintar, Š. (2004). Comparative Evaluation of C-value in the Treatment of Nested Terms. In *Proceedings of Memura 2004 – Methodologies and Evaluation of Multi-word Units in Real-World Applications,* (LREC 2004 Workshop), pp. 54-57.

Discovering and Explaining Knowledge in Multilingual Historical Documents

Cristina Vertan, Walther v. Hahn

University of Hamburg, Research Group “Computerphilology”
Vogt-Kölln Straße 30, 22527 Hamburg, Germany
E-mail: cristina.vertan@uni-hamburg.de, vhahn@informatik.uni-hamburg.de

Abstract

The recent digitization efforts had as its result a relatively big number of digital versions of ancient manuscripts and old printed books. By this digitization process the preservation and availability of the text is secured but it is still restricted to a small number of scientists who understand the language and are familiar with obsolete geographical names, historical figures etc. Digital repositories, however, offer for the first time the opportunity, that a broad public can get familiar with this cultural testimony. It is thus of great importance that some effort is invested in enhancing historical texts with explanations and links to modern knowledge bases. Text technology offers methods, which can support this activity. In this paper we present a first attempt of creating an annotation schema for tagging linguistic, domain specific and general knowledge in historical texts. We exemplify this on a multilingual text written originally in Latin, translated into German and containing a big number of lexical materials in Romanian, Ancient Greek and Latin. We explain the necessity of such annotations, describe the challenges and present first result.

Keywords: historical languages, multilinguality, knowledge discovery.

1. Introduction

The digitization campaign during the last years for the first time made available historical documents from different regions of the world and different time periods to a non-scientific public. This is a unique opportunity for investigating and rediscovering history of places and countries being less in focus currently, like e.g. central and Eastern Europe or the Balkan countries.

Digital libraries include some documents from this part of Europe, many others wait to be digitized but even then, processing and understanding them is a challenge, as:

- Laymen, except researchers in the respective fields, (Turcology, Romanian history, German dialects in Transylvania) cannot understand the language or, much worse, misinterpret its contents.
- The sources are not searchable beyond strings, esp. on a content level.
- No links can be used between parts of the same text or to other texts.

For other researchers (ethnology, general linguistics, medieval history) or for the public the old documents are not really meaningful because

- the language changed (morphology / syntax /semantics)
- geographical entities changed their name or borders
- political entities are not any longer known to everybody

In order to make these texts usable it is necessary to design a platform where

- the content of historical texts is annotated with linguistic and multimodal information, thus prepared for further processing and informed understanding for readers,
- multilingual editions are aligned,
- readers have the opportunity of semantic queries and cross-lingual text retrieval,
- text-mining tools are available for researchers.

In order to explain the challenges of the processing of such documents, we exemplify it by the works of Dimitrie Cantemir, prince of Moldavia, and member of the Berlin Academy of Science, who offered - in the XVIIIth century - first scientific works about the history of the Ottoman Empire as well as his own country „Moldavia“ (nowadays corresponding with the territories in eastern Romania, the Republic of Moldova and some parts of Ukraine). He himself wrote in Latin, but his works were translated into German, French and English, and later into Romanian. His work about the Ottoman Empire was the reference literature until middle of XIXth century. The other volume „Descriptio Moldaviae“ is the first complete historical description of Moldavia and includes also the first map of the described territory. In this sense his works are of particular interest for the history of the Balkan countries, however, we will illustrate in this article that the translations of the books do not correspond exactly to the original and even the translations are not open to straightforward understanding, moreover, an automatic processing of the texts without pre-processing and an annotation step is practically impossible.

In the following sections we exemplify our findings with passages from „Descriptio Moldaviae“, as use-case for other extensive multilingual documents. In section 2 we present the challenges in a structured manner, in section 3 we introduce an annotation schema, which may be used for embedding general supportive knowledge. Section 4 contains the conclusions and envisaged steps for further work.

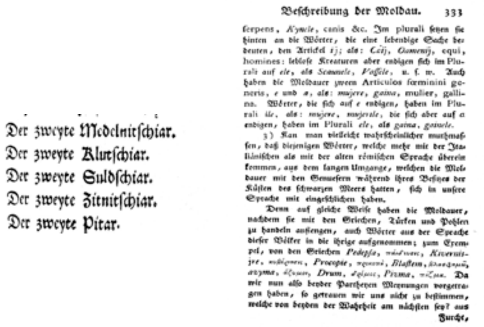
2. Challenges for processing multilingual historical documents

In this section we describe the challenges we identified during the analysis of Dimitre Cantemir's „Descriptio Moldaviae", in its German translation from 1771. We inspected a facsimile edition from 1973 (Cantemir 1771), which includes together with the facsimile, a foreword and a glossary of terms compiled by the editor of the facsimile 1973. This detail is relevant for he discussion of various available knowledge sources.

2.1. Digitization challenges

As we intended to annotate content, the first requirement is to have material available in Unicode-format (text formats and not images). There are two options: Either the text is typed in: This involves a lot of human work and it is error prone, especially due to the presence of words in several languages, or OCR-Software is used. We have chosen the second solution, also to test the applicability of OCR-Software for old multilingual texts.

Most parts of the text is printed with black-letter typeface. We used ABBYY-software[1] which offers to our experience the best solution for such typeface. To illustrate possible sources of errors we present in figure 1 an input example for the OCR system



Der zweyte Medelnitſchiar.
Der zweyte Klutſchiar.
Der zweyte Suldſchiar.
Der zweyte Zitnitſchiar.
Der zweyte Pitar.

(a)

Beſchreibung der Moldau. 333

ſerpens, *Kyncle*, canis &c. Im plurali ſetzen ſie hinten an die Wörter, die eine lebendige Sache bedeuten, den Artickel *ij*; als: *Ceij*, *Oamenij*, equi, homines: lebloſe Kreaturen aber endigen ſich im Plurali auf *ele*, als *Scaunele*, *Vaſele*, u. ſ. w. Auch haben die Moldauer zween Articulos feminini generis, *e* und *a*, als: *mujere*, *gaina*, mulier, gallina. Wörter, die ſich auf *e* endigen, haben im Plurali *ile*, als: *mujere*, *mujerile*, die ſich aber auf *a* endigen, haben im Plurali *ele*, als *gaina*, *gainele*.

3) Kan man vielleicht wahrſcheinlicher muthmaſſen, daß diejenigen Wörter, welche mehr mit der Italiäniſchen als mit der alten römiſchen Sprache überein kommen, aus dem langen Umgange, welchen die Moldauer mit den Genueſern während ihres Beſitzes der Küſten des ſchwarzen Meers hatten, ſich in unſere Sprache mit eingeſchlichen haben.

Denn auf gleiche Weiſe haben die Moldauer, nachdem ſie mit den Griechen, Türken und Pohlen zu handeln anfiengen, auch Wörter aus der Sprache dieſer Völker in die ihrige aufgenommen; zum Exempel, von den Griechen *Pedepſa*, [illegible], *Kivernisire*, [illegible], *Procopie*, [illegible], *Blaſtem*, [illegible], *azyma*, [illegible], *Drum*, [illegible], *Pizma*, [illegible]. Da wir nun alſo beyder Partheyen Meynungen vorgetragen haben, ſo getrauen wir uns nicht zu beſtimmen, welche von beyden der Wahrheit am nächſten ſey? aus Furcht,

(b)

Figure 1. Input example of for OCR

We distinguish between two error types

1. Errors related to mixed type-face which relate in fact with words in other languages. Here it is the case of words as *„mujere"* in Romanian or *„προκοπη"* in Greek (see Figure 1(a))

2. Errors related to foreign words written after German phonetic rules (see Figure 1(b)). Such words are printed with black-letter type face but the language model witch underlines the OCR-system cannot recognize them.

The error rate in case 2 is nearly 100%: Only German spelling of foreign named entities (if they are valid also nowadays, e.g. the name of the river *Pruth)* is correctly recognized.

On homogenous pages (with less foreign words) the OCR-rate was about 25%. The reason is the high number of out of current vocabulary words or morphological forms. On pages with higher percentage of foreign words the error rate raises to more than 50%. For such pages we conclude that it takes less time and effort to type-in the material.

2.2. Linguistic challenges

It is already known that human understanding as well as automatic processing of historical texts implies several linguistic challenges. In this work we focused on identifying those diachronic aspects, which have to be annotated. The purpose of annotation is twofold:

1. As support information for further automatic processing and
2. To mark-up those elements for which further explanations are necessary to make the document understandable for non-specialist readers.

Following the linguistic levels we distinguish between the following cases and corresponding linguistic annotations:

a) At the morphological and syntactical level:

Unknown named entities for any modern system: e.g. *„in dem bergigten Theile von Moramor, (*)*". Even in text the *„Moramor"* region is not clearly identified and the text passage contains two footnotes "(*)", one from Cantemir himself and one from the German translator from 1771.

Transliteration variations of named entities with respect to modern forms: e.g. „W*alachey"* (modern form *„Walachei*"), *„Dragosch*" (modern form *„Dragos*").

Old morphological forms in normal language e.g. *„zweyten"* or *„Theil"* instead of *"zweiten"* and *"Teil"*

b) At the semantic level:

There are either words which still exist in the modern vocabulary but mostly used with a different meaning. An example is the word *„flüchtigen"* used in the XVIIIth century exclusively with the meaning of *„running away from somebody*" whereas nowadays it is mostly used with the meaning of *„volatile substance*". The main challenge here is that both meanings were and are still valid through the whole period from XVIIIth century until now, just the usage frequency of one or the other meaning changed.

Time references are often relative. In an expression like *„von dem heutigen Ungarn"* (engl. *„from Hungary nowadays*") one should understand and interpret the temporal expression „nowadays" as referring to the time when the text was written (even not: published). This also implies that the corresponding political or geographical unit, in this case *„Hungary"* may have changed over time.

c) At the knowledge-level

Here we give just two examples, which are not exhaustive at all. But we consider it representative for the type of challenges encountered in processing such historical texts.

[1] http://www.abbyy.com

The first example refers to geographical units / population groups, which changed their denomination or may refer to different entities depending of the historical/geographical context. In the sentence

> *„Die auf der andern Seite angränzende Polen und Russen nennen die Moldauer Wolochen, d. i.* ***Wälsche*** *oder Italiäner, die Walachen aber, die auf dem Gebirge wohnen, heissen sie die Berg-Walachen, oder die Leute jenseits des Gebirges*“

we find the term „***Wälsche“***. In Central Germany up to the last century „*Wälsche“* was the name for Frenchmen, in Southern Germany for Italians and still today in Eastern Austria it is still the name for Slovenians. Thus the term depends on the historical and geographical context and is not fixed to one population. However, readers may be confused without this background knowledge.

The second example refers to pieces of knowledge, which for modern user are erroneous, and thus cannot be found in any modern knowledge base: Cantemir refers to a Roman emperor „*Nerva-Trajan“*. This was indeed the knowledge of XVIIIth century; nowadays historical writers differentiate between the two emperors „*Trajan“* and „*Nerva*“, the latter being the adoptive father of Trajan. Thus the reader will not find any entry „*Nerva Trajan*“ in a modern knowledge source. Here a manual annotation and explanation is necessary.

Finally the footnotes of Cantemir as well as the critical notes of the translator and further those of the editor have to be included and embedded as knowledge sources. The main problem here is that they often contradict. This is no problem for a specialist, but may be confusing for a normal reader.

d) In case of Dimitrie Cantemir works we have a fourth level of confusion related to the multilingual editions. There are available versions in German, English, Romanian and the Latin original of „Descriptio Moldaviae“. All translations do not respect the original to the same extent, but include personal interpretations and knowledge of the translators, resp.. The versions are rather comparable corpora than parallel ones in a proper sense. Thus, it is extremely important that paragraphs referring to identical events, people, and regions are aligned and displayed to the reader. Methods of extracting knowledge from comparable corpora (Smith, Toutanova) rely on identification of common units as well as similar lengths of textual units. In this case the links could be the named entities. However, automatic processing will run into problems, because named entities are differently lexicalised in each of the languages and thus their spelling differs. A very simple example is illustrative: The German edition speaks about „*Moldau*“, the Romanian one about „*Moldova*“ as in the Latin original the same geographical unit is called „*Moldavia”*.

3. A Knowledge Annotation Scheme

Given all challenges presented in Section 2 we propose an intermediate annotation level, between the TEI level, i.e. the one which marks all layout aspects, and the deep ones (e.g. the linguistic annotation). This intermediate annotation level delivers:

1. The terms or multiword- expressions, for which links with external knowledge bases and/or additional explanations, multimodal information have to be provided in order to increase the readability and

2. The possible sources of errors for further language processing tools

Our annotation-schema is XML-conform and we intend to link the attribute–values, tags and attributes with a domain ontology, which will be defined in a second stage. Thus, the annotations will have a meaning and could be used further for interlinking among and reasoning over documents.

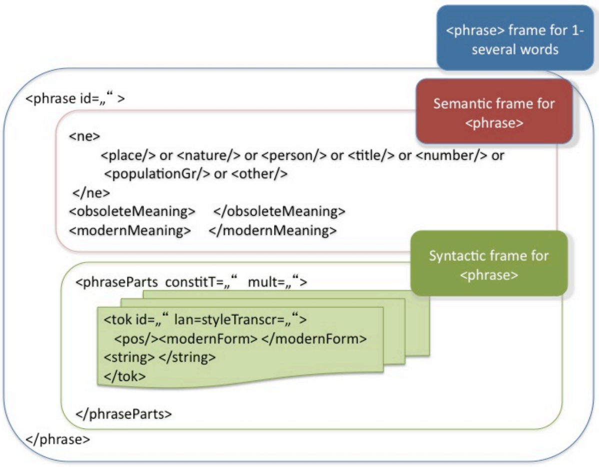
The main unit of the annotation is called „phrase“. By phrase we understand a word or a multi-word expression. For each phrase we distinguish a syntactic and a semantic frame. The syntactic frame contains the information

- whether the „phrase“ is one word or a multi-word expression

- Details about each token, namely: part-of-speech, the string as well as the modern writing form of the string, if necessary

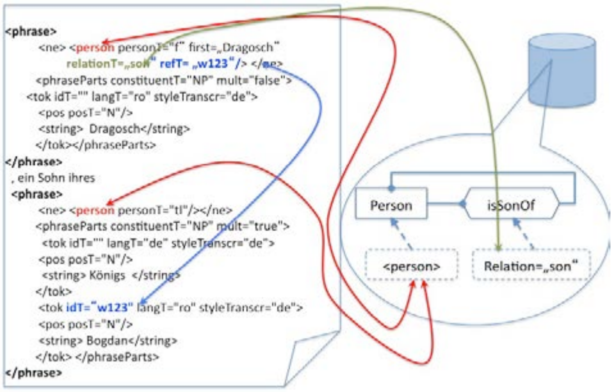
The semantic frame includes information about the name-entity (if any) as well as the obsolete meaning and the modern meaning of the word.

In figure 2 we present the structure of the annotation, while in figure 3 we present an example of an annotation as well as the possible linkage with a domain ontology


<phrase> frame for 1-several words
Semantic frame for <phrase>
Syntactic frame for <phrase>

```
<phrase id=„“ >
    <ne>
        <place/> or <nature/> or <person/> or <title/> or <number/> or
        <populationGr/> or <other/>
    </ne>
    <obsoleteMeaning>   </obsoleteMeaning>
    <modernMeaning>   </modernMeaning>
    <phraseParts  constitT=„“   mult=„“>
        <tok id=„“  lan=styleTranscr=„“>
          <pos/><modernForm> </modernForm>
        <string> </string>
        </tok>
    </phraseParts>
</phrase>
```

Figure 2. Structure of the annotation scheme



```
<phrase>
      <ne> <person personT="f" first=„Dragosch"
            relationT=„son" refT= „w123"/> </ne>
     <phraseParts constituentT="NP" mult="false">
   <tok idT="" langT="ro" styleTranscr="de">
      <pos posT="N"/>
      <string> Dragosch</string>
     </tok></phraseParts>
</phrase>
 , ein Sohn ihres
 <phrase>
      <ne> <person personT="tl"/></ne>
     <phraseParts constituentT="NP" mult="true">
       <tok idT="" langT="de" styleTranscr="de">
      <pos posT="N"/>
       <string> Königs  </string>
     </tok>
      <tok idT="w123" langT="ro" styleTranscr="de">
      <pos posT="N"/>
      <string> Bogdan</string>
     </tok> </phraseParts>
</phrase>
```

Figure 3. Example of annotation scheme and ontology linking

4. Conclusions and further work

In this paper we presented a first analysis attempt for multilingual historical documents, that means documents which do not include only Latin and/or ancient Greek words / paragraphs, but also a reasonable amount of words in another language than its proper language. We argue that such documents can be made available for a broader public and even for specialists from domains related to history (like ethnology, history of music) only by means of additional knowledge, which is linked with the document. We discuss the challenges for processing such texts and introduce an annotation scheme, which may serve as base for such enrichment. We argue that the annotation can be done only in a semi-automated way.

For the moment we are in the process of annotating a first set of several hundred pages which will serve as control test set for further automatic processes. Annotation of similar data (identical tokens or lemmas) will be performed also automatically for further texts.

5. References

Cantemir, D,1771, *Beschreibung der Moldau*. Faksimildruck der Original Ausgabe von 1771, Maciuca C. (Ed). , Bukarest, Kriterion Verlag, 1973

Smith, J.R. and Toutanova, K. , *Extracting parallel sentences from Corpora using Document Level Alignement* , http://research.microsoft.com/pubs/140708/n10-1063.pdf