Language Technology for Normalisation of Less-Resourced Languages
SALTMIL 8 - AFLAT 2012

Workshop Programme

Tuesday, May 22, 2012

09:15–09:30	**Welcome / Opening Session**

09:30–10:30	**Invited Talk** - *How to build language technology resources for the next 100 years* Sjur Moshagen Nørstebø, Sámi Parliament

10:30–11:00	**Coffee Break**

11:00–13:00	**Resource Creation**
11:00–11:30	*Issues in Designing a Spoken Corpus of Irish* Elaine Uí Dhonnchadha, Alessio Frenda and Brian Vaughan
11:30–12:00	*Learning Morphological Rules for Amharic Verbs Using Inductive Logic Programming* Wondwossen Mulugeta and Michael Gasser
12:00–12:30	*The Database of Modern Icelandic Inflection* Kristín Bjarnadottir
12:30–13:00	*Natural Language Processing for Amazigh Language: Challenges and Future Directions* Fadoua Ataa Allah and Siham Boulaknadel

13:00–14:00	**Lunch Break**

Tuesday, May 22, 2012 (continued)

14:00–16:00	**Resource Use**
14:00–14:30	*Compiling Apertium morphological dictionaries with HFST and using them in HFST applications* Tommi A. Pirinen and Francis M. Tyers
14:30–15:00	*Automatic structuring and correction suggestion system for Hungarian clinical records* Borbála Siklósi, György Orosz, Attila Novák and Gábor Prószéky
15:00–15:30	*Constraint Grammar based Correction of Grammatical Errors for North Sámi* Linda Wiechetek
15:30–16:00	*Toward a Rule-Based System for English-Amharic Translation* Michael Gasser

16:00–16:30	**Coffee Break**

16:30–17:30	**Poster Session**

- *Technological Tools for Dictionary and Corpora Building for Minority Languages: Example of the French-based Creoles* – Paola Carrion Gonzalez and Emmanuel Cartier
- *Describing Morphologically-rich Languages using Metagrammars: a Look at Verbs in Ikota* – Denys Duchier, Brunelle Magnana Ekoukou, Yannick Parmentier, Simon Petitjean and Emannuel Schang
- *A Corpus of Santome* – Tjerk Hagemeijer, Iris Hendrickx, Abigail Tiny and Haldane Amaro
- *The Tagged Icelandic Corpus (MM)* – Sigrún Helgadóttir, Ásta Svavarsdóttir, Eiríkur Rögnvaldsson, Kristín Bjarnadóttir and Hrafn Loftsson
- *Semi-automated extraction of morphological grammars for Nguni with special reference to Southern Ndebele* – Laurette Pretorius and Sonja Bosch
- *Tagging and Verifying an Amharic News Corpus* – Björn Gambäck
- *Resource-Light Bantu Part-of-Speech Tagging* – Guy De Pauw, Gilles-Maurice de Schryver and Janneke van de Loo
- *POS Annotated 50M Corpus of Tajik Language* – Gulshan Dovudov, Vít Suchomel and Pavel Šmerk

17:30–17:45	**Closing Session**

Editors

Guy De Pauw	University of Antwerp
Gilles-Maurice de Schryver	Ghent University
Mikel L. Forcada	Universitat d'Alacant
Kepa Sarasola	University of the Basque Country
Francis M. Tyers	Universitat d'Alacant
Peter Waiganjo Wagacha	University of Nairobi

Organizing Committee

Mikel L. Forcada (SALTMIL)
Departament de Llenguatges i Sistemes Informàtics, Universitat d'Alacant

Guy De Pauw (AfLaT)
CLiPS - Computational Linguistics Group, University of Antwerp

Gilles-Maurice de Schryver (AfLaT)
African Languages and Cultures, Ghent University
Xhosa Department, University of the Western Cape

Kepa Sarasola (SALTMIL)
Dept. of Computer Languages, University of the Basque Country

Francis M. Tyers (SALTMIL)
Departament de Llenguatges i Sistemes Informtics, Universitat d'Alacant

Peter Waiganjo Wagacha (AfLaT)
School of Computing & Informatics, University of Nairobi

Workshop Programme Committee

Iñaki Alegria - University of the Basque Country, Spain
Núria Bel - Universitat Pompeu Fabra, Barcelona, Spain
Lars Borin - Göteborgs universitet, Sweden
Sonja Bosch - University of South Africa, South Africa
Khalid Choukri - ELRA/ELDA, France
Guy De Pauw - Universiteit Antwerpen, Belgium
Gilles-Maurice de Schryver - Universiteit Gent
Mikel L. Forcada - Universitat d'Alacant, Spain
Dafydd Gibbon - Universität Bielefeld, Germany
Lori Levin - Carnegie Mellon University, USA
Hrafn Loftsson - University of Reykjavik, Iceland
Girish Nath Jha - Jawaharlal Nehru University, India
Ọdẹ́túnjí Ọdẹ́jọbí - Obafemi Awolowo University, Nigeria
Laurette Pretorius - University of South Africa, South Africa
Benoît Sagot - INRIA, France
Felipe Sánchez-Martínez - Universitat d'Alacant, Spain
Kepa Sarasola - University of the Basque Country, Spain
Kevin Scannell - Saint Louis University, United States
Trond Trosterud - University of Tromsø, Norway
Francis M. Tyers - Universitat d'Alacant, Spain
Peter Waiganjo Wagacha - University of Nairobi, Kenya

All contributions found in the present proceedings were peer-reviewed by at least two members of the programme committee.

Table of Contents

Author Index

Preface

The 8th International Workshop of the ISCA Special Interest Group on Speech and Language Technology for Minority Languages (SALTMIL)[1] and the Fourth Workshop on African Language Technology (AfLaT2012)[2] is held as a joint effort as part of the 2012 International Language Resources and Evaluation Conference (LREC 2012). Entitled "*Language technology for normalisation of less-resourced languages*", the workshop is intended to continue the series of SALTMIL/LREC workshops on computational language resources for minority languages, held in Granada (1998), Athens (2000), Las Palmas de Gran Canaria (2002), Lisbon (2004), Genoa (2006), Marrakech (2008) and Malta (2010), and the series of AfLaT workshops, held in Athens (EACL2009), Malta (LREC2010) and Addis Ababa (AGIS11).

The Istanbul 2012 workshop aims to share information on tools and best practices, so that isolated researchers will not need to start from scratch. An important aspect will be the forming of personal contacts, which can minimize duplication of effort. There will be a balance between presentations of existing language resources, and more general presentations designed to give background information needed by all researchers.

While less-resourced languages and minority languages often struggle to find their place in a digital world dominated by only a handful of commercially interesting languages, a growing number of researchers are working on alleviating this linguistic digital divide, through localisation efforts, the development of BLARKs (basic language resource kits) and practical applications of human language technologies. The joint SALTMIL/AfLaT workshop on "Language technology for normalisation of less-resourced languages" provides a unique opportunity to connect these researchers and set up a common forum to meet and share the latest developments in the field.

The workshop takes an inclusive approach to the word normalisation, considering it to include both technologies that help make languages more "normal" in society and everyday life, as well as technologies that normalise languages, i.e. help create or maintain a written standard or support diversity in standards. We particularly focus on the challenges less-resourced and minority languages face in the digital world.

The Workshop Organizers
Mikel L. Forcada
Guy De Pauw
Gilles-Maurice de Schryver
Kepa Sarasola
Francis M. Tyers
Peter Waiganjo Wagacha

[1] http://ixa2.si.ehu.es/saltmil

[2] http://AfLaT.org

Issues in Designing a Corpus of Spoken Irish

Elaine Uí Dhonnchadha, Alessio Frenda, Brian Vaughan
Centre for Language and Communication Studies,
Trinity College Dublin, Ireland.
E-mail: {uidhonne; frendaa; bvaughan}@tcd.ie

Abstract

This paper describes the stages involved in implementing a corpus of spoken Irish. This pilot project (consisting of approximately 140K words of transcribed data) implements part of the design of a larger corpus of spoken Irish which it is hoped will contain approximately 2 million words when complete. It hoped that such a corpus will provide material for linguistic research, lexicography, the teaching of Irish and for development of language technology for the Irish language.

Keywords: spoken language, corpus design, Irish

1. Introduction

This paper describes the design of a corpus of spoken Irish. The proposed spoken corpus, consisting of approximately 2 million words will provide material for linguistic research, lexicography, the teaching of Irish and for development of language technology for the Irish language. Also described are various stages involved in the pilot implementation of part of the proposed corpus.

In order to create a comprehensive corpus of spoken Irish, the design includes dialectal and chronological variation, as well as different registers and contexts of language use. In addition to new recordings, material will also be drawn from existing collections and archives (radio and TV broadcast and folklore archives). The corpus is in line with current standards in terms of time-alignment of transcripts, XML formatting and part-of-speech tagging for electronic searchability and querying. It will also be available online.

2. Linguistic Background

Irish is the first official language of Ireland with English being the second official language. In practice Irish is spoken as a first language in only a small number of areas known as *Gaeltachtaí* which are mainly on the western seaboard. For the remainder of the population Irish is learned at school (compulsorily) as a second language. While 1.6 million[1] of the 3.9 million population report proficiency in the spoken language, the number of native speakers is much lower, at 64 thousand[2] and dwindling in the *Gaeltachtaí* (although numbers are increasing in urban areas). These sociolinguistic conditions mean that a comprehensive spoken corpus has a vital role to play in promoting and preserving the spoken language.

3. Corpus Design

In order to design a corpus that is representative and authoritative, it is useful to take into account the design adopted by recent, state-of-the art corpora for other languages. We examined the design of a number of corpora (London-Lund Corpus of Spoken English[3], Lancaster/IBM Spoken English Corpus (SEC)[4], Corpus of Spoken New Zealand English[5] British National Corpus[6], COREC (Corpus oral de referencia del Español Contemporáneo)[7], CLIPS (Corpora e Lessici dell'Italiano Parlato e Scritto)[8], ICE (The International Corpus of English)[9] and CGN (Corpus Gesproken Nederlands)[10]
One common feature shared by the more recent corpora surveyed here is the extent of naturalistic conversational material they include.
There is no existing corpus of spoken Irish which meets our criteria of including dialectal and chronological variation. The most substantial collection of spoken language transcripts, *Caint Chonamara* (Wigger, 2000) (1.2 million words approx.) relates to one dialect only (Conamara) and one year, 1964, and is not linguistically annotated.

Our design considers the following variables:

- time frame: we aim to create a diachronic corpus by including spoken Irish from the earliest available recordings to the present day. We have decided upon

[1] Census 2006 http://www.cso.ie/en/newsandevents/pressreleases/2007pressreleases/2006censusofpopulation-volume9-irishlanguage/

[2] Census 2006 http://census.cso.ie/Census/TableViewer/tableView.aspx?ReportId=96447

[3] London-Lund Corpus of Spoken English http://khnt.hit.uib.no/icame/manuals/LONDLUND/INDEX.HTM

[4] Lancaster/IBM Spoken English Corpus (SEC) http://khnt.hit.uib.no/icame/manuals/sec/INDEX.HTM.

[5] Corpus of Spoken New Zealand English http://icame.uib.no/wsc/index.htm.

[6] British National Corpus http://www.natcorp.ox.ac.uk/corpus/index.xml.

[7] A reference corpus for contemporary spoken Spanish http://www.lllf.uam.es/~fmarcos/informes/corpus/corpulee.html.

[8] Corpora and Lexica for Spoken and Written Italian http://www.clips.unina.it/it/.

[9] International Corpus of English http://ice-corpora.net/ice/.

[10] Spoken Dutch Corpus http://lands.let.kun.nl/cgn/ehome.htm.

the three time periods, P1: 1930-1971, P2: 1972-1995 and P3 1996-present. In our pilot corpus we concentrated on contemporary speech (P3 1996-present).

- dialectal variation: we aim to cover the three main dialects of Irish in equal measure: i.e. not proportionally to the number of speakers of each dialect, given that the corpus is diachronic and the relative proportions might have varied over the years, but to provide equal documentation of each dialect insofar as possible.
- sociolinguistic variation: we aim to include Irish speakers from all linguistic backgrounds (a) 'traditional' native speakers, (b) non-native speakers and (c) 'non-traditional' native speakers, i.e. those who describe themselves as native speakers having being raised through Irish by L1 or L2 parents, typically in a non-*Gaeltacht* setting, and who have subsequently attended Irish-medium schools (Ó Giollagáin & Mac Donnacha, 2008, p. 111f.).
- gender and age: we aim to represent both males and females proportionally, and to include a spread of ages (e.g. young adults, middle aged and elderly).
- context and subject matter: we aim to include conversations recorded in a variety of contexts (home, work, leisure, education etc.) and cover a variety of topics.

The corpus design for Period 3 (1996-present), which is inspired by the ICE and CGN corpus designs, proposes 70% dialogic speech (420 X 8-10 min recordings) and 30% monologic speech (180 X 8-10 min recordings). Each of the 600 recording transcriptions will contain 1500-2000 words giving a corpus of between 900,000 and 1,200,000 words. The dialogic speech is further categorised into 'private' (e.g. face-to-face conversations, phonecalls and interviews) and public (e.g. broadcast discussions and interviews, parliamentary debates, classroom lessons, business meetings etc.) speech. The monologic speech is categorised as either scripted (news broadcasts, speeches etc.) or unscripted (e.g. sports commentaries, unscripted speeches, demonstrations, legal presentations etc.) speech.

For Period 1 (1930-1971) and Period 2 (1972-1995), not all of the required types of material will be available. We will aim to keep the same proportions as for Period 3 but the quantities will necessarily be less. In order to ensure that as many of the design categories are repesented as fully as is possible, a thorough investigation of available archival material will have to be undertaken.

4. Data Collection and Recording

In the case of dialogic speech (70%) there is ample public broadcast material available in the form of radio podcasts and archives. Other categories such as classroom lessons and business meetings will have to be recorded. All private dialogue speech will have to be recorded in the various dialectal regions. In the case of monologic speech (30%), the majority of this can be sourced from broadcast media and archives, with some categories such as legal presentations being recorded.

Funding was obtained from Foras na Gaeilge (the cross-border body responsible for promoting the Irish language on the island of Ireland) to carry out a pilot study. We decided to concentrate our initial efforts in the contemporary period (P3) and on dialogic speech. As time and resources were limited we used readily available public broadcast dialogues (radio interviews and discussions).

We also carried out a small amount of video recording of private dialogue conversations. Four pairs of volunteers agreed to be video recorded in informal conversation in the Speech Communications Laboratory[11], TCD. The interactions were video recorded using a Sony HDR-XR500v High Definition Handycam. The audio was recorded in two ways: 1) using the onboard camera microphone and 2) using two Sennheiser MKH-60 shotgun microphones and an Edirol 4-channel HD Audio recorder. Audio was recorded at a sampling rate of 96KhZ with a bit rate of 24 bits. For practical purposes, the audio was bounced down to a sampling rate of 44.1KhZ with a bit rate of 16bits (the Redbook audio standard), with the higher 96KhZ files being used for archiving.

In total, 70 x 8 min. recording extracts were transcribed giving 102,000 words of transcribed speech. By also aligning and formatting some existing transcripts[12], the overall total is currently 140,000 words (approximately).

5. Transcription

Spoken and written language differ in a number of important respects. The syntactic structure of spontaneous spoken utterances is usually simpler, but any faithful transcript of a spoken conversation will not look as orderly as a written dialogue. It is natural in spontaneous speech to produce repetitions, make false starts, to hesitate or simply to leave part of a message unfinished, relying instead on non-verbal communication such as a gesture or the tone of voice. This together with dialectal pronunciations which deviate substantially from standard orthographical representations means that transcribing spoken language presents immediate challenges.

5.1 Guidelines

We examined a number of transcription conventions already in use including CHAT[13], LINDSEI[14], and LDC[15]. The CHAT (Codes for the Human Analysis of Transcripts) System is a comprehensive standard for transcribing and encoding the characteristics of spoken language (MacWhinney, 2000). These guidelines were developed for the transcription of spoken interactions between children and their carers in order to study child language

[11] http://www.tcd.ie/slscs/clcs/scl/

[12] Frenda (2011) material transcribed for PhD research TCD (20K); Wigger (2000) Caint Chonamara (10K); Dillon, G., material transcribed for PhD research TCD (5K).

[13] CHAT http://childes.psy.cmu.edu/manuals/chat.pdf

[14] Louvain International Database of Spoken English Interlanguage Transcription guidelines http://www.uclouvain.be/en-307849.html

[15] Linguistic Data Consortium http://www.ldc.upenn.edu /Creating/creating_annotated.shtml#Transcription

acquisition. They give detailed guidelines for marking up such phenomena as inaudible segments, phonetic fragments, repetitions, overlaps, interruptions, trailing off, foreign words, proper nouns and numbers etc. While the guidelines are very comprehensive there are a few drawback to implementing the guidelines in full; it can slow down the transcription process considerably; some are quite subjective (short, medium and long pauses) while others are difficult to implement (retracings and reformulations).

At the other end of the scale, the LDC guidelines advocate simplicity. The philosophy here is to keep the rules to a minimum in order to make transcription as easy as possible for the transcriber, which increases transcription speed, accuracy and consistency. In addition automatic procedures are used when possible.

We also consulted researchers in other universities and research institutes[16] who have worked on the transcription of spoken Irish and obtained advice and good-practice guidelines with regard to the orthographic rendition of dialect-specific features of spoken Irish.

From our experience, it takes on average 30 minutes to orthographically transcribe 1 minute of audio material. Considering that transcription is a slow and painstaking process we believe that in order to achieve a sufficient quantity of accurately transcribed material, the transcription process must be as straightforward and intuitive as possible. This means that codes should be kept to a minimum and those codes which are necessary should use a minimum number of keystrokes.

Some aspects of speech do not need to be recorded in the transcription as they can be automatically generated at a later stage, e.g. the length of pauses. The standard orthography is morphologically transparent, i.e. it shows the internal structure of a word, which is a distinct advantage for the automatic treatment of the text, e.g. for part-of-speech tagging and the generation of a broad phonemic transcription (Ó Raghallaigh, 2010, p. 76).

We have chosen to use standard orthographic representation, for which *Caighdeán Oifigiúil* (1979) and *Foclóir Gaeilge-Béarla* (Ó Dónaill, 1977) are taken as references, and to avoid invented and ad hoc spellings at all times. There are a number of advantages to using standard orthography:

- It makes the job of transcription easier and quicker for transcribers
- It helps mimimise spelling inconsistencies among transcribers as only standard spelling is used, apart from a predefined lists permitted exceptions
- Attempting to represent actual pronunciation in orthography is difficult and prone to inconsistency. It requires specialist knowledge and can be more accurately captured in a separate phonetic transcription layer (which may be partially generated from the orthography).
- Standard orthography facilitates corpus querying and lexical searches
- Standard orthography facilitates automatic text processing, such as part-of-speech tagging and parsing
- Transcription codes for some linguistic features (e.g. co-articulation effects, ellision etc.) would require specialist training for transcribers, in order to ensure accuracy and consistency, and are better undertaken as a separate task.

Based on the above principles a set of guidelines for the transcription of Irish was developed which is available online on the project website.[17] In addition to these general guidelines are lists of prescribed spellings for filled pauses, contracted wordforms, multi-word fixed phrases (including several English fixed phrases, e.g. you know, so, just etc.) and some common dialectal forms not included in the reference dictionary (Ó Dónaill, 1977).

5.2 Software

Creating a corpus of spoken language requires transcribing audio or video recordings (spoken conversations, interviews, speeches etc). These transcriptions should ideally be time-aligned with the speech signal. There are a variety of freely available software packages to carry out this task and to aid the transcription process in general.

We tested several pieces of freely-available transcription and annotation software (e.g. Praat, ELAN, Anvil, CLAN, Xtrans, Transcriber) and chose *Transcriber*[18] as the most suitable software for the orthographic transcription of audio speech at this stage of the project, for the following reasons:

- It has a straightforward user interface which means transcribers can become proficient users is a short amount of time;
- It facilitates alignment of the audio and text transcription in XML format;
- It provides audio duration and word count information at a glance;
- Transcripts can be conveniently exported as text;
- It handles a variety of audio file types, including wav, mp3 (podcasts) and ogg which were used in this project.
- The later version of the software (TranscriberAG) can handle video as well as audio;
- It facilitates the annotation of various features of spontaneous speech (overlap, interruptions, coughs, laughs, etc.) as well as linguistics categories (e.g. proper nouns, human/animate etc. etc.) if desired.
- It can be used with foot pedals for increased speed if necessary;

This decision will be kept under review in future phases as new and improved software regularly becomes available,

[16] Pauline Welby, Laboratoire Parole et Langage CNRS - Aix-Marseille Université (personal communication); Brian Ó Raghallaigh, Fiontar, Dublin City University (p.c.) ; Eoghan Ó Raghallaigh (Doegen Project, http://dho.ie/doegen/); McKenna, M. (2005).

[17] GaLa Project http://www.tcd.ie/slscs/assets/documents/research/gala/Treoirlinte_agus_Transcriber.pdf

[18] http://trans.sourceforge.net/

and project requirements may change.

5.3 Transcribers

As there were no available experienced transcribers of Irish, it was necessary to recruit and train transcribers in the use of the transcription software and transcription guidelines.

Notices were posted in both Irish and Linguistics departments of universities around Ireland and a good response was received. We required a panel of transcribers covering the various dialects, therefore applicants were asked to nominate their preferred dialect and their second choice (if any). A dialect-specific test workpackage (consisting of a one minute audio file and transcription guidelines) was sent to all suitable applicants. Based on the results of the test piece, a panel of twenty-two transcribers was established.

We organised a transcription workshop which was attended by a number of the transcribers, together with interested parties from Foras na Gaeilge, the Royal Irish Academy as well as post graduate researchers. This proved to be very beneficial to all and discussions about transcriptions issues lead to modifications in the transcription guidelines.

Audio segments of 8 min. in duration containing broadcast discussions and interviews were selected mainly from *Raidio na Gaeltachta* podcasts. Workpackages were sent via e-mail to members of the panel of transcribers who worked from home. They returned a time-aligned transcription and timesheet for each workpackage completed.

5.4 Checking and Anonymising

Each transcript was checked for accuracy against the audio file by a member of the project team. In the case of new video-recordings, the transcripts were also anonymised, i.e. names and places which could identify the participants were replaced by fictitious names to ensure anonynity. Anonymising is not carried out for existing recordings which are available on the internet as podcasts or which have been broadcast on radio or TV.

6. Corpus Processing

6.1 Corpus Metadata

All relevant details related to speakers, transcripts and transcribers are recorded in a database. Each speaker is given a speaker code which is used in the transcript in place of the speaker's name, in order to make speakers less recognisable. Speaker attributes such as dialect, language acquisition type, i.e. whether native Gaeltacht speaker (L1 Gaeltacht), native non-Gaeltacht speaker (L1 non-Gaeltacht) or a non-native speaker (L2), gender and age, etc are recorded where known.

This data is used to generate XML corpus headers, and to facilitate onging monitoring of word counts of the various corpus design categories.

6.2 Corpus Encoding Standards

For each transcript, the output of the Transcriber software was transformed into TEI compliant XCES (XML Corpus Encoding Standard) format using a Perl script and data from the corpus database. The script also computed word counts per speaker which were fed back into the database.

All of the transcripts to date are conversations or interviews involving at least two participants. It is quite common, particularly in radio interviews, for spoken interactions to take place between speakers with different dialects or between native and non-native speakers. As we would like to be able to create sub-corpora on the basis of dialect, native/non-native status, speaker, age, gender etc. then these features must be recorded at the level of speaker-turn rather than for the transcript as a whole.

Therefore, as well as having a detailed transcript header which includes time of recording and source of audio/video file etc. we also include speaker attributes on the <speaker_turn> tag, as shown in Figure 1.

```
<doc id = "irbs0012" title =
"Barrscéalta 08 October 2010" period
= "1996-pres" medium =
"broadcast-radio"spokentype =
"interview" text_source = "GALA-TCD"
av_source = "RnaG podcast">

<speaker_turn id = "200" code =
"RNG_ANC" dialect = "Ulaidh" gender =
"Bain" actype = "L1 Gaeltacht" year =
"2010">
caidé méid airgid a chosnódh sé na bádaí
seo a thabhairt suas chun dáta agus
cloígh lena rialacha úra atá tagtha
isteach?
</speaker_turn>

<speaker_turn id = "559" code =
"RNG_LCI" dialect = "Mumhan" gender =
"Fir" actype = "L1 Gaeltacht?" year =
"2010" >
Bhuel ehm braitheann sé sin ar
chaighdeán an bháid, abair, agus níl
aon dabht faoi ach go bhfuil sé
costasach, abair, [tá tá] tá tuairiscí
faighte agamsa ar daoine go raibh orthu
eh [céad míle ar] céad míle euro a
chaitheamh eh ag tabhairt a mbád suas
chun caighdeáin. …
</speaker_turn>
```

Figure 1 Fragment XCES formatted spoken transcript

6.3 Part-of-speech Tagging

The XML transcripts have been part-of-speech tagged. Additional codes and lexical items were added to the finite-state tokenizer and morphological analyser (Uí Dhonnchadha, 2006) to handle some features specific to spoken language such filled pauses (em, eheh etc,) fixed phrases (*an dtuigeann tú* 'do you understand', *mar a déarfá* 'as you say' etc.), as well as codes for non-verbal events (coughs, laughs, sneezes etc.), phonetic fragments (*b- b- bosca* 'b- b- box') and indecipherable material (xxx). Dialectal varriants (Ó Dónaill, 1977) e.g. *gleamaigh* 'lobster', *aoinne* 'anyone' etc. proved useful as these forms are perhaps more common in spoken language than written language.

Spoken transcripts contain more English words than would be found in written Irish, therefore a list of English vocabulary items would be useful addition to the morphological analyser, but this was not carried out in the current phase project. Detailed analysis of the accuracy of the POS tagging on spoken language as compared to accuracy on written language also has not yet been carried out.

6.4 SketchEngine Corpus Query Engine

All POS tagged transcripts have been converted to vertical format and loaded into the SketchEngine [19] Corpus Query System. For each transcript the following information is available: document id, title, time period, text_source (source of transcription) and av_source (source of audio/video file). For each speaker turn the following information is available: speaker code, dialect, actype (language acquisition type), gender, year of recording. Sub-corpora can be created by selecting particular values for any selection of the above variables, i.e. dialect = Ulster, actype=L1, etc.

7. Conclusion

In this paper we have outlined the issues involved in designing a spoken corpus, including data collection and transcription and initial stages of corpus processing. Through implementing a pilot corpus, we believe that we have overcome most difficulties likely to be encountered in a fullscale project, and are in a position to make infomed decisions about costings and timings of a larger scale project.

8. Future Work

The main tasks for the future, are to collect additional data particularly through the recording of spontaneous conversations from volunteers in various *Gaeltacht* locations around the country, and also to improve the part-of-speech tools to better handle the particular characteristics of spoken langauage. Quality control measures would also need to be put in place to ensure the quality and consistency of future transcriptions.

9. Acknowledgements

We would like to thank Foras na Gaeilge for funding this phase of the project.

10. References

Caighdeán Oifigiúil, An. (1979[1958]). *Gramadach na Gaeilge agus litriú na Gaeilge: An caighdeán oifigiúil.* Baile Átha Cliath: Oifig an tSoláthair.

Frenda, A. (2011). *Gender in Insular Celtic: A functionalist account of variation and change in Irish and Welsh.* PhD thesis, Trinity College Dublin.

Ó Curnáin, B. (2007). *The Irish of Iorras Aithneach County Galway*. Dublin: Dublin Institute for Advanced Studies.

Ó Dónaill, N. (1977). *Foclóir Gaeilge-Béarla*. Baile Átha Cliath: Roinn Oideachais agus Eoilaíochta.Ó Giollagáin, C. & S. Mac Donnacha (2008). The Gaeltacht today. In C. Nic Pháidín and S. Ó Cearnaigh (Eds.), A new view of the Irish language, pp. 108–120. Dublin: Cois Life.

Ó Raghallaigh, B. (2010). *Multi-dialect phonetisation for Irish text-to-speech synthesis: A modular approach.* PhD thesis. Trinity College Dublin.

MacWhinney, B. (2011). *The CHILDES Project: Tools for Analyzing Talk.* Electronic Edition. Available online at http://childes.psy.cmu.edu/manuals/chat.pdf.

McKenna, M. (2005). *Seanchas Rann na Feirste: is fann guth an éin a labhras leis féin*, pp.169-180. Dublin: Coiscéim.

Uí Dhonnchadha, E. and van Genabith, J. (2006). Scaling an Irish FST morphology engine for use on unrestricted text, In: Yli-Jyrä, A., Karttunen, L., Karhumäki, J. (Eds.). *Finite-State Methods in Natural Language Processing* (Book Series: Lecture Notes in Artificial Intelligence), Springer-Verlag, pp. 247 – 258.

Wigger, A. (Ed.) (2000). *Caint Chonamara: Bailiúchán Hans Hartmann. Imleabhar IX. Ros Muc*. Universität Bonn.

[19] http://the.sketchengine.co.uk/

Learning Morphological Rules for Amharic Verbs Using Inductive Logic Programming

Wondwossen Mulugeta[1] and Michael Gasser[2]
[1]Addis Ababa University, Addis Ababa, Ethiopia
[2]Indiana University, Bloomington, USA
E-mail: [1]wondgewe@indiana.edu, [2]gasser@cs.indiana.edu

Abstract

This paper presents a supervised machine learning approach to morphological analysis of Amharic verbs. We use Inductive Logic Programming (ILP), implemented in CLOG. CLOG learns rules as a first order predicate decision list. Amharic, an under-resourced African language, has very complex inflectional and derivational verb morphology, with four and five possible prefixes and suffixes respectively. While the affixes are used to show various grammatical features, this paper addresses only subject prefixes and suffixes. The training data used to learn the morphological rules are manually prepared according to the structure of the background predicates used for the learning process. The training resulted in 108 stem extraction and 19 root template extraction rules from the examples provided. After combining the various rules generated, the program has been tested using a test set containing 1,784 Amharic verbs. An accuracy of 86.99% has been achieved, encouraging further application of the method for complex Amharic verbs and other parts of speech.

1. Introduction

Amharic is a Semitic language, related to Hebrew, Arabic, and Syriac. Next to Arabic, it is the second most spoken Semitic language with around 27 million speakers (Sieber, 2005; Gasser, 2011). As the working language of the Ethiopian Federal Government and some regional governments in Ethiopia, most documents in the country are produced in Amharic. There is also an enormous production of electronic and online accessible Amharic documents.

One of the fundamental computational tasks for a language is analysis of its morphology, where the goal is to derive the root and grammatical properties of a word based on its internal structure. Morphological analysis, especially for complex languages like Amharic, is vital for development and application of many practical natural language processing systems such as machine-readable dictionaries, machine translation, information retrieval, spell-checkers, and speech recognition.

While various approaches have been used for other languages, Amharic morphology has so far been attempted using only rule-based methods. In this paper, we applied machine learning to the task.

2. Amharic Verb Morphology

The different parts of speech and their formation along with the interrelationships which constitute the morphology of Amharic words have been more or less thoroughly studied by linguists (Sieber, 2005; Dwawkins, 1960; Bender, 1968). In addition to lexical information, the morphemes in an Amharic verb convey subject and object person, number, and gender; tense, aspect, and mood; various derivational categories such as passive, causative, and reciprocal; polarity (affirmative/negative); relativization; and a range of prepositions and conjunctions.

For Amharic, like most other languages, verbs have the most complex morphology. In addition to the affixation, reduplication, and compounding common to other languages, in Amharic, as in other Semitic languages, verb stems consist of a root + vowels + template merger (e.g., *sbr* + ee + CVCVC, which leads to the stem *seber* ሰበር[1] 'broke') (Yimam, 1995; Bender, 1968). This non-concatenative process makes morphological analysis more complex than in languages whose morphology is characterized by simple affixation. The affixes also contribute to the complexity. Verbs can take up to four prefixes and up to five suffixes, and the affixes have an intricate set of co-occurrence rules.

For Amharic verbs, grammatical features are not only shown using the affixes. The intercalation pattern of the consonants and the vowels that make up the verb stem will also be used to determine various grammatical features of the word. For example, the following two words have the same prefixes and suffixes and the same root while the pattern in which the consonants and the vowels intercalated is different, resulting in different grammatical information.

*?-**sebr**-alehu→ 1s pers.sing. simplex imperfective*
*?-**seber**-alehu→ 1st pers.sing.passive imperfective*

Figure 1: Stem template variation example

In this second case, the difference in grammatical feature is due to the affixes rather than the internal root template structure of the word.

*te-**deres**-ku → 1st pers. sing. passive perfective*
***deres**-ku → 1st pers. sing. simplex perfective*

Figure 2: Affix variation example

[1] *Amharic is written in the Geez writing system. For our morphology learning system we romanize Amharic orthography, and we cite these romanized forms in this paper.*

As in many other languages, Amharic morphology is also characterized by alternation rules governing the form that morphemes take in particular environments. The alternation can happen either at the stem affix intersection points or within the stem itself. Suffix-based alternation is seen, for example, in the second person singular feminine imperfect and imperative, shown in Table 1. The first two examples in Table 1 shows that, the second person singular feminine imperative marker ***'i'***, if preceded by the character ***'l'***, is altered to ***'y'***. The last two examples show that the same alternation rule applies for imperfect roots.

No.	Word	Root	Feature
1	**gdel**	gdl	2nd person sing. masc. imperative
2	**gdey (gdel-i)**	gdl	2nd person sing. fem. imperative
3	**t-gedl-aleh**	gdl	2nd person sing. masc. imperfect
4	**t-gedy-alex**	gdl	2nd person sing. fem. imperfect

Table 1: Example of Amharic Alternation Rule

3. Machine Learning of Morphology

Since Koskenniemi's (1983) ground-breaking work on two-level morphology, there has been a great deal of progress in finite-state techniques for encoding morphological rules (Beesley & Karttunen, 2003). However, creating rules by hand is an arduous and time-consuming task, especially for a complex language like Amharic. Furthermore, a knowledge-based system is difficult to debug, modify, or adapt to other similar languages. Our experience with HornMorpho (Gasser, 2011), a rule-based morphological analyser and generator for Amharic, Oromo, and Tigrinya, confirms this. For these reasons, there is considerable interest in robust machine learning approaches to morphology, which extract linguistic knowledge automatically from an annotated or un-annotated corpus. Our work belongs to this category.

Morphology learning systems may be unsupervised (Goldsmith, 2001; Hammarström & Borin, 2011; De Pauw & Wagacha, 2007) or supervised (Oflazer *et al* 2001; Kazakov, 2000). Unsupervised systems are trained on unprocessed word forms and have the obvious advantage of not requiring segmented data. On the other hand, supervised approaches have important advantages of their own where they are less dependent on large corpora, requires less human effort, relatively fast which makes it scalable to other languages and that all rules in the language need not be enumerated.

Supervised morphology learning systems are usually based on two-level morphology. These approaches differ in the level of supervision they use to capture the rules. A weakly supervised approach uses word pairs as input (Manandhar *et al*, 1998; Mooney & Califf, 1995; Zdravkova *et al*, 2005). Other systems may require segmentation of input words or an analysis in the form of a stem or root and a set of grammatical morphemes.

4. ILP and Morphology Learning

Inductive Logic Programming (ILP) is a supervised machine learning framework based on logic programming. In ILP a hypothesis is drawn from background knowledge and examples. The examples (E), background knowledge (B) and hypothesis (H) all take the form of logic programs. The background knowledge and the final hypothesis induced from the examples are used to evaluate new instances.

Since logic programming allows for the expression of arbitrary relations between objects, ILP is more expressive than attribute-value representations, enabling flexible use of background knowledge (Bratko & King, 1994; Mooney & Califf, 1995). It also has advantages over approaches such as n-gram models, Hidden Markov Models, neural networks and SVM, which represent examples using fixed length feature vectors (Bratko & King, 1994). These techniques have difficulty representing relations, recursion and unbounded structural representation (Mooney, 2003). ILP, on the other hand, employs a rich knowledge representation language without length constraints. Moreover, the first order logic that is used in ILP limits the amount of feature extraction required in other approaches.

In induction, one begins with some data during the training phase, and then determines what general conclusion can logically be derived from those data. For morphological analysis, the learning data would be expected to guide the construction of word formation rules and interactions between the constituents of a word.

There have been only a few attempts to apply ILP to morphology, and most of these have dealt with languages with relatively simple morphology handling few affixations (Kazakov, 2000; Manandhar et al, 1998; Zdravkova et al, 2005). However, the results are encouraging.

While we focus on Amharic verb morphology, our goal is a general-purpose ILP morphology learner. Thus we seek background knowledge that is plausible across languages that can be combined with language-specific examples to yield rule hypotheses that generalize to new examples in the language.

CLOG is a Prolog based ILP system, developed by Manandhar *et al* (1998)[2], for learning first order decision lists (rules) on the basis of positive examples only. A rule in Prolog is a clause with one or more conditions. The right-hand side of the rule (the body) is a condition and the left-hand side of the rule (the head) is the conclusion. The operator between the left and the right hand side (the sign '***:-***') means ***if***. The body of a rule is a list of goals separated by commas, where commas are understood as conjunctions. For a rule to be true, all of its conditions/goals must be evaluated to be true. In the expression below, ***p*** is true if ***q*** and ***r*** are true or if ***s*** and ***t*** are true.

[2] *CLOG is freely available ILP system at:* http://www-users.cs.york.ac.uk/suresh/CLOG.html)

$$\left.\begin{array}{l} p :- q, r. \\ p :- s, t. \end{array}\right\} \; p \Leftrightarrow (q \wedge r) \vee (s \wedge t)$$

*Where **q**, **r**, **s** and **t** could be facts or predicates with any arity and **p** is a predicate with any number of arguments.*

CLOG relies on output completeness, which assumes that every form of an object is included in the example and everything else is excluded (Mooney & Califf, 1995). We preferred CLOG over other ILP systems because it requires only positive examples and runs faster than the other variants (Manandhar *et al*, 1998). CLOG uses a hill climbing strategy to build the rules, starting from a simple goal and iteratively adding more rules to satisfy the goal until there are no possible improvements. The evaluation of the rules generated by the learner is validated using a gain function that compares the number of positively and negatively covered examples in the current and previous learning stages (Manandhar et al, 1998).

5. Experiment Setup and Data

Learning morphological rules with ILP requires preparation of the training data and background knowledge. To handle a language of the complexity of Amharic, we require background knowledge predicates that can handle stem extraction by identifying affixes, root and vowel identification and grammatical feature association with constituents of the word.

The training data used during the experiment is of the following form:

```
stem([s,e,b,e,r,k,u],[s,e,b,e,r],[s,b,r] [1,1]).
stem([s,e,b,e,r,k],[s,e,b,e,r],[s,b,r], [1,2]).
stem([s,e,b,e,r,x],[s,e,b,e,r],[s,b,r], [1,3]).
```

Figure 3: Sample examples for stem and root learning

The predicate *'stem'* provides a word and its stem to permit the extraction of the affixes and root template structure of the word. The first three parameters specify the input word, the stem of the word after affixes are removed, and the root of the stem respectively. The fourth parameter is the codification of the grammatical features (tense-aspect-mood and subject) of the word.

Taking the second example in Figure 3, the word ***seberk*** has the stem ***seber*** with the root ***sbr*** and is perfective (the first element of the third parameter which is 1) with second person singular masculine subject (the second element of the third parameter is 2).

We codified the grammatical features of the words and made them parameters of the training data set rather than representing the morphosyntactic description as predicates as in approaches used for other languages (Zdravkova et al, 2005).

The background knowledge also includes predicates for string manipulation and root extraction. Both are language-independent, making the approach adaptable to other similar languages. We run three separate training experiments to learn the stem extraction, root patterns, and internal stem alternation rules.

a) Learning stem extraction:

The background predicate *'set_affix'* uses a combination of multiple '*split'* operations to identify the prefix and suffixes attached to the input word. This predicate is used to learn the affixes from examples presented as in Figure 3 by taking only the *Word* and the *Stem* (the first two arguments from the example).

```
set_affix(Word, Stem, P1,P2,S1,S2):-
   split(Word, P1, W11),
   split(Stem, P2, W22),
   split(W11, X, S1),
   split(W22, X, S2),
   not( (P1=[],P2=[],S1=[],S2=[])).
```

Figure 4: Affix extraction predicate

The predicate makes all possible splits of *Word* and *Stem* into three segments to identify the prefix and suffix substitutions required to unify *Stem* with *Word*. In this predicate, P1 and S1 are the prefix and suffix of the *Word*; while P2 and S2 are the prefix and suffix of the *Stem* respectively. For example, if *Word* and *Stem* are ***tgedyalex*** and ***gedl*** respectively, then the predicate will try all possible splits, and one of these splits will result in P1=[***t***], P2=[], S1=[***yalex***] and S2=[***l***]. That is, ***tgedyalex*** will be associated with the stem ***gedl***, if the prefix P1 is replaced with P2 and the suffix S1is replaced with S2.

The ultimate objective of this predicate is to identify the prefix and suffix of a word and then extract the valid stem (*Stem*) from the input string (*Word*).

Here, we have used the utility predicate '*split'* that segments any input string into all possible pairs of substrings. For example, the string ***sebr*** could be segmented as {([]-[*sebr*]), ([*s*]-[*ebr*]), ([*se*]-[*br*]), ([*seb*]-[*r*]), or ([*sebr*]-[])}.

b) Learning Roots:

The root extraction predicate, '*root_vocal'*, extracts *Root* and the *Vowel* with the right sequence from the *Stem*. This predicate learns the root from examples presented as in Figure 3 by taking only the *Stem* and the *Root* (the second and third arguments).

```
root_vocal(Stem,Root,Vowel):-
        merge(Stem,Root,Vowel).

merge([X,Y,Z|T],[X,Y|R],[Z|V]):-
                merge(T,R,V).
merge([X,Y|T],R,[X,Y|V]):-
                merge(T,R,V).
merge([X|Y],[X|Z],W) :-
                merge(Y,Z,W).
merge([X|Y],Z,[X|W]) :-
                merge(Y,Z,W).
```

Figure 5: Root template extraction predicate

The predicate '*root_vocal'* performs unconstrained permutation of the characters in the *Stem* until the first part of the permutated string matches the *Root* character pattern provided during the training. The goal of this predicate is to separate the vowels and the consonants of a *Stem*. In this predicate we have used the utility predicate '*merge'* to perform the permutation. For example, if *Stem* is ***seber*** and the example associates this stem with the *Root* **sbr**, then *'root_temp'*, using '*merge',* will generate many patterns, one of which would be ***sbree***. This, ultimately, will learn that the vowel pattern [***ee***] is valid within a stem.

c) Learning stem internal alternations:
Another challenge for Amharic verb morphology learning is handling stem internal alternations. For this purpose, we have used the background predicate *'set_internal_alter'*:

```
set_internal_alter(Stem,Valid_Stem,St1,St2):-
            split(Stem,P1,X1),
            split(Valid_Stem,P1,X2),
            split(X1,St1,Y1),
            split(X2,St2,Y1).
```

Figure 6: stem internal alternation extractor

This predicate works much like the '*set_affix*' predicate except that it replaces a substring which is found in the middle of *Stem* by another substring from *Valid_Stem*. In order to learn stem alternations, we require a different set of training data showing examples of stem internal alternations. Figure 7 shows some sample examples used for learning such rules.

```
alter([h,e,d],[h,y,e,d]).
alter([m,o,t],[m,e,w,o,t]).
alter([s,a,m],[s,e,?,a,m]).
```

Figure 7: Examples for internal stem alternation learning

The first example in Figure 7 shows that for the words ***hed*** and ***hyed*** to unify, the ***e*** in the first argument should be replaced with ***ye***.

Along with the three experiments for learning various aspects of verb morphology, we have also used two utility predicates to support the integration between the learned rules and to include some language specific features. These predicates are '*template'* and '*feature'*:

- '*template'*: used to extract the valid template for *Stem*. The predicate manipulates the stem to identify positions for the vowels. This predicate uses the list of vowels (vocal) in the language to assign '0' for the vowels and '1' for the consonants.

```
template([],[]).
template([X|T1],[Y|B]):-
        template(T1,B),
        (vocal(X)->Y=0;Y=1).
```

Figure 8: CV pattern decoding predicate

For the stem ***seber*** this predicate tries each character separately and finally generates the pattern [1,0,1,0,1] and for the stem ***sebr***, it generates [1,0,1,1] to show the valid template of Amharic verbs.

- '*feature'*: used to associate the identified affixes and root CV pattern with the known grammatical features from the example. This predicate uses a codified representation of the eight subjects and four tense-aspect-mood features ('tam') of Amharic verbs, which is also encoded as background knowledge. This predicate is the only language-dependent background knowledge we have used in our implementation.

```
feature([X,Y],[X1,Y1]):-
        tam([X],X1),
        subj([Y],Y1).
```

Figure 9: Grammatical feature assignment predicate

6. Experiments and Result

For CLOG to learn a set of rules, the predicate and arity for the rules must be provided. Since we are learning words by associating them with their stem, root and grammatical features, we use the predicate schemas ***rule(stem(_,_,_,_))*** for *set_affix* and *root_vocal*, and ***rule(alter(_,_))*** for *set_internal_alter*. The training examples are also structured according to these predicate schemas.

The training set contains 216 manually prepared Amharic verbs. The example contains all possible combinations of tense and subject features. Each word is first romanized, then segmented into the stem and grammatical features, as required by the '*stem'* predicate in the background knowledge. When the word results from the application of one or more alternation rules, the stem appears in the canonical form. For example, for the word ***gdey***, the stem specified is ***gdel*** (see the second example in Table 1).

Characters in the Amharic orthography represent syllables, hiding the detailed interaction between the consonants and the vowels. For example, the masculine imperative verb 'ግደል' ***gdel*** can be made feminine by adding the suffix 'i' (gdel-i). But, in Amharic, when the dental 'l' is followed by the vowel 'i', it is palatalized, becoming 'y'. Thus, the feminine form would be written 'ግደይ', where the character 'ይ' 'y' corresponds to the sequence 'l-i'.

To perform the romanization, we have used our own Prolog script which maps Amharic characters directly to sequences of roman consonants and vowels, using the familiar SERA transliteration scheme. Since the mapping is reversible, it is straightforward to convert extracted forms back to Amharic script.

After training the program using the example set, which took around 58 seconds, 108 rules for affix extraction, 18 rules for root template extraction and 3 rules for internal stem alternation have been learned. A sample rule generated for affix identification and associating the word constituents with the grammatical features is shown below:

```
stem(Word, Stem, [2, 7]):-
        set_affix(Word, Stem, [y], [], [u], []),
        feature([2, 7], [imperfective, tppn]),
        template(Stem, [1, 0, 1, 1]).
```

Figure 10: Learned affix identification rule example

The above rule declares that, if the word starts with ***y*** and ends with ***u*** and if the stem extracted from the word after stripping off the affixes has a CVCC ([1,0,1,1]) pattern, then that word is imperfective with third person plural neutral subject (tppn).

```
alter(Stem,Valid_Stem):-
  set_internal_alter(Stem,Valid_Stem, [o], [e, w, o]).
```

Figure 11: Learned internal alternation rule example

The above rule will make a substitution of the vowel ***o*** in a specific circumstances (which is included in the program) with ***ewo*** to transform the initial stem to a valid stem in the language. For example, if the Stem is ***zor***, then ***o*** will be replaced with ***ewo*** to give ***zewor.***

The other part of the program handles formation of the root of the verb by extracting the template and the vowel sequence from the stem. A sample rule generated to handle the task looks like the following:

```
root(Stem, Root):-
        root_vocal(Stem, Root, [e, e]),
         template(Stem, [1, 0, 1, 0, 1]) .
```

Figure 12: Learned root-template extraction rule example

The above rule declares that, as long as the consonant vowel sequence of a word is CVCVC and both vowels are *e*, the stem is a possible valid verb. Our current implementation does not use a dictionary to validate whether the verb is an existing word in Amharic.

Finally, we have combined the background predicates used for the three learning tasks and the utility predicates. We have also integrated all the rules learned in each experiment with the background predicates. The integration involves the combination of the predicates in the appropriate order: stem analysis followed by internal stem alternation and root extraction.

After building the program, to test the performance of the system, we started with verbs in their third person singular masculine form, selected from the list of verbs transcribed from the appendix of Armbruster (1908)[3]. We then inflected the verbs for the eight subjects and four tense-aspect-mood features of Amharic, resulting in 1,784 distinct verb forms. The following are sample analyses of new verbs that are not part of the training set by the program:

```
InputWord: [a, t, e, m, k, u]
      Stem: [?, a, t, e, m]
      Template: [1,0, 1, 0, 1]
      Root: [?, t, m]
      GrammaticalFeature: [perfective, fpsn*]
```

Figure 13: Sample Test Result (with boundary alternation)

*fpsn: first person singular neuter

The above example shows that the suffix that needs to be stripped off is ***[k,u]*** and that there is an alternation rule that changes ***'a'*** to ***'?,a'*** at the beginning of the word.

```
InputWord: [t, k, e, f, y, a, l, e, x]
    Stem: [k, e, f, l]
    Template: [1,0, 1, 1]
    Root: [k, f, l]
    GrammaticalFeature: [imperfective, spsf*]
```

Figure 14: Sample Test Result (Internal alternation)

*spsf: second person singular feminine

The above example shows that the prefix and suffix that need to be stripped off are ***[t]*** and ***[a,l,e,x]*** respectively and that there is an alternation rule that changes ***'y'*** to ***'l'*** at the end of the stem after removing the suffix.

The system is able to correctly analyze 1,552 words, resulting in 86.99% accuracy. With the small set of training data, the result is encouraging and we believe that the performance will be enhanced with more training examples of various grammatical combinations.

The wrong analyses and test cases that are not handled by the program are attributed to the absence of such examples in the training set and an inappropriate alternation rule resulting in multiple analysis of a single test word.

Test Word	***Stem***	***Root***	***Feature***
[s,e,m,a,c,h,u]	*[s,e,m,a,?]*	*[s,m,?]*	*perfective, sppn*
[s,e,m,a,c,h,u]	*[s,e,y,e,m]*	*[s,y,m]*	*gerundive, sppn*
[l,e,g,u,m,u]	*[l,e,g,u,m]*	*NA*	*NA*

Table 2: Example of wrong analysis

Table 2 shows some of the wrong analyses and words that are not analyzed at all. The second example shows that an alternation rules has been applied to the stem resulting in wrong analysis (the stem should have been the one in the first example). The last example generated a stem with vowel sequence of '***eu'*** which is not found in any of the training set, categorizing the word in the not-analyzed category.

7. Future work

ILP has proven to be applicable for word formation rule extraction for languages with simple rules like English. Our experiment shows that the approach can also be used for complex languages with more sophisticated background predicates and more examples. While Amharic has more prefixes and suffixes for various morphological features, our system is limited to only subject markers. Moreover, all possible combinations of subject and tense-aspect-mood have been provided in the training examples for the training. This approach is not practical if all the prefix and suffixes are going to be included in the learning process.

One of the limitations observed in ILP for morphology learning is the inability to learn rules from incomplete examples. In languages such as Amharic, there is a range of complex interactions among the

[3] *Available online at:* http://nlp.amharic.org/resources/lexical/word-lists/verbs/c-h-armbruster-initia-amharica/ *(accessed February 12, 2012).*

different morphemes, but we cannot expect every one of the thousands of morpheme combinations to appear in the training set. When examples are limited to only some of the legal morpheme combinations, CLOG is inadequate because it is not able to use variables as part of the body of the predicates to be learned.

An example of a rule that could be learned from partial examples is the following: *"if a word has the prefix* ***'te',*** *then the word is passive no matter what the other morphemes are"*. This rule (not learned by our system) is shown in Figure 15.

```
stem(Word, Stem, Root, GrmFeatu):-
        set_affix(Word, Stem, [t,e], [], S, []),
        root_vocal(Stem, Root, [e, e]),
        template(Stem, [1, 0, 1, 0, 1]),
        feature(GrmFeatu, [Ten, passive, Sub]).
```

Figure 15: Possible stem analysis rule with partial feature

That is, ***S*** is one of the valid suffixes, ***Ten*** is the Tense, and ***Sub*** is the subject, which can take any of the possible values.

Moreover, as shown in section 2, in Amharic verbs, some grammatical information is shown by various combinations of affixes. The various constraints on the co-occurrence of affixes are the other problem that needs to be tackled. For example, the 2nd person masculine singular imperfective suffix ***aleh*** can only co-occur with the 2nd person prefix ***t*** in words like ***t-sebr-aleh***. At the same time, the same prefix can occur with the suffix ***alachu*** for the 2nd person plural imperfective form. To represent these constraints, we apparently need explicit predicates that are specific to the particular affix relationship. However, CLOG is limited to learning only the predicates that it has been provided with.

We are currently experimenting with genetic programming as a way to learn new predicates based on the predicates that are learned using CLOG.

8. Conclusion

We have shown in this paper that ILP can be used to fast-track the process of learning morphological rules of complex languages like Amharic with a relatively small number of examples. Our implementation goes beyond simple affix identification and confronts one of the challenges in template morphology by learning the root-template extraction as well as stem-internal alternation rule identification exhibited in Amharic and other Semitic languages. Our implementation also succeeds in learning to relate grammatical features with word constituents.

9. References

Armbruster, C. H. (1908). *Initia Amharic: an Introduction to Spoken Amharic.* Cambridge: Cambridge University Press.

Beesley, K. R. and L. Karttunen. (2003). *Finite State Morphology*. Stanford, CA, USA: CSLI Publications.

Bender, M. L. (1968). *Amharic Verb Morphology: A Generative Approach.* Ph.D. thesis, Graduate School of Texas.

Bratko, I. and King, R. (1994). *Applications of Inductive Logic Programming*. *SIGART Bull. 5*, 1, 43-49.

Dawkins, C. H., (1960). *The Fundamentals of Amharic*. Sudan Interior Mission, Addis Ababa, Ethiopia.

De Pauw, G. and P.W. Wagacha. *(2007). Bootstrapping Morphological Analysis of Gĩkũyũ Using Unsupervised Maximum Entropy Learning.* Proceedings of the Eighth INTERSPEECH Conference, Antwerp, Belgium.

Gasser, M. (2011). *HornMorpho: a system for morphological processing of Amharic, Oromo, and Tigrinya*. Conference on Human Language Technology for Development, Alexandria, Egypt.

Goldsmith, J. (2001). *The unsupervised learning of natural language morphology*. Computational Linguistics, 27: 153-198.

Hammarström, H. and L. Borin. (2011). *Unsupervised learning of morphology*. Computational Linguistics, 37(2): 309-350.

Kazakov, D. (2000). *Achievements and Prospects of Learning Word Morphology with ILP*, Learning Language in Logic, Lecture Notes in Computer Science.

Kazakov, D. and S. Manandhar. (2001). *Unsupervised learning of word segmentation rules with genetic algorithms and inductive logic programming*. Machine Learning, 43:121–162.

Koskenniemi, K. (1983). *Two-level Morphology: a General Computational Model for Word-Form Recognition and Production*. Department of General Linguistics, University of Helsinki, Technical Report No. 11.

Manandhar, S. , Džeroski, S. and Erjavec, T. (1998). *Learning multilingual morphology with CLOG.* Proceedings of Inductive Logic Programming. 8th International Workshop in Lecture Notes in Artificial Intelligence. Page, David (Eds) pp.135–44. Berlin: Springer-Verlag.

Mooney, R. J. (2003). *Machine Learning*. Oxford Handbook of Computational Linguistics, Oxford University Press, pp. 376-394.

Mooney, R. J. and Califf, M.E. (1995). *Induction of first-order decision lists: results on learning the past tense of English verbs*, Journal of Artificial Intelligence Research, v.3 n.1, p.1-24.

Oflazer, K., M. McShane, and S. Nirenburg. (2001). *Bootstrapping morphological analyzers by combining human elicitation and machine learning*. Computational Linguistics, 27(1):59–85.

Sieber, G. (2005). *Automatic Learning Approaches to Morphology*, University of Tübingen, International Studies in Computational Linguistics.

Yimam, B. (1995). *Yamarigna Sewasiw (Amharic Grammar).* Addis Ababa: EMPDA.

Zdravkova, K., A. Ivanovska, S. Dzeroski and T. Erjavec, (2005). *Learning Rules for Morphological Analysis and Synthesis of Macedonian Nouns*. In Proceedings of SIKDD 2005, Ljubljana.

The Database of Modern Icelandic Inflection (Beygingarlýsing íslensks nútímamáls)

Kristín Bjarnadóttir

The Árni Magnússon Institute for Icelandic Studies
Iceland
kristinb@hi.is

Abstract

The topic of this paper is the Database of Modern Icelandic Inflection (DMII), containing about 270,000 paradigms from Modern Icelandic, with over 5.8 million inflectional forms. The DMII was created as a multipurpose resource, for use in language technology, lexicography, and as an online resource for the general public. Icelandic is a morphologically rich language with a complex inflectional system, commonly exhibiting idiosyncratic inflectional variants. In spite of a long history of morphological research, none of the available sources had the necessary information for the making of a comprehensive and productive rule-based system with the coverage needed. Thus, the DMII was created as a database of paradigms showing all and only the inflectional variants of each word. The initial data used for the project was mostly lexicographic. The creation of a 25 million token corpus of Icelandic, the MÍM Corpus, has made it possible to use empirical data in the development of the DMII, resulting in extensive additions to the vocabulary. The data scarcity in the corpus, due to the enormous number of possible inflectional forms, proves how important it is to use both lexicographic data and a corpus to complement each other in an undertaking such as the DMII.

Keywords: Morphology, Inflectional database, Icelandic

1. Introduction

This paper describes the Database of Modern Icelandic Inflection (DMII; Beygingarlýsing íslensks nútímamáls), a collection of (at present) 270,000 paradigms with about 5.8 million inflectional forms, i.e., word forms with grammatical tags.[1] The DMII was initially created to serve two purposes, i.e., to produce data for use in LT projects, and to make the resulting paradigms available to the general public on the website of The Árni Magnússon Institute for Icelandic Studies (AMI).[2] From the outset the aim was to present Icelandic inflection 'as is', with as full a description of variants as possible. With this in mind, the decision was made to produce a full paradigm for as large a proportion of the vocabulary as possible, instead of producing a rule system for the generation of inflection by inflectional classes. It turned out that in spite of centuries of research on Icelandic morphology, the necessary data for a productive rule system was simply not available. The problem is that for analysis it may be acceptable to use an overgenerating rule-system, but for production it is not, if the end result, i.e., a text, is expected to be correct. This is a very relevant point, as demonstrated by the facts that the data from the DMII is used for context sensitive grammar correction, and the paradigms are also widely used online by the general public for reference. Native speakers of Icelandic need guidance to cope with a very complex inflectional system.[3] Work on the DMII started in 2002, as a part of an LT Program launched by the Minister of Education, Science and Culture (Rögnvaldsson et al., 2009). The first version of the data was made available for LT use in 2004, and the online version was opened the same year. Data from the DMII has been used in various LT projects, such as search engines, PoS tagging, context sensitive correction, in language teaching, lexicography, etc. Both the DMII and the Tagged Icelandic Corpus (The MÍM Corpus) (Helgadóttir et al., 2012) are being produced at the AMI, and the two projects run in tandem. The original sources for the DMII were lexicographic, i.e., the electronic version of the classic *Dictionary of Icelandic* (Árnason, 2000), containing 135,000 headwords, and the AMI's lexicographic archives. Various other sources are now used, but the next stage is to include the vocabulary contained in the MÍM Corpus. This work is now in progress.

The paper is structured as follows. Section 2 contains a short description of the richness of Icelandic morphology, followed in Section 3 by an account of the method used in creating the DMII and the two accessible versions of it, one for LT purposes and one online, for the general public. Section 4 contains an account of the limitation of the sources of information on Icelandic inflection, followed by Section 5, a description of the independent research needed to fill gaps in the sources. The inclusion of the vocabulary from the MÍM Corpus is described in Section 6, with the lesson learned on data scarcity in a language with a very rich morphology in Section 7. The conclusion is in Section 8.

2. The richness of Icelandic morphology

As can be inferred from the ratio of inflectional forms to paradigms in the DMII, i.e., 5.8 million inflected forms in 270,000 paradigms, the inflectional system of Icelandic is rich, with up to 16 inflectional forms to a noun, 120 to an

[1] The term Modern Icelandic is here used of contemporary Icelandic, i.e., 21st century usage.

[2] http://bin.arnastofnun.is/

[3] In November 2011 there were 268,011 pagewiews, and 52,570 visits from 60 countries to the online DMII. Iceland has about 320,000 inhabitants and most of the visits are domestic. The users are also in contact via email, with queries, additions and corrections.

adjective, and 107 to a verb, not including variants. This is reflected by the size of the tagset used in the PoS tagging of Icelandic, with over 700 tags (Pind et al., 1991) and (Helgadóttir et al., 2012).

The Icelandic inflectional system is also quite complex, as the endings that mark grammatical categories can, in some instances, have a number of variants, e.g., -s/-ar/-ur in the genitive singular of masculine nouns with a certain structure of base form, i.e., the ending -ur in the nominative singular. The result is a proliferation of inflectional variants, e.g., *þröskuldar/þröskulds*, genitive singular of the masculine noun *þröskuldur* 'threshold'. Furthermore, stem changes are common, both in vowels and consonants.

In the case of inflectional variants, the grammatical tradition in Iceland is to say that a word can belong to more than one inflectional class. However, the method of producing the paradigms for the DMII does not allow that; each lemma is shown in full in one paradigm, including all variants.[4] An inflectional class arrived at this way is in fact a unique bundle of inflectional rules, specific to a word or group of words.

3. The production of the paradigms

Initially, the paradigms in the DMII were produced with simple Unix shell scripts, by merging a matrix of inflectional endings containing slots for numbered variants of stems with records for individual words. The result was a set of XML files. The concept of the database now in use is similar. The record for each word contains all variants of the stem, and information on which parts of the full paradigm are applicable in each case ("flags"), e.g., no singular for pluralia tantum, no active voice in mediopassive verbs, no past participles for some verbs, etc. These records are merged with a matrix for the appropriate inflectional class and the resulting inflectional forms are then stored with morphosyntactic tags according to their place in the matrix.[5]

	Indefinite (+)			
	Singular (+)		Plural (+)	
Nom.	1+0	*akur*	3+ar	*akrar*
Acc.	1+0	*akur*	3+a	*akra*
Dat.	3+i	*akri*	2+um	*ökrum*
Gen.	1+s	*akurs*	4+a	*akra*
	Definite (+)			
Nom.	1+inn	*akurinn*	3+arnir	*akrarnir*
Acc.	1+inn	*akurinn*	3+ana	*akrana*
Dat.	3+inum	*akrinum*	2+unum	*ökrunum*
Gen.	1+sins	*akursins*	4+anna	*akranna*

Table 1: Matrix for the noun *akur*.

Table 1 shows a matrix for one class of nouns, with the resulting inflectional forms in italics. The word *akur* 'field, meadow' is flagged for the grammatical categories number and definiteness, i.e., +sg., +pl., +indef., +def., which specifies that there are no gaps in the paradigm. (For pluralia tantum (flagged −sg.), the singular would be left blank.) The table contains English translations of the Icelandic abbreviations used in the online version, which is similar to Table 1, leaving out the columns 2 and 4 (numbers for stems and the endings), retaining abbreviations and inflectional forms. The metalanguage is Icelandic.

The most commonly used output for LT purposes is a simple list with 6 fields, as in the 16 inflectional forms for the word *akur* 'field, meadow' in Table 2. The fields are lemma, identifier (number), word class or gender of nouns, type (i.e., common language, named entity, terminology, etc.; 'com' in Table 2 signifies 'common language'), inflectional form, and tag. The tags shown here are English translations, as in Table 1.

```
akur;472164;masc;com;akur;NOM-SG
akur;472164;masc;com;akurinn;NOM-SG-DEF
akur;472164;masc;com;akur;ACC-SG
akur;472164;masc;com;akurinn;ACC-SG-DEF
akur;472164;masc;com;akri;DAT-SG
akur;472164;masc;com;akrinum;DAT-SG-DEF
akur;472164;masc;com;akurs;GEN-SG
akur;472164;masc;com;akursins;GEN-SG-DEF
akur;472164;masc;com;akrar;NOM-PL
akur;472164;masc;com;akrarnir;NOM-PL-DEF
akur;472164;masc;com;akra;ACC-PL
akur;472164;masc;com;akrana;ACC-PL-DEF
akur;472164;masc;com;ökrum;DAT-PL
akur;472164;masc;com;ökrunum;DAT-PL-DEF
akur;472164;masc;com;akra;GEN-PL
akur;472164;masc;com;akranna;GEN-PL-DEF
```

Table 2: Output for LT: Example from a CSV file.

The matrices were (and still are) produced at need, every time a new variant makes a new inflectional pattern necessary, thus creating a new unique bundle of inflectional rules. This makes the description truly 'bottom-up', as it is purely based on the actual inflection of individual words.

There are at present over 630 such inflectional classes in the DMII, some with tens of thousands of words, but others showing the idiosyncracy of individual words, sometimes due to historical remnants of obsolete inflectional classes still attested in common phrases and idioms in the modern language.[6]

Out of the inflectional classes for nouns, adjectives and verbs, 42% contain no variants or other complicating features, such as internal inflection or unsystematic gapping.[7]

[4] The base form of the lemma is decisive in the division of paradigms. A variant base form will therefore produce two paradigms, as in the nouns *sannleikur/sannleiki* 'truth'. The base form for nouns is the nominative singular.

[5] For each word, intuition is used to choose between possible inflectional variants (i.e., between inflectional classes) and to assign values to the flags deciding the structure of the paradigm, e.g., +/−plural for nouns, +/−degree for adjectives, and +/−mediopassive for verbs, etc., when data is not available.

[6] The inflectional system of Icelandic has undergone some changes through the centuries, both structural changes (i.e., changes of inflectional classes per se), and drift of vocabulary between inflectional classes. As a whole, these changes are fairly minor, but the result is nevertheless very apparent in individual word forms.

[7] The DMII contains 49,296 pairs of variants and 262 triplets (Feb. 2012).

Lemmas	Infl. classes	Lemmas	Infl. classes
<10.000	5		
5.000-9.999	7		
1.000-4.999	29	<1000	41
500-999	21	<500	62
100-499	72	<100	134
50-99	47	<50	181
10-49	126	<10	307
2-9	146	<2	453
1	156		

Table 3: The productivity of inflectional classes (nouns, adjectives and verbs, 609 classes).

The number of words in each inflectional class varies greatly, as shown in Table 3, with about 25% of the inflectional classes of nouns, adjectives and verbs showing bundles of rules describing the truly idiosyncratic inflection of single words.

4. The source material and the lack of information

At the outset, the bulk of the source material for the project was from lexicographic resources, such as data from the digitized version of the classic *Dictionary of Icelandic* (Árnason, 2000), first published in 1963, as well as 20th century headwords from the AMI Written Language Archive (WLA, Ritmálssafn Orðabókar Háskólans[8]), a collection of citations created for a historical dictionary of Icelandic from the 16th century to modern times, which contains over 700,000 headwords. The third initial source was a book of personal names (Kvaran and Jónsson, 1991), containing about 4,800 personal names. The first version of the DMII contained 176,000 paradigms, mainly of vocabulary from these three sources. The additional material in later versions of the DMII comes from various other sources, and it is sometimes the result of cooperation on projects creating search engines capable of finding all inflectional forms of a word, by entering either the base form (the headword) or any inflectional form. This is true of the online version of the new translation of the *Bible* (2007), and the Icelandic telephone directory, a good source of named entities. These sources are, however, not to be relied upon for actual information on inflection, except for random forms, and it is only the published lexicographic work (*The Dictionary of Icelandic* and the book of names (Kvaran and Jónsson, 1991)) that contain a systematic coding for inflection, and only a partial one at that. It was therefore clear from the beginning that grammatical descriptions would be relied on, but the fact that these would prove to be incomprehensive was not immediately obvious.

The tradition in Icelandic dictionaries is to give certain inflectional forms as indicators of inflectional class, such as the genitive singular and nominative plural for nouns, either by showing the endings, e.g., *bátur*, (masc.) -s, -ar 'boat', or, in the case of wowel change, by showing the whole inflectional form, e.g., *köttur*, (masc.) *kattar, kettir* 'cat'. The remaining inflectional forms of nouns are very rarely shown in dictionaries, although some of these can be unpredictable, as in the masculine noun *bátur* 'boat', dative singular indefinite *báti* or *bát*, and dative singular definite *bátnum* (not **bátinum*) (cf. *köttur* 'cat', dat.sg.indef. *ketti*, dat.sg.def. *kettinum*). Information on the inflection in other word classes is also fragmentary, with the description of verbs usually confined to the principal parts, i.e., three or four inflectional forms, depending on inflectional class. The inflection of adjectives is very often omitted altogether, although it is not wholly predictable from the base forms. The dictionaries cover a large vocabulary, but they only give information on a part of the inflectional forms needed for complete paradigms.

The grammatical descriptions, on the other hand, show full paradigms of selected examples to give a survey of the system, i.e., they present the general structure 'top-down'.[9] This is true throughout the history of the description of Icelandic inflection, from the first one, which is fragmentary, usually referred to as *Grammaticæ islandicæ rudimenta*, first published in Copenhagen in 1651 (Jónsson, 1688), to the first definitive one, Rasmus Rask's *Vejledningen* (1811), up until now (cf. Kvaran, 2005). All the inflectional descriptions share the same characteristics, i.e., they present a set of generalized inflectional classes, mentioning exceptions at times.[10]

The grammar books are therefore not a good source on individual words, apart from the few selected examples, which usually are from the classic Icelandic core vocabulary. The emphasis on the core vocabulary means that data on loanwords, informal language, and slang is mostly absent from the sources, even though that is where changes to the system will first appear. Such material is ignored, perpaps for reasons of language purism,[11] even when such words seem to be fully adapted to the language, appearing in any syntactic context and being fully inflected, sometimes exhibiting major systematic differences from the traditional inflectional classes.

A case in point is the ending -i which the grammatical literature claims to be universal in the dative singular of neuter nouns, with the exceptions of four words. However, the data used for the DMII shows that the dative ending fluctuates between -i and -0 in multisyllabic neuter loanwords. This is the case for the the word *fennel* which is adopted from English, instead of the Icelandic version *fennika* (fem.), preferred by the purists. Even though some loanwords exhibiting this kind of fluctuation were adopted in the 18th century, they are still absent from the grammatical surveys. The word *arsenik* 'arsenic' is a case in point:[12]

[8] `http://arnastofnun.is/page/arnastofnun\_gagnasafn\_ritmal`

[9] The notable exception is Svavarsdóttir (1993), a monograph on the productivity of the inflectional classes of nouns, based on an empirical study of a corpus of 2.5 million running words.

[10] The most comprehensive one, *Islandsk grammatik* (Guðmundsson, 1922), proved to be immensely helpful, especially in the conjugation of verbs, but nevertheless it shares the characteristics of being a survey with the rest of the grammatical literature.

[11] There is a strong tradition of language purism in Iceland, i.e., a strong bias in favour of neologisms coined from Icelandic words in preference to loanwords.

[12] Citations from `http://timarit.is`, The National and

Hvönn er svolítið lík fennel<dat.>
'Angelica is a little bit like fennel'
. . . ásamt brytjuðu fenneli<dat.>
'. . . with diced fennel'
. . . byrlað eitur, drepinn með arsenik<dat.>
'. . . poisoned with arsenic'
Hann var myrtur með arseniki<dat.>
'He was murdered with arsenic'

The lack of information in the two types of sources is therefore as follows: The dictionaries give partial information on quite a large vocabulary, but the grammatical descriptions give exhaustive information on a part of the vocabulary. Additional data is clearly needed.

5. Research for the DMII

It is of course only necessary to research possible ambiguous inflectional forms, but considerable research was (and is) needed to fill the above-mentioned gaps in the sources used for the DMII, using all the available sources at the AMI Department of Lexicography, i.e., the archives, citations in printed dictionaries, and digitized text collections, both at the AMI and at the National Library, as well as native speaker intuition, both from linguists and others. The users of the online DMII are also very generous with their opinions and suggestions for additions and improvement. At a pinch, Google is also used for reference, time-consuming though that may be. It still remains a fact that some problematic inflectional forms simply cannot be found anywhere, by any means.

To name an example, the word *Yggdrasill* (from Old Norse mythology, 'the great tree whose branches and roots extend through the universe') does not appear in the dative in any of the Old Icelandic sources. There are two possible dative forms, *Yggdrasil* and *Yggdrasli* and neither of them are attested in the literature. The second variant would be the regular inflection, but confusion with the neuter noun *drasl* (dat. *drasli*) 'rubbish' makes modern speakers cringe (or laugh), although the first variant is not quite acceptable either. This seems to make speakers avoid referring to a shop in today's Reykjavík named *Yggdrasill* in the dative, making do with syntactic context where another case can be used.[13] In the case of unattested inflectional forms, the choice is between a blank and an educated guess; both occur in the DMII. In the case of *Yggdrasill*, the first dative variant is shown (*Yggdrasil*), with a note in the online version stating that the form is unattested in the trustworthy sources.

Not all attested forms find their way into the DMII, although the purpose is description rather than prescription. Attested but totally unacceptable inflectional variants are not included in the DMII, such as the plural *fótar* of *fótur* 'foot', instead of *fætur*. (The plural *fótar* is sometimes heard in the speech of children and foreign learners, cf. English *foot*, pl. *feet*, not *foots*.) Such forms can often be excluded on the grounds of frequency, but the balance between description and prescription is a difficult one and the choices made can be subjective. For LT analysis, it might be better to include unacceptable variants, but the users of the online version would be up in arms to see them, as inclusion in the DMII is taken to be a kind of recognition of correctness.[14]

The limitations on the kind of research described here are of course the fact that the researcher has to rely on intuition to a great extent when deciding what to search for. It is simply not possible to use this method to search exhaustively in unannotated material, as the word forms themselves are ambiguous, both within paradigms and between lemmas. The identical nominative and accusative singular forms of the noun *akur* in Table 1. are examples of ambiguity within a paradigm, and the word form *minni* is an example of ambiguity between lemmas, as it appears in four different paradigms as 36 different inflectional forms, i.e., as the comparative of *lítill* 'small' (20 different tags), in the verb *minna* 'remind; remember' (10 different tags), in the neuter noun *minni* 'memory' (5 different tags), and as the feminine dative singular of the 1st person possessive pronoun *minn* 'my'.

The ambiguity is extensive. There are 2.8 million word forms (unique character strings) contained in the 5.8 million inflectional forms in the DMII, 1.8 million of the word forms are unique inflectional forms and thus unambiguous, but 1 million word forms are ambiguous. The figures for the ambiguity of inflectional forms within and between lemmas is shown in Table 4.[15]

Inflectional forms in DMII	5,881,374	
Unambiguous	1,850,090	31.5%
Ambiguous within 1 lemma	3,619,482	61.5%
Ambiguous between lemmas	63,641	1.1%
Ambiguous within and between lemmas	348,161	5.9%

Table 4: Ambiguity of inflectional forms in the DMII.

An annotated corpus, such as the MÍM Corpus, is therefore an extremely important resource in the work on the DMII, both to resolve ambiguity and as a source of additional vocabulary.

6. The MÍM Corpus and the DMII

The MÍM Corpus is due to be completed later this year, and it will contain about 25 million running words. The first stage in comparing the vocabulary in the MÍM Corpus and the DMII is now in progress, with inclusion in the DMII in mind. Word forms from the ca. 17.7 million running words available from the MÍM Corpus when the process began have been compared to the DMII.[16] When all strings

University Library of Iceland's digital library of journals and newspaper texts.

[13] Examples of the word form *Yggdrasli* can be found on the web, often in disputes about the dative form.

[14] The solution in the DMII is data driven, which is much too liberal for some, but not descriptive enough for others. There are, however, extensive notes on the variants on the website to help the users cope. The DMII could therefore perhaps be said to be prescriptive, as it is used for guidance, but that prescription is by data, rather than by fiat. Still, the question of acceptability is an acute one.

[15] Figures from October 2011.

[16] It should be noted that the version of the MÍM Corpus used in the comparison contains additional material which will not be a

containing numerals, symbols and punctuation have been removed, the total number of tokens used in the comparison stands at 16,245,429.

Tokens	16,245,429
Unique tagged forms	737,856
In DMII	425,238
Not in DMII	312,618

Table 5: Tokens and unique tagged word forms in 1st batch of MÍM.

The lemmatization of the tagged word forms not found in the DMII has been checked and corrected manually. The corrections range from changes in the form of the lemma to changes of word class and/or gender of nouns. An example of both occurs in the lemmatization of the inflectional form *Miklagarði* with the lemma form *Miklagarð*, tagged as a feminine noun (fem.sg.acc.proper name). The correct lemma is *Mikligarður*, i.e., a masculine noun, with the nominative ending -ur, and a sound change in the stem.[17] It should be noted that the lemmatizer is used on PoS tagged text, and thus it inherits the mistakes made in the assignment of word class or the gender of nouns made by the tagger. This in turn is a source of mistakes in the assignment of the form of the lemma.

Of the 312,618 tagged word forms inspected at this stage of the work, 34.5% were found to be correctly tagged, both for the form of the lemma and word class. The number of inflectional forms in the major word classes (and gender for nouns) is shown in Table 6, where the first column of figures is the number of inflectional forms as tagged in the MÍM Corpus, the second one is the number of correctly assigned inflectional forms in the MÍM Corpus, and the last one is the result of the correction, i.e., the actual number of inflectional forms in the word class.

	MÍM	Correct	Result
n.neut.	70,770	31,942	54,317
n.masc.	142,630	42,910	64,709
n.fem.	78,908	46,207	64,499
adj.	23,835	11,713	18,345
verb	7,639	1,035	3,210

Table 6: Number of inflectional forms per word class, before and after correction.

60% of the word forms from MÍM not found in the DMII were found to be true Icelandic inflectional forms. These are the candidates for inclusion in the DMII. The remaining 40% were classified into a few categories, simultaneously with the correction of the lemmatization. The bulk of these uninflectable word forms and extraneous material is foreign words, 24.6% of the total, but other categories include errors (5.8%), abbreviations (1.6%), and computer-oriented strings (email addresses, urls, etc.), (0.7%).

This material has no direct relevance to the DMII, but it may come in useful in projects using the DMII data, such as context sensitive spelling correction. Hopefully, this data can also be of use in further work on the MÍM Corpus.

The preliminary results indicate that just over 125,000 paradigms should be added to the DMII. With an average number of inflectional forms per paradigm just exceeding 20, the necessary additions to the DMII could consist of over 2.5 million inflectional forms, bringing the total number of inflectional forms in the DMII to over 8.3 million. These figures are, however, preliminary only; the corrected list of lemmas from the MÍM Corpus has to be checked against the lemma list from the DMII again. This can not be said to be completed until the new material has been added to the DMII, as that process will serve to verify the lemmatization.

Work on the inclusion of the additional paradigms in the DMII is in progress. Inflectional class is assigned to each lemma, by comparison to previous DMII material, with the aid of a recently developed compound splitter.[18] Values are then assigned to the flags for grammatical categories for each lemma (cf. Table 1). The actual inflectional forms from the MÍM Corpus will be used for both processes, along with additional data from other sources at need. A revision of paradigms presently in the DMII will also take place, when the MÍM Corpus yields additional forms.

7. Data scarcity in a rich morpholgy

The part of the MÍM Corpus used in the project described here yields 737,856 word forms, and the estimated number of inflectional forms is just under 623,000 (84%). This is only a small part 5.8 million inflectional forms presently found in the DMII.[19] These figures give an indication of how large a corpus would have to be in order to be a sufficient base for a description (or a rule-system) of Icelandic inflection. A description of inflection based solely on the MÍM Corpus would be very meager indeed.[20] This problem is no surprise to Icelandic lexicographers, who have always had to cope with a similar problem in a different context (Pind et al., 1993), as the same kind of scarcity problem is seen in the search for inflectional forms as in the search for words in specific syntactic context.

The proposed addition of a further 125,000 paradigms to the DMII, on the basis of the comparison with the MÍM Corpus, brings the total of inflectional forms in the DMII to 8.3 million, i.e., the 623,000 forms from the Corpus spawn

part of the final version. The tagging is not the final product either, with some texts not tagged with the final combination of taggers described in Loftsson et al. (2010).

[17] *Mikligarður* is the Old Norse name of Constantinople, sometimes still used of Istanbul in Modern Icelandic. The word is a compound, from *mikill* 'great' and *garður* 'seat (of a king); city'.

[18] Work in progress, by Jón Friðrik Daðason, in cooperation with the AMI.

[19] The total number of lemmas in the MÍM Corpus cannot be compared to the number of lemmas in the DMII, as the comparison of word forms described above was defined to word forms not present in the DMII. The lemmatization of the remaining word forms, i.e., those appearing both in the MÍM Corpus and the DMII, is known to contain too many errors to be meaningful at this stage.

[20] For comparison, the *Icelandic Frequency Dictionary* (Pind et al., 1991), which is based on a corpus of 500,000 running words, contains 31,876 lemmas and 59,343 inflectional forms.

a estimated 2.5 million new forms. This would bring the total number of paradigms in the DMII to 395,000, still far short of the over 700,000 headwords found in the largest lexicographic archive (WLA) at the AMI. The indication is, therefore, that the lexicographic data and the corpus complement each other as valuable sources.

8. Conclusion

As pointed out in the introduction, the DMII was intitially made mainly for two purposes, i.e., to produce data for LT use and as reference material for the general public. In the process of the work, the role of the DMII in language research has become increasingly important. In spite of the reputation of Icelandic as a relatively well-researched language, with centuries of history of morphological description, it turned out that there was and still is a lot of work to be done in the field. The fact that the inflectional system is both complex and irregular, with enormous fluctuation between variant forms and inflectional classes, made it necessary to produce a complete set of paradigms, which also allows for notes to be included on individual words at need. The notes contain data on the underlying research, and indications on usage. The usage notes are published on the website, and they are aimed at the general public, as there can often be semantic or stylistic restrictions on the choice of variants.

The production of a rule system of Icelandic inflection is gradually becoming more feasible, as the scope of the DMII is expanded. A morphological analyzer based on the DMII would also be useful, as the DMII will of course never be all-inclusive. It should however be emphasized that the DMII is still work in progress.

The data from the DMII is available online for LT projects on the AMI website, free of charge. Conditions on the use of the data are published on the website.

9. Acknowledgements

The project was initially financed by The Language Technology Program of the Ministry of Education, Science and Culture which ran during the years 2002–2004. The creation of the DMII database was financed by Technology Development Fund in 2005–2006. The company Já has cooperated with the AMI since 2008 on maintaining the database, as well as giving a grant in order to facilitate the use of the DMII for LT purposes. The author would like to thank numerous contributors to the project since 2002, including the well-wishers contributing additions and corrections to the online version by email. Special thanks are due to Þórdís Úlfarsdóttir, who worked on the project for two years at a critical time, and to Sigrún Helgadóttir, from the MÍM Corpus, and Jón Friðrik Daðason, M.Sc. student in Computer Science at the University of Iceland, without whose help the comparison of data from the DMII and MÍM would not have been made. Thanks are also due to the anonymous SALTMIL reviewers for their helpful comments.

10. References

M. Árnason. *Íslensk orðabók [Dictionary of Icelandic]*. 2000. Edda hf., Reykjavík, 3rd edition, electronic version.

Biblían [The Bible]. 2007. Hið íslenska biblíufélag, Reykjavík, 11th edition.

V. Guðmundsson. *Islandsk grammatik [Grammar of Icelandic]*. 1922. Hagerup, Copenhagen.

S. Helgadóttir, Á. Svavarsdóttir, E. Rögnvaldsson, K. Bjarnadóttir, and H. Loftsson. 2012. The Tagged Icelandic Corpus (MÍM). In *Proceedings of "Language Technology for Normalization of Less-Resourced Languages", workshop at the 8th International Conference on Language Resources and Evaluation, LREC 2012*, Istanbul, Turkey. Submitted.

R. Jónsson. *Grammaticæ islandicæ rudimenta*. 1688. E theatro Sheldoniano, Oxford, 2nd edition [1st edition, 1651].

G. Kvaran. *Íslensk tunga 2. Orð. Handbók um beygingar- og orðmyndunarfræði*. 2005. Almenna bókafélagið, Reykjavík.

G. Kvaran and S. Jónsson. *Nöfn Íslendinga [The Names of the Icelanders]*. 1991. Heimskringla, Reykjavík, 1st edition.

H. Loftsson, J. H. Yngvason, S. Helgadóttir, and E. Rögnvaldsson. 2010. Developing a PoS-tagged corpus using existing tools. In *Proceedings of "Creation and use of basic lexical resources for less-resourced languages", workshop at the 7th International Conference on Language Resources and Evaluation, LREC 2010*, Valetta, Malta.

J. Pind, F. Magnússon, and S. Briem. 1991. *Íslensk orðtíðnibók [The Icelandic Frequency Dictionary]*. The Institute of Lexicography, University of Iceland, Reykjavík.

J. Pind, K. Bjarnadóttir, J. H. Jónsson, G. Kvaran, F. Magnússon, and Á. Svavarsdóttir. 1993. Using a Computer Corpus to Supplement a Citation Collection for a Historical Dictionary. *International Journal of Lexicography*, 6(1):1–18.

R. K. Rask. 1811. *Vejledningen til det Islandske eller gamle nordiske Sprog*. Schubothes Forlag, Copenhagen.

E. Rögnvaldsson, H. Loftsson, K. Bjarnadóttir, S. Helgadóttir, A. B. Nikulásdóttir, M. Whelpton, and A. K. Ingason. 2009. Icelandic Language Resources and Technology: Status and Prospects. In R. Domeij, K. Koskenniemi, S. Krauwer, B. Maegaard, E. Rögnvaldsson, and K. de Smedt, editors, *Proceedings of the NODALIDA 2009 Workshop Nordic Perspectives on the CLARIN Infrastructure of Language Resources*. Odense, Denmark.

Á. Svavarsdóttir. *Beygingakerfi nafnorða í nútímaíslensku [The Inflectional System of Nouns in Modern Icelandic]*. 1993. Málvísindastofnun Háskóla Íslands, Reykjavík.

Natural Language Processing for Amazigh Language: Challenges and Future Directions

Ataa Allah Fadoua, Boulaknadel Siham
CEISIC, IRCAM
Avenue Allal El Fassi, Madinat Al Irfane, Rabat, Morocco
E-mail: {ataaallah, boulaknadel}@ircam.ma

Abstract

Amazigh language, as one of the indo-European languages, poses many challenges on natural language processing. The writing system, the morphology based on unique word formation process of roots and patterns, and the lack of linguistic corpora make computational approaches to Amazigh language challenging.

In this paper, we give an overview of the current state of the art in Natural Language Processing for Amazigh language in Morocco, and we suggest the development of other technologies needed for the Amazigh language to live in "information society".

Keywords: Amazigh Language, Natural Language Processing, Less-resourced language

1. Introduction

During the last few decades, most researches have focused on automatic natural language processing in European and East Asian languages at the expense of native languages of many countries that are under or less resourced. The Moroccan Amazigh language is a part of this list. For many years, it has been neglected and less studied from computational point of view.

However, the official status and the institutional one have enabled Amazigh language to get an official spelling, proper coding in Unicode Standard, appropriate standards for keyboard realization, and linguistic structures that are being developed with a phased approach. This process was initiated and undertaken by spelling standardization and establishment of segmentation rules of the spoken chain (Ameur et al., 2006), character encoding specified by extended ASCII, Alphabetical Arrangement (Outahajala, 2007), incorporation into Unicode standard (Andries, 2008; Zenkouar, 2008), implementation of a standard keyboard layout , building new Tifinaghe fonts (Ait Ouguengay, 2007), vocabularies' construction (Ameur et al., 2006-a; Kamel, 2006; Ameur et al., 2009-a; Ameur et al., 2009-b), and elaboration of grammar rules (Boukhris et al., 2008).

Nevertheless, all these stages of standardization are not sufficient for a less-resourced language as Amazigh to join the well-resourced languages in information technology. In this context, many scientific researches are undertaken at national level to improve the current situation. Primarily, they focus on optical character recognition (Amrouch et al., 2010; Es Saady et al., 2010; Fakir et al., 2009). But those concentrated on natural language processing are limited (Iazzi and Outahajala, 2008; Ataa Allah and Jaa, 2009; Boulaknadel, 2009, Es Saady et al., 2009; Ataa Allah and Boulaknadel, 2010; Outahajala et al., 2010; Boulaknadel and Ataa Allah, 2011).

The remainder of this paper is divided into four main sections. In the first, we give a brief overview of the Moroccan Amazigh language features. In the second section, we present and discuss some of Amazigh linguistic challenges. In the third section, we survey existing systems and resources built for Amazigh languages at the Royal Institute of Amazigh Culture (IRCAM). While in the fourth section, we try to identify needs and suggest some future directions on Amazigh natural language processing.

2. Moroccan Amazigh language features

The Amazigh language, known as Berber or Tamazight, is a branch of the Afro-Asiatic (Hamito-Semitic) languages (Greenberg, 1966; Ouakrim, 1995). Nowadays, it covers the Northern part of Africa which extends from the Red Sea to the Canary Isles, and from the Niger in the Sahara to the Mediterranean Sea.

In Morocco, this language is divided, due to historical, geographical and sociolinguistic factors, into three main regional varieties, depending on the area and the communities: Tarifite in North, Tamazight in Central Morocco and South-East, and Tachelhite in the South-West and the High Atlas.

Since the ancient time, the Amazigh language has its own writing system that has been undergoing many slight modifications. In 2003, it has also been changed, adapted, and computerized by IRCAM, in order to provide the Amazigh language an adequate and usable standard writing system. This system is called Tifinaghe-IRCAM (Ameur et al., 2004).

2.1 Tifinaghe-IRCAM graphical system

Since 2003, Tifinaghe-IRCAM has become the official graphic system for writing Amazigh in Morocco. This system contains:

- 27 consonants including: the labials (ⴱ, ⴼ, ⵎ), the dentals (ⵜ, ⴷ, ⵟ, ⴹ, ⵏ, ⵔ, ⵕ, ⵍ), the alveolars (ⵙ, ⵣ, ⵚ, ⵥ), the palatals (ⵛ, ⵊ), the velar (ⴽ, ⴳ), the labiovelars (ⴽⵯ, ⴳⵯ), the uvulars (ⵇ, ⵅ, ⵖ), the pharyngeals (ⵄ, ⵃ) and the laryngeal (ⵀ);
- 2 semi-consonants: ⵢ and ⵡ;
- 4 vowels: three full vowels ⴰ, ⵉ, ⵓ and neutral vowel (or schwa) ⴻ which has a rather special status in Amazigh phonology.

2.2 Punctuation and numeral

No particular punctuation is known for Tifinaghe. IRCAM has recommended the use of the international symbols: “ ” (space), “.”, “,”, “;”, “:”, “?”, “!”, “…”, for punctuation markers; and the standard numeral used in Morocco (0, 1, 2, 3, 4, 5, 6, 7, 8, 9) for Tifinaghe writing.

2.3 Directionality

Historically, in ancient inscriptions, the Amazigh language was written horizontally from left to right, from right to left, vertically upwards, downwards or in boustrophedon (as illustrated in Figure 1) . However, the orientation most often adopted in Amazigh language script is horizontal and from left to right, which is also adopted in IRCAM-Tifinaghe writing.

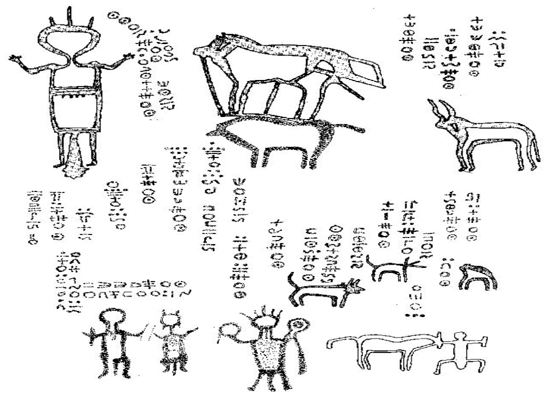

Figure 1: Plate 9 Anou Elias Valley Mammanet (Niger). Henri Lhote, The engravings of Wadi Mammanet. Les Nouvelles Editions Africaines. 1979

3. The complexity of Amazigh in Natural Language Processing

Amazigh is the second official language of Morocco. However, it has been less studied from computational point of view for many years. It has only a limited set of users over the world which makes the interest in developing NLP applications less attractive for foreign developers. Moreover, Amazigh is among the languages having rich morphology and different forms of writing. Below we describe the difficulties that the Amazigh language confronts in developing natural language applications.

3.1 Amazigh script

Amazigh is one of the languages with complex and challenging pre-processing tasks. Its writing system poses three main difficulties:

- -Writing forms' variation that requires a transliterator to convert all writing prescriptions into the standard form 'Tifinaghe – Unicode'. This process is confronted with spelling variation related to regional varieties ([tfucht] [tafukt] (sun)), and transcription systems ([tafuct] [tafukt]), especially when Latin or Arabic alphabet is used.
- -The standard form adopted 'Tifinaghe – Unicode' requires special consideration even in simple applications. Most of the existed NLP applications were developed for Latin script. Therefore, those that will be used for Tifinaghe – Unicode require localization and adjustment.
- -Different prescriptions differ in the style of writing words using or elimination of spaces within or between words ([tadartino] [tadart ino] (my house)).

3.2 Phonology and phonetic

The main problem of Amazigh phonology and phonetic consists on allophones. This problem depends particularly on the regional varieties, when a single phoneme realized in different ways, such as /ll/ that is realized as [dj] in the North.

3.3 Amazigh morphology

An additional reason for the difficulties of computational processing of the Amazigh language is its rich and complex morphology. Inflectional processes in Amazigh are based primarily on both prefix and suffix concatenations. Furthermore, the base form itself can be modified in different paradigms such as the derivational one. Where in case of the presence of geminated letter in the base form, this later will be altered in the derivational form (qqim → svim (make sit)).

4. The State of Amazigh language technology

In the context of promoting Amazigh language, many works have been done to provide this language with linguistic resources and tools in the aim to enable its automatic processing and its integration in the field of Information and Communication Technology. In this section, we describe existing works on Amazigh language processing.

4.1 Tifinaghe Encoding

Over several years, the Amazigh language has been writing in Latin alphabet supported by diacritics and phonetic symbols, or in Arabic script. While after adopting Tifinaghe as an official script in Morocco, the Unicode encoding of this script was became a necessity. To this end considerable efforts have been invested. However, this process took ample time to be done, which required the use of ANSI encoding as a first step to integrate the Amazigh language into the educational system at time.

Considering Tifinaghe variants used in all parts of the Amazigh world, the Unicode encoding is composed of four character subsets: the basic set of IRCAM, the extended IRCAM set, other Neo-Tifinaghe letters in use, and modern Touareg letters. The two first subsets constitute the sets of characters chosen by IRCAM. While, the first is used to arrange the orthography of different Moroccan Amazigh dialects, the second subset is used for historical and scientific use. The letters are classified in accordance with the order specified by IRCAM. Other Neo-Tifinaghe and Touareg letters are interspersed according to their pronunciation. Thus, the UTC accepts the 55 Tifinaghe characters for encoding in the range U+2D30..U+2D65, U+2D6F, with Tifinaghe block at U+2D30..U+2D7F (Andries, 2008).

4.2 Optical character recognition

In the aim to achieve perfection on Amazigh optical character recognition systems many studies have been undertaken using different approaches. Most of these approaches have achieved a recognition rate around 92%. In the following, we present briefly some Amazigh optical character recognition systems. (Es Saady et al., 2011) focused on isolated printed characters recognition based on a syntactic approach using finite automata. (Amrouch et al., 2010) proposed a global approach based on Hidden Markov Models for recognizing handwritten characters. (El Ayachi et al., 2010) presented a method using invariant moments for recognizing printed script. (Ait Ouguengay et al., 2009) proposed an artificial neural network approach to recognize printed characters.

4.3 Fundamental processing tools

In this section, we describe natural language processing systems that have been developed for the Amazigh language at the Royal Institute of Amazigh Culture.

- **Transliterator:** The Amazigh language has known through its existence different forms of writing: Latin alphabet supported by diacritics and phonetic symbols, Arabic script, and Tifinaghe character based on ANSI and Unicode encoding. In the aim to facilitate the passage from one form to another, and to convert all writing prescriptions into a standard unique form in order to simplify the text processing a transliterator tool has been developed (Ataa Allah and Boulaknadel, 2011).

- **Tagging assistance tool:** The use of corpora in natural language processing, especially those annotated morphosyntactically, has become an indispensable step in the language tools' production and in the process of language computerization. In this context, we have lead to build a morphosyntactic corpus; which has elicited the development of a tool, providing support and linguists' assistance (Ataa Allah and Jaa, 2009).

- **Stemmer:** To enhance the performance of information retrieval systems for the Amazigh language a computationally stemming process was realized. This process consists in splitting Amazigh words into constituent stem part and affix parts without doing complete morphological analysis, in order to conflate word variants into a common stem (Ataa Allah and Boulaknadel, 2010-a).

- **Search engine:** As the number of Amazigh documents grew, searching algorithms have become one of the most essential tools for managing information of Amazigh documents. Thus, a first attempt has been proposed in order to develop a search engine that could support the Amazigh language characteristics. The proposed search engine is designed to crawl and index the Amazigh web pages written in Tifinaghe. Moreover, it is based on some natural language processing such as stop words removal and light stemming in retrieval task (Ataa Allah and Boulaknadel, 2010-b).

- **Concordancer:** Amazigh linguistics corpora are currently enjoying a surge activity. As the growth in the number of available Amazigh corpora continues, there is an increased need for robust tools that can process this data, whether it is for research or teaching. One such tool that is useful for both groups is the concordancer, which is a simple tool for displaying a specified target word in its context. However, obtaining one that can reliably support all Moroccan Amazigh language scripts has proved an extreme difficulty. In this aim, an online concordancer was developed (Ataa Allah and Boulaknadel, 2010-c).

4.4 Language resources

Natural language processing is showing more interest in the Amazigh language in recent years. Suitable resources for Amazigh are becoming a vital necessity for the progress of this research. In this context some efforts are currently underway.

- **Corpora:** Corpora are a very valuable resource for NLP tasks, but the Amazigh language lacks such resources. Therefore, researchers at IRCAM have tried to build an Amazigh corpora in progressive way until reaching a large-scale corpus that follows TREC's standards. Thus, two parallel works are undertaking (Outahajala et al., 2010; Boulaknadel and Ataa Allah, 2011).

The first consists in building a general corpus based on texts dealing with different literary genres: novels, poems, stories, newspaper articles, and covering various topics. While the second is based on POS tagged data that was collected from IRCAM's newspapers, websites and pedagogical supports.

- **Dictionary:** Although many paper dictionaries are available for the Amazigh language, none of them is computational. To deal with this lack, an application that is helping in collecting and accessing Amazigh words has been elaborated (Iazzi and Outahajala, 2008). This application has provided all necessary information such as definition, Arabic French and English equivalent words, synonyms, classification by domains, and derivational families.

- **Terminology database:** While the Amazigh language is given new status, it becomes necessary, even inevitable to own a terminology covering the largest number of lexical fields. Thus, a tool managing terminology database has been developed to facilitate the work of researchers allowing an efficient exploitation of users. This tool allows the processing of new terminology data, the compilation, and the management of existing terminology (El Azrak and El Hamdaoui, 2011).

5. Conclusion and Future Directions

In this paper, we discussed the main challenges in processing the Amazigh language, and we attempted to survey the research work on Amazigh NLP in Morocco. In the aim to convert Amazigh language from a less resourced language into a resourced, studied language from computational point of view, we need to expedite the basic research on Amazigh NLP tools development by addressing the following issues:

-Building a large and representative Amazigh corpus which will be helpful for spelling and grammar checking, speech generation, and many other related topics.

-Developing a machine translation system which will immensely contribute to promote and disseminate the Amazigh language.

-Creating a pool of competent human resources to carry out research work on Amazigh NLP by offering scholarship for higher degrees and attracting young researchers with attractive salary.

6. References

Ait Ouguengay Y. (2007). Quelques aspects de la numérisation des polices de caractères : Cas de Tifinaghe. *La typographie entre les domaines de l'art et de l'informatique*. Rabat, Maroc, pp. 159--181.

Ait Ouguengay Y., Taalabi M. (2009). Elaboration d'un réseau de neurones artificiels pour la reconnaissance optique de la graphie amazighe: Phase d'apprentissage, *Systèmes intelligents-Théories et applications*. Europia productions.

Andries P. (2008). *Unicode 5.0 en pratique, Codage des caractères et internationalisation des logiciels et des documents.* Dunod, France, Collection InfoPro.

Ameur M., Bouhjar A., Boukhris F., Boukous A., Boumalk A., Elmedlaoui M., Iazzi E., Souifi H. (2004). *Initiation à la langue amazighe*. IRCAM, Rabat, Maroc.

Ameur M., Bouhjar A., Boukhris F., Boumalk A., Elmedlaoui M., Iazzi E. (2006). *Graphie et orthographe de l'amazighe*. IRCAM, Rabat, Maroc.

Ameur M., Bouhjar A., Boukhris F., Elmedlaoui M., Iazzi E. (2006). *Vocabulaire de la langue amazighe (Français-Amazighe).* série : Lexiques N°1, IRCAM, Rabat, Maroc.

Ameur M., Bouhjar A., Boumalk A., El Azrak N., Laabdelaoui R. (2009). *Vocabulaire des médias (Français-Amazighe-Anglais-Arabe).* série : Lexiques N°3, IRCAM, Rabat, Maroc.

Ameur M., Bouhjar A., Boumalk A., El Azrak N., Laabdelaoui R. (2009). *Vocabulaire grammatical*. série : Lexiques N°5, IRCAM, Rabat, Maroc.

Amrouch M., Rachidi A., El Yassa M., Mammass D. (2010). Handwritten Amazigh Character Recognition Based On Hidden Markov Models. *International Journal on Graphics, Vision and Image Processing*. 10(5), pp.11--18.

Ataa Allah F., Boulaknadel S. (2010). Amazigh Search Engine: Tifinaghe Character Based Approach. In *Proceeding of International Conference on Information and Knowledge Engineering*, Las Vegas, Nevada, USA, pp. 255-259.

Ataa Allah F., Boulaknadel S. (2010). Pseudo-racinisation de la langue amazighe. In *Proceeding of Traitement Automatique des Langues Naturelles*. Montréal, Canada.

Ataa Allah F., Boulaknadel S. (2010). Online Amazigh Concordancer. In *Proceedings of International Symposium on Image Video Communications and Mobile Networks*. Rabat, Maroc.

Ataa Allah F., Boulaknadel S. (2011). Convertisseur pour la langue amazighe : script arabe - latin – tifinaghe. In *Proceedings of the 2ème Symposium International sur le Traitement Automatique de la Culture Amazighe*. Agadir, Marocco, pp. 3--10.

Ataa Allah F., Jaa H,. (2009). Etiquetage morphosyntaxique : Outil d'assistance dédié à la langue amazighe. In *Proceedings of the 1er Symposium international sur le traitement automatique de la culture amazighe*, Agadir, Morocco, pp. 110- -119.

Boukhris F., Boumalk A., Elmoujahid E., Souifi H. (2008). *La nouvelle grammaire de l'amazighe*, IRCAM, Rabat, Maroc.

Boulaknadel S. (2009). Amazigh ConCorde: an appropriate concordance for Amazigh. In *Proceedings of the 1er Symposium international sur le traitement automatique de la culture amazighe*, Agadir, Morocco, pp. 176--182.

Boulaknadel S., Ataa Allah F. (2011). Building a standard Amazigh corpus. In *Proceedings of the International Conference on Intelligent Human Computer Interaction*. Prague, Tchec.

EL Azrak N., EL Hamdaoui A. (2011). Référentiel de la Terminologie Amazighe : Outil d'aide à l'aménagement linguistique. In *Proceedings of theème atelier international sur l'amazighe et les TICs*, Rabat, Morocco.

El Yachi R., Moro K., Fakir M., Bouikhalene B. (2010). On the Recognition of Tifinaghe Scripts. *Journal of*

Theoretical and Applied Information Technology, 20(2), pp. 61--66.

Es Saady Y., Ait Ouguengay Y., Rachidi A., El Yassa M., Mammass D. (2009). Adaptation d'un correcteur orthographique existant à la langue Amazighe : cas du correcteur Hunspell. In *Proceedings of the 1er Symposium International sur le Traitement Automatique de la Culture Amazighe.* Agadir, Morocco, pp. 149--158.

Es Saady Y., Rachidi A., El Yassa M., Mammas D. (2010). Printed Amazigh Character Recognition by a Syntactic Approach using Finite Automata. *International Journal on Graphics, Vision and Image Processing*, 10(2), pp.1--8.

Fakir M., Bouikhalene B., Moro K. (2009). Skeletonization methods evaluation for the recognition of printed tifinaghe characters. In *Proceedings of the 1er Symposium International sur le Traitement Automatique de la Culture Amazighe*. Agadir, Morocco, pp. 33--47.

Greenberg J. (1966). *The Languages of Africa*. The Hague.

Iazzi E., Outahajala M. (2008). Amazigh Data Base. In *Proceedings of HLT & NLP Workshop within the Arabic world: Arabic language and local languages processing status updates and prospects.* Marrakech, Morocco, pp. 36--39.

Ouakrim O. (1995). Fonética y fonología del Bereber, *Survey at the University of Autònoma de Barcelona*.

Outahajala M. (2007). Les normes de tri, Du clavier et Unicode. *La typographie entre les domaines de l'art et de l'informatique*. Rabat, Morocco, pp. 223--237.

Outahajala M., Zekouar L., Rosso P., Martí M.A. (2010). Tagging Amazigh with AnCoraPipe. In *Proceeding of the Workshop on Language Resources and Human Language Technology for Semitic Languages*. Valletta, Malta, pp. 52--56.

Zenkouar L. (2008). Normes des technologies de l'information pour l'ancrage de l'écriture amazighe. *Etudes et documents berbères*. 27, pp. 159--172.

Compiling Apertium morphological dictionaries with HFST and using them in HFST applications

Tommi A Pirinen, Francis M. Tyers

University of Helsinki, Universitat d'Alacant
FI-00014 University of Helsinki Finland, E-03071 Alacant Spain
tommi.pirinen@helsinki.fi, ftyers@dlsi.ua.es

Abstract

In this paper we aim to improve interoperability and re-usability of the morphological dictionaries of Apertium machine translation system by formulating a generic finite-state compilation formula that is implemented in HFST finite-state system to compile Apertium dictionaries into general purpose finite-state automata. We demonstrate the use of the resulting automaton in FST-based spell-checking system.

Keywords: finite-state, dictionary, spell-checking

1. Introduction

Finite-state automata are one of the most effective format for representing natural language morphologies in computational format. The finite-state automata, once compiled and optimised via process of minimisation are very effective for parsing running text. This format is also used when running morphological dictionaries in machine-translation system Apertium (Forcada et al., 2011)[1]. In this paper we propose a generic compilation formula to compile the dictionaries into weighted finite state automata for use with any FST tool or application. We implement this system using a free/libre open-source finite-state API HFST (Lindén et al., 2011)[2]. HFST is a general purpose programming interface using a selection of freely-available finite-state libraries for the handling of finite-state automata. While Apertium uses the dictionaries and the finite-state automata for machine translation, HFST is used in multitude of other applications ranging from basic morphological analysis (Lindén et al., 2011) to end-user applications such as spell-checking (Pirinen and Lindén, 2010) and predictive text-entry for mobile phones (Silfverberg et al., 2011). In this article we show how to generate automatically a spell-checker from an Apertium dictionary and evaluate roughly the usability of the automatically generated spell-checker.

The rest of the article is laid out as follows: In section 2. we describe the generic compilation formula for the HFST-based compilation of Apertium dictionaries and the formula for induction of spell-checkers error model from Apertium's dictionary. In section 3. we introduce the Apertium dictionary repository and the specific dictionaries we use to evaluate our systems. In section 4. we evaluate speed and memory usage of compilation and application of our formula against Apertium's own system and show that our system has roughly same coverage and explain the differences arise from.

2. Methods

The compilation of Apertium dictionaries is relatively straight-forward. We assume here standard notations for finite-state algebra. The morphological combinatorics of Apertium dictionaries are defined in following terms: There is one set of root morphs (finite strings) and arbitrary number of named sets of affix morphs called `pardefs`. Each set of affix morphs is associated with a name. Each morph can also be associated with a paradigm reference pointing to a named subset of affixes. As an example, a language of singular and plural of *cat* and *dog* in English would be described by root dictionary consisting of morphs `cat` and `dog`, both of which point on the right-hand side to pardef named `number`. The number affix morphs are defined then as set of two morphs, namely `s` for plural marker and empty string for singular marker.

Each morph can be compiled into single-path finite-state automaton[3] containing the actual morph as string of UTF-8 arcs m. The morphs in the root dictionary are extended from left or right sides by joiner markers iff they have a pardef definition there and each affix dictionary is extended on the left (for suffixes) or right (for prefixes) by the pardef name marker. In the example of *cats, dogs* language this would mean finite state paths `c a t NUMBER`, `d o g NUMBER`, `NUMBER s` and `NUMBER` ϵ, where ϵ as usual marks zero-length string[4]. These sets of roots and affixes can be compiled into disjunction of such joiner delimited morphs. Now, the morphotactics can be defined as related to joiners by any such path that contains joiners only as pairs of adjacent identical paradigm references, such as `c a t NUMBER NUMBER s` or `d o g NUMBER NUMBER` ϵ, but not `c a t NUMBER d o g NUMBER` or `NUMBER s NUMBER`

[1] http://www.apertium.org

[2] http://hfst.sf.net

[3] the full formula allows any finite-state language as morph, compiled from regular expressions, the extension to this is trivial but for readability we present the formula for string morphs

[4] In the current implementation we have used temporarily a special non-epsilon marker as this decreases the local indeterminism and thus compilation time

s. The finite-state formula for this morphotactics is defined by

$$M_x = (\Sigma \cup \bigcup_{x \in p} xx)^\star, \tag{1}$$

where p is set of pardef names and Σ the set of symbols in morphs not including the set of pardef names. Now the final dictionary is simply composition of these morphotactic rules over the repetion of affixes and roots:

$$(M_a \cup M_r)^\star \circ M_x, \tag{2}$$

where M_a is the disjunction of affixes with joiners, M_r the disjunction of roots with joiners, and M_x the morphotactics defined in formula 1. This is a variation of morphology compilation formula presented in various HFST documentation, such as (Lindén et al., 2011).

2.1. Implementation Details

There are lot of finer details we will not thoroughly cover in this article, as they are mainly engineering details. In this section we shortly summarise specific features of HFST-based FST compilation that result in meaningful differences in automaton structure or working. One of the main source of differences is that HFST automata are two-sided and compiled only ones from the source code whereas Apertium generates two different automata for analysis and generation. In these automata the structure may be different, since Apertium dictionaries have ways of marking morphs limited to generation or analysis only, so they will only be included in one of the automatons. Our approach to this is to use special symbols called flag-diacritics (Beesley and Karttunen, 2003) to limit the paths as analysis only or generation only on runtime, but still including all paths in the one transducer that gets compiled.

Another main difference in processing comes from the special word-initial, word-final and separate morphs that in Apertium are contained in separate automata altogether, but HFST tools do not support use of multiple automata for analysis, so these special morphs will be concatenated optionally to beginning or end of the word, or disjuncted to the final automata respectively. These special morphs include things like article *l'* in French as bound form.

2.2. Creating a Spell-Checker Automatically

To create a finite-state spell-checker we need two automata, one for the language model, for which the dictionary compiled as described earlier will do, and one for the error model (Pirinen and Lindén, 2010). A classic baseline error model is based on the edit distance algorithm (Levenshtein, 1966; Damerau, 1964), that defines typing errors of four types: pressing extra key (insertion), not pressing a key (deletion), pressing wrong key (change) and pressing two keys in wrong order (swap). There have been many finite-state formulations of this, we use the one defined in (Schulz and Mihov, 2002; Pirinen and Lindén, 2010). The basic version of this where the typing errors of each sort have equal likelihood for each letters can be induced from the compiled language model, and this is what we use in this paper. The induction of this model is relatively straightforward; when compiling the automaton, save each unique UTF-8 codepoint found in the morphs[5]. For each character generate the identities in start and end state to model correctly typed runs. For each of the error types the generate one arc from initial state to the end state modelling that error, except for swap which it requires one auxiliary state for each character pair.

3. Materials

The Apertium project hosts a large number of morphological dictionaries for each of the languages translated. From these we have selected three dictionaries to be tested: Basque from Basque-Spanish pair as it is released dictionary with the biggest on-disk size, Norwegian Nynorsk from the Norwegian pair as a language that has some additional morphological complexity, such as compounding, and Manx from as a language that currently lacks spell-checking tools to demonstrate the plausibility of automatic conversion of Apertium dictionary into a spell-checker[6].

To evaluate the use of resulting morphological dictionaries and spell-checkers we use following Wikipedia database dumps[7]: `euwiki-20120219-pages-articles.xml.bz2`, `nnwiki-20120215-pages-articles.xml.bz2`, and `gvwiki-20120215-pages-articles.xml.bz2`. For the purpose of this article we performed very crude cleanup and preprocessing to Wikipedia data picking up the text elements of the article and discarding most of Wikipedia markup naïvely[8].

4. Test Setting and Evaluation

To get one view on differences made by generic compilation formula instead of direct automata building used by Apertium we look at the created automata, this will also give us a rough idea of what its efficiency might be. In table 1 we give the counts of nodes and edges, in that order, in the graphs compiled from the dictionaries. Note, that in case of Apertium it is the sum of all the separate automata states and edges that is counted. The small differences in sizes of graphs are mostly caused by the different handling of generation vs. analysis mode. The difference in sizes of automata on disk in is shown in table 2. The size of HFST au-

[5] The description format of Apertium requires declaration of exemplar character set as well, but as this is only used in the tokenisation algorithm (Garrido-Alenda et al., 2002) , which is not going to be used, we induce the set from the morphs

[6] We also provide a Makefile script to recreate results of this article for any language in Apertium's repository

[7] `http://download.wikipedia.org/`

[8] For details see the script in `http://hfst.svn.sourceforge.net/viewvc/hfst/trunk/lrec-2011-apertium/`.

tomata can be attributed to the clever compression algorithm used by HFST (Silfverberg and Lindén, 2009).

Lang.	Apertium LR	Apertium RL	HFST
Basq.	30,114	34,005	34,824
	59,321	68,030	68,347
Norg.	56,226	55,722	56,871
	138,217	132,475	139,259
Manx	13,055	12,955	12,920
	28,220	27,062	27,031

Table 1: Size of HFST-based system against original (count of nodes first, then edges)

Lang.	Apertium LR	Apertium RL	HFST
Basq.	252 KiB	289 KiB	1,7 MiB
Norg.	558 KiB	535 KiB	3,7 MiB
Manx	108 KiB	110 KiB	709 KiB

Table 2: Size of HFST-based system against original (as B on disk)

To test efficiency we measure times of running various tasks. The times and memory usage have been measured using GNU `time` utility and `getrusage` system call's `ru_utime` field, averaged over three test runs. The tests were performed on quad-core Intel Xeon E5450 @ 3.00 GHz with 64 GiB of RAM.

First we measure speed of analysing a full corpus with the result automaton. The speed is measured in the table 3, in seconds to precision that was available in our system. Curiously the results do not give direct advantage to either of the system but it seems to depend on the language which system is a better choice for corpus analysis.

Language	Apertium	HFST
Basque	32.0 s	18.4 s
Norwegian	2.4 s	5.5 s
Manx	1.6 s	2.2 s

Table 3: Speed of HFST-based system against original in corpus analysis (as s in user time)

Similarly we measure the speed of current compilation process in table 4. In here there's an obvious advantage to manual building of the automaton (see (Rojas et al., 2005) for the precise algorithm used) over the finite-state algebra method, as is in line with earlier results for lexc building in (Lindén et al., 2009).

Language	Apertium time	HFST time
Basque	35.7 s	160.0 s
Norwegian	6.6 s	200.2 s
Manx	0.8 s	11.2 s

Table 4: Speed of HFST-based system against original in compilation (as seconds of user time)

Finally we evaluate the usability of dictionaries meant for machine translation as spell-checkers by running the finite-state spell checkers we produced automatically through a large corpus and show the measure both speed and quality of the results. The errors were automatically generated to Wikipedia text's correct words using simple algorithm that may generate one Levenshtein error per each character position at probability of $\frac{1}{33}$. This test shows only rudimentary results on the plausibility of using machine translation dictionary for spell-checking; for more thorough evaluation of efficiency of finite-state spell-checking see (Hassan et al., 2008).

Language	Speed (words/sec)
Basque	7,900
Norwegian	9,200
Manx	4,700

Table 5: Efficiency of spelling correction in artificial test setup, average over three runs.

5. Conclusions

In this article we have shown a general formula to compile morphological dictionaries from machine-translation system Apertium in generic FST system of HFST and using the result in HFST-based application of spell-checking.

6. Future Work

In this article we showed a basic method to gain more inter-operability between generic FST system of HFST and a specialised morphological dictionary writing formalism of machine-translation system Apertium by implementing a generic compilation formula to compile the language descriptions. In future research we are leveraging this and other related formulas into automatic optimisation of the final automata using the information present in the language description to optimise instead of relying generic graph algorithms for the final minimised result automata.

We demonstrated importing the compiled dictionary as a language model and inducing error model for real-world spell-checking applications. Further development in this direction should aim for interoperable formalisms, formats and mechanisms for language models and end applications of all relevant language technology tools.

Acknowledgements

We thank the HFST and Apertium contributors for fruitful internet relayed chats, and the two anonymous reviewers for their helpful suggestions.

7. References

Kenneth R Beesley and Lauri Karttunen. 2003. *Finite State Morphology.* CSLI publications.

F J Damerau. 1964. A technique for computer detection and correction of spelling errors. *Commun. ACM*, (7).

Mikel L. Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M. Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, jul.

Alicia Garrido-Alenda, Mikel L. Forcada, and Rafael C. Carrasco. 2002. Incremental construction and maintenance of morphological analysers based on augmented letter transducers. In *Proceedings of TMI 2002 (Theoretical and Methodological Issues in Machine Translation, Keihanna/Kyoto, Japan)*, pages 53–62.

Ahmed Hassan, Sara Noeman, and Hany Hassan. 2008. Language independent text correction using finite state automata. In *Proceedings of the Third International Joint Conference on Natural Language Processing*, volume 2, pages 913–918.

V. I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics—Doklady 10, 707–710. Translated from Doklady Akademii Nauk SSSR*, pages 845–848.

Krister Lindén, Miikka Silfverberg, and Tommi Pirinen. 2009. Hfst tools for morphology—an efficient open-source package for construction of morphological analyzers. In Cerstin Mahlow and Michael Piotrowski, editors, *sfcm 2009*, volume 41 of *Lecture Notes in Computer Science*, pages 28—47. Springer.

Krister Lindén, Miikka Silfverberg, Erik Axelson, Sam Hardwick, and Tommi Pirinen, 2011. *HFST—Framework for Compiling and Applying Morphologies*, volume Vol. 100 of *Communications in Computer and Information Science*, pages 67–85. Springer.

Tommi A Pirinen and Krister Lindén. 2010. Finite-state spell-checking with weighted language and error models. In *Proceedings of the Seventh SaLTMiL workshop on creation and use of basic lexical resources for less-resourced languagages*, pages 13–18, Valletta, Malta.

Sergio Ortiz Rojas, Mikel L. Forcada, and Gema Ramírez Sánchez. 2005. Construcción y minimización eficiente de transductores de letras a partir de diccionarios con paradigmas. *Procesamiento del Lenguaje Natural*, (35):51–57.

Klaus Schulz and Stoyan Mihov. 2002. Fast string correction with levenshtein-automata. *International Journal of Document Analysis and Recognition*, 5:67–85.

Miikka Silfverberg and Krister Lindén. 2009. Hfst runtime format—a compacted transducer format allowing for fast lookup. In Bruce Watson, Derrick Courie, Loek Cleophas, and Pierre Rautenbach, editors, *FSMNLP 2009*, 13 July.

Miikka Silfverberg, Mirka Hyvärinen, and Tommi Pirinen. 2011. Improving predictive entry of finnish text messages using irc logs. pages 69–76.

Automatic structuring and correction suggestion system for Hungarian clinical records

Borbála Siklósi, György Orosz, Attila Novák, Gábor Prószéky
Pázmány Péter Catholic University Faculty of Information Technology
H-1083 Budapest, Práter street 50/a
E-mail: siklosi.borbala@itk.ppke.hu, oroszgy@itk.ppke.hu, novak.attila@itk.ppke.hu, proszeky@itk.ppke.hu

Abstract

The first steps of processing clinical documents are structuring and normalization. In this paper we demonstrate how we compensate the lack of any structure in the raw data by transforming simple formatting features automatically to structural units. Then we developed an algorithm to separate running text from tabular and numerical data. Finally we generated correcting suggestions for word forms recognized to be incorrect. Some evaluation results are also provided for using the system as automatically correcting input texts by choosing the best possible suggestion from the generated list. Our method is based on the statistical characteristics of our Hungarian clinical data set and on the HUMor Hungarian morphological analyzer. The conclusions claim that our algorithm is not able to correct all mistakes by itself, but is a very powerful tool to help manually correcting Hungarian medical texts in order to produce a correct text corpus of such a domain.

Keywords: spelling correction, clinical text mining, language models, morphology, agglutinative, biomedical corpora

1. Introduction

In most hospitals medical records are only used for archiving and documenting a patient's medical history. Though it has been quite a long time since hospitals started using digital ways for written text document creation instead of handwriting and they have produced a huge amount of domain specific data, they later use them only to lookup the medical history of individual patients. Digitized records of patients' medical history could be used for a much wider range of purposes. It would be a reasonable expectation to be able to search and find trustworthy information, reveal extended knowledge and deeper relations. Language technology, ontologies and statistical algorithms make a deeper analysis of text possible, which may open the prospect of exploration of hidden information inherent in the texts, such as relations between drugs and other treatments and their effects. However, the way clinical records are currently stored in Hungarian hospitals does not even make free text search possible, the look-up of records is only available referring to certain fields, such as the name of the patient.

Aiming at such a goal, i.e. implementing an intelligent medical system requires a robust representation of data. This includes well determined relations between and within the records and filling these structures with valid textual data. In this paper we describe how the structure of the medical records is established and the method of automatic transformation. Basic links between individual records are also recognized, such as medical prehistory of a patient. Then, after the elimination of non-textual data, we demonstrate a basic method for correcting spelling errors in the textual parts with an algorithm that is able to handle both the language and domain specific phenomena.

2. Representation of medical texts

We were provided anonymized clinical records from various departments, we chose one of them, i.e. ophthalmology to build the system that can be extended later to other departments as well. The first phase of processing raw documents is to compensate the lack of structural information. Due to the lack of a sophisticated clinical documentation system, the structure of raw medical documents can only be inspected in the formatting or by understanding the actual content. Besides basic separations - that are not even unified through documents - there were no other aspects of determining structural units. Moreover a significant portion of the records were redundant: medical history of a patient is sometimes copied to later documents at least partially, making subsequent documents longer without additional information regarding the content itself. However these repetitions will provide the base of linking each segment of a long lasting medical process.

2.1 XML structure

Wide-spread practice for representing structure of texts is to use XML to describe each part of the document. In our case it is not only for storing data in a standard format, but also representing the identified internal structure of the texts which are recognized by basic text mining procedures, such as transforming formatting elements to structural identifiers or applying recognition algorithms for certain surface patterns. After tagging the available metadata and performing these transformations the structural units of the medical records are the followings:

- the *whole copy* in its original form of the document is stored to be used at later stages.
- *content*: parts of the records that are in free text form are further divided into sections such as header, diagnoses, applied treatments, status,

operation, symptoms, etc.

- *metadata*: applying basic pattern recognition methods we automatically tagged such units as the type of the record, name of the institution and department where it was written, diagnoses represented in tabular forms and standard encodings of health related concepts.
- *simple named entities*: at this stage of our work we only tagged basic named entities, such as dates, doctors, operations, etc. The medical language is very sensitive to named entities, that is why handling them requires much more sophisticated algorithms, which are a matter of further research.
- *medical history*: with the help of repeated sections of medical records related to one certain patient, we have been able to build a simple network of medical processes. Since the documentation of medical history is not standardized and not consequent even for the same patient, the correspondence is determined by partial string matching and comparing algorithms. Thus we can store the identifiers of the preceding and following records.

raw medical records (ophthalmology)	6.741.435
relevant content, marked with the tag content	1.452.216
textual data in content parts	422.611

Table 1: Size of each resource
(measured in number of tokens).

2.2 Separating textual and non-textual data

The resulting structure defines the separable parts of each record; however there are still several types of data within these structural units. Thus it is not possible to apply standard text processing methods on such noisy texts. Such non-textual information inserted into free word descriptions are laboratory test results, numerical values, delimiting character series and longer chains of abbreviations and special characters. We filtered out these expressions to get a set of records containing only natural text, making it possible to use natural language processing algorithms for preprocessing. Since non-textual fragments were quite diverse especially in documents originating from different doctors, or even assistants, it was impossible to develop a robust rule-based algorithm to recognize them. To solve this issue we applied the unsupervised methods of clustering algorithms. The first assumption was to tokenize documents into sentences, however due to the domain specific behaviour and the non-well-formed written representation of the texts, there are hardly any sentence boundaries in the classical sense. Our basic units were lines (i.e. units separated with newline character) and concatenations of multiple lines where neighbouring lines were suspected to be continuation of each other. This continuation does not apply to the semantic content of the lines, rather to their behaviour regarding textual or non-textual form of information. We concatenated two lines if the end of the line was a non-sentence closing punctuation mark and the beginning of the following line was not a capital letter and not a number or if the end of the first line was an inner sentence punctuation mark. Thus such short textual fragments were kept together with more representative neighbours avoiding them to be filtered out by themselves, since their feature characteristics are very similar to those of non-textual lines. We applied k-means clustering to these concatenated lines. The goal was to split into k=2 groups, however this proved to be inefficient and could not be improved by modifying the feature set. Thus we applied k=7 clustering, where two groups were flowing texts and five were different types of non-textual fragments. The labeling of the seven groups were done offline by hand, however applying a classifier trained on these data is able to recognize and separate new sets of data automatically. Testing the efficiency of our feature set and clustering algorithm, a simple Naive Bayes classifier performed 98% accuracy on a data set of 100 lines. Portions considered to be textual information need to be normalized in terms of punctuation, spelling and the used abbreviations. A fault tolerant tokenization is applied to the running text that takes into account domain specific phenomena.

text	Zavartalan korai posztoperatív szakot követően otthonába bocsátjuk, ahol javasolt kímélő életmód mellett naponta 5x1 Tobradex (tobramycin, dexamethasone) szemcsepp alkalmazása az operált szembe. Kontroll vizsgálat megbeszélés szerint, 2010. jún. 09.-n délelőtt, Dr.Benedek Szabolcsnál klinikánk ambulanciáján, illetve panasz esetén azonnal.
non-text	V.: 0,63 +0,75 Dsph -1,00Dcyl 180° = 0,8 V: 1.0 -0.5 Dsph-al élesebb V közeli: +1.5 Dsph -al Csapody III.

Table 2: Examples of textual and non-textual fragments.

3. Spelling correction

3.1 Language and domain specific difficulties

Research in the field of clinical records processing have advanced considerably in the past decades and similar applications exist for records written in English, however, these tools are not readily applicable to other languages. In our case, the problem is not only that Hungarian is another language, but agglutination and compounding, which yield a huge number of different word forms and free word order in sentences render solutions applicable to English unfeasible. E.g. while the number of different word tokens in a 10 million word English corpus is generally below 100,000, in Hungarian it is well above 800,000. However, the 1:8 ratio does not correspond to the ratio of the number of possible word forms between the two languages: while there are about 4–5 different inflected forms for an English word, there are about a 1000 for Hungarian, which indicates that a corpus of the same size is much less representative for Hungarian than it is for English (Oravecz et al., 2002.).

Moreover, medical language contains additional difficulties. Since these records are not written by clinical experts (and there is no spell-checker in the software they use) they contain many errors of the following types:

- typing errors occurring during text input mainly by accidentally swapping letters, inserting extra letters or just missing some,
- the misuse of punctuation,
- substandard spelling with especially many errors arising from the use of special medical language with a non-standard spelling that is a haphazard mixture of what would be the standard Latin and Hungarian spelling (e.g. tension / tenzió / tenzio / tensió). Though there exists a theoretical standard for the use of such medical expressions, doctors tend to develop their own customs and it is quite difficult for even an expert to choose the right form.

Besides these errors, there are many additional difficulties that must be handled in a text mining system, which are also consequences of the special use of the language. When writing a clinical record, doctors or assistants often use short incomplete phrases instead of full sentences. The use of abbreviations does not follow any standards in the documents. Assistants do not only use standard abbreviations but abbreviate many common words as well in a rather random manner and abbreviations rarely end in a period as they should in standard orthography. Moreover, the set of abbreviations used is domain specific and also varies with the doctor or assistant typing the text. In some extreme situations it might happen that a misspelled word in one document is an intentional abbreviation or short form in the other.

For the identification of an appropriate error model of the spelling errors, a corpus of corrected clinical records is needed. There is no such corpus at all for Hungarian medical language, thus we needed to create a corrected version of our real-life medical corpus. This was necessarily a partly manual process for a subset of the corpus, but we wanted to make the correction process as efficient as possible. Our goal was to recognize misspelled word forms and automatically present possible corrections in a ranked order. Additional algorithms with manual validation could then choose the final form, which is much easier than correcting the whole corpus by hand, moreover the baseline system might be easily extended to be able to carry out the whole process trained on the already corrected corpus.

3.2 Combination of language models

Aiming at such a goal, a simple linear model was built to provide the most probable suggestions for each misspelled word. We combined several language models built on the original data set and on external resources, that are the followings (the first two used as prefilters before suggesting corrections, the rest were used for generating the suggestions):

- *stopword list*: a general stopword list for Hungarian was extended with the most common words present in our medical corpus. After creating a frequency list, these words were manually selected.
- *abbreviation list*: after automatically selecting possible abbreviations in the corpus, the generated list was manually filtered to include the possible abbreviations. Since we have not applied expert knowledge, this list should be more sophisticated for further use.
- *list of word forms licensed by morphology*: those word forms that are accepted by our Hungarian morphology (HUMor (Prószéky et al. 2005.)) were selected from the original corpus, creating a list of potentially correct word forms. To be able to handle different forms of medical expressions, the morphology was extended with lists of medicine names, substances and the content of the Hungarian medical dictionary. We built a unigram model from these accepted word forms.
- *list of word forms not licensed by morphology*: the frequency distribution of these words were taken into consideration in two ways when generating suggestions. Ones appearing just a few times in the corpus remained as unaccepted forms (transforming their frequency value to 1 - original frequency). Those ones however, whose frequency was higher than the predefined threshold were considered to be good forms, even though they were denied by the morphology. Our assumption was that it is less possible to consequently use the same erroneous word form than being that form correct and contradicting our morphology.
- *general and domain specific corpora*: we built unigram models similar to that of the above described licensed word forms from the Hungarian Szeged Korpusz and from the descriptions of the entities in the ICD coding system documentation. We assumed that both of these corpora contains only correct word forms.

After having these models created, the text to be corrected was tokenized with a language independent tokenizer that is able to handle abbreviations keeping the punctuations and letters together as one token if necessary and is robust in this aspect. The tokenizer is insensitive for punctuation errors, at the presence of any non-alphanumeric character it creates a new token. The creation of such a tool was motivated by the special language requirements and the frequent occurrence of punctuation errors. The tokenizer uses a general list of abbreviations and the aforementioned domain specific list.

Model	size	example
stopword list	36	az
abbreviation list	1.251	alk
licensed by morphology	4.850	pupilla
not licensed by morphology	1.660	látsziuk
Szeged Korpusz	114.205	szeretnék
ICD corpus	3.209	betegségekben

Table 3: Size of each language model and resource (measured in number of tokens) with examples.

3.3 Generating possibly correct suggestions

As the next step we filtered out those word forms that are not to be corrected. These were the ones contained in the stopword and abbreviation lists. For the rest of the words the correction suggestion algorithm is applied. For each word a list of suggestion candidates is generated that contains the word forms with one unit of Levenshtein distance difference (Levenshtein, 1965) and the possible suggestions generated by the morphology. Then these candidates are ordered with a weighted linear combination of the different language models, the weight of the Levenshtein generation and the features of the original word form. Thus a weighted suggestion list is generated to all words in the text (except for the abbreviations and stopwords), but only those will be considered to be relevant, where the score of the best weighted suggestion is higher than that of the original word. At the end we considered the ten best suggestions.

4. Results

We investigated the performance of the system as a standalone automatic correcting tool, accepting the best weighted suggestion as the correction, but also as an aiding system that is only to help manual correction at this initial state. Since we did not have a correct corpus, we had to create one manually by correcting a portion of our medical corpus. Our test set contained 100 paragraphs randomly chosen from the corpus. When creating the gold standard from this set, there were disagreements even between human correctors, that is why in several cases we had to accept more than one word form as correct. The normalization of these forms is a task of further research. We used three metrics for evaluation:

- *precision*: measures how the number of properly corrected suggestions relates to the number of all corrections, considering the best weighted suggestion as correction.
- *recall*: measures the ratio of the number of properly corrected suggestions and the number of misspelled word forms in the original text.
- *f-measure*: the average of the above two

We investigated the result measures for several combinations of weighting the above described models and features:

- *Models based on justification of morphology (VOC, OOV):* since these models are the most representatives for the given corpus, these models were considered with the highest weight.
- *Models built from external resources (ICD, Szeged):* these models are bigger, but they are more general, thus word forms are not that relevant for our raw texts. Our results reflect that though these models contribute to the quality of the corrections, they should have lower weights in order to keep the scores of medical words higher.
- *Original form (ISORIG, ORIG):* the original forms of the words received two kinds of weighting. First we scored whether if the word to be corrected is licensed by the morphology or not. The second weight was given to the original word form in the suggestion list, regardless of its correctness. This was introduced so that the system would not “correct” an incorrect word form to another incorrect form, but rather keep the original one, if no real suggestions can be provided.
- *Morphological judgment on suggestions (HUMor):* each generated suggestion licensed by the morphology received a higher weight to ensure that the final suggestions are valid words.
- *Weighted Levenshtein generation (LEV):* when generating word forms that are one Levenshtein distance far from the original one, we gave special weighting for more probable phenomena, such as swapping letters placed next to each other on the keyboard of a computer (e.g.: n-m, s-d, y-z), improper use of long and short forms of Hungarian vowels (e.g.: o-ó, u-ú, ö-ő), mixing characteristic letters of Latin (e.g.: t-c, y-i).

The best combination of weights resulting in the best result for automatic correction, i.e. evaluated on the first highest scored suggestions is displayed in table 4.

Model	weights
OOV	0.05
VOC	0.25
Szeged	0.15
ICD	0.2
HUMOR	0.15
PRECISION	**70%**
RECALL	**75%**
F-MEASURE	**72%**

Table 4: Evaluation results for the best combinations of the applied models.

Example sentence 1 and correction:
A beteg intraorbitalis *implatatumot* is kapott ezért klinikánkon szeptember végén,október elején előzetes *telefonnegbeszélés* után kontrollvizsgálat javasolt.
A beteg intraorbitalis *implantatumot* is kapott ezért klinikánkon szeptember végén,október elején előzetes *telefonmegbeszélés* után kontrollvizsgálat javasolt.
Example sentence 2 and correction:
Meibm mirgy nyílások helyenként sárgás *kupakszeráűen* elzáródtak, ezeket megint *túvel* megnyitom
Meibom mirigy nyílások helyenként sárgás *kupakszerűen* elzáródtak, ezeket megint *tűvel* megnyitom

Table 5.a: Examples of automatically corrected sentences.

implatatumot	telefonnegbeszélés	Meibm	mirgy	kupakszeráűen	túvel
'implantatumot' : 5.60144363762e-05	'telefonmegbeszélés' : 5.87158540802e-05	'meibom' : 0.000105652387431	'mirigy' : 9.03702080337e-05	'kupakszerűen' : 5.87158540802e-05	'tűvel' : 5.88118697623e-05
'implatatumot' : 5.33130186722e-05	'telefonnegbeszélés' : 5.33130186722e-05	'meibm' : 5.06116009682e-05	'miragy' : 5.87158540802e-05	'kupakszervűen' : 5.87158541753e-05	'tevel' : 5.87158540802e-05
'ímplatatumot' : 1.875e-05	'telefónnegbeszélés' : 1.875e-05	'meíbm' : 1.875e-05	'mirgy' : 5.06116009682e-05	'kupakszeráűen' : 5.06116009682e-05	'tővel' : 5.87158540802e-05
'implatatumót' : 1.875e-05	'telefonnegbeszéléz' : 1.40625e-05	'meybm' : 1.40625e-05	'mírgy' : 1.875e-05	'kúpakszeráűen' : 1.875e-05	'túvel' : 5.06116009682e-05
'implatatúmot' : 1.875e-05	'telefonnegbessélés' : 1.40625e-05	'meilbm' : 4.6875e-06	'myrgy' : 1.40625e-05	'kupakszeráüen' : 1.875e-05	'tuvel' : 1.875e-05

Table 5.b: Detailed results of suggestions for the misspelled words in the above sentences.

The low numerical values in the table can be explained by several phenomena. The relatively small size of our test set does not reflect all types of errors. However manually creating a larger corrected text is very time and effort consuming. The system though provides great help in this task as well, so the evaluation of a generalized application will be much more accurate. Domain specific ambiguities also cause trouble at the time of evaluation. We allow the system to accept more than one correction as appropriate, but still there are several cases, where this is still a problem to decide. Thus the system might reject some correct forms while accepting other erroneous ones. The precise handling of abbreviations is still a problem, but is to be solved later on, thus it is unavoidable to fail on such fragments like “szemhéjszél idem, mérs. inj. conj, l.sin.” or “Vitr. o.s. (RM) abl. ret. miatt”. Human evaluation instead of the used metrics predicts much better results, which means that the readability of the texts has significantly improved.

Regarding the ranking of the suggestions, in 99.12% of the words of the test set, the 5 best suggestions contained the real correction. This means that using the system as an aiding tool for manual correction of medical texts is very powerful. An interactive user interface has been created to exploit the possibilities provided by such a feature, where the user can paste portions of medical texts, than the system highlights the words that it judged to be misspelled and offers the 5 best suggestions, from among which the user can choose. The scores are also displayed to give a hint to the user about the difference between each suggestions.

5. Further plans

The system at its early phase has several shortcomings regarding the generation and weighting of suggestions. Several problems are discussed above, besides which two more problems are to be solved in the near future. The first is that we are not yet taking into account the context of a word. This could solve some ambiguous cases, where no decision can be made on the word level. Considering the context as an affecting feature is also related to the task of deciding whether if a word form is an abbreviation or an incorrect word. The main difficulty for introducing this factor in the model is that a proper n-gram model is needed, which points back to the need of a correct corpus. The other important issue is that of multiple-word expressions. At its present stage the system is not able to correct such cases, when two words are written together without space between them, or vice versa. There is however a theoretical disagreement about such events, since several multiword expressions are used by doctors as one word expressions, though the standard would require them to use separately. Still these phenomena should be handled. As our test set contains examples for all these unhandled appearances, the evaluation metrics would surely be improved if these problems were solved.

6. Conclusion

The primary goal of developing our baseline algorithm was to aid the creation of a correct, reliable Hungarian medical text corpus. Having reached this goal, a more precise error model can be built to use for training a more improved system. As the results reflected, this motivation is fulfilled, since our correcting algorithm is quite efficient for such a basic aspect. The result of the system by itself could lead to several useful applications, such as at the background of a medical search engine, where both the query, and the actual result texts could be extended by other suggested forms of each word, making it possible to retrieve valuable information even if some misspellings are present on either side. The basic tagging and structuring described in the first part of this paper is also useful for storing, organizing and easier retrieving of the data. We demonstrated that the creation of an intelligent clinical system built on the knowledge lying in medical records is not trivial even in the preprocessing phase. However after some iterative application of the combination of automatic and manual work, a gradually improved corpus can be available, finally making the whole process automatic.

7. Acknowledgment

This work was partially supported by TÁMOP-4.2.1.B-11/2/KMR-2011-0002.

8. References

Brill, E., Moore, R.C. (2000). An improved error model for noisy channel spelling correction. *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pp. 286—293.

Contractor, D., Faruquie, T.A., Subramaniam,L.V. (2010). Unsupervised cleansing of noisy text. *Proceedings of the 23rd International Conference on Computational Linguistics*, pp. 189--196.

Farkas, R., Szarvas, Gy. (2008). Automatic Construction of rule-based ICD-9-CM coding systems. *BMC Bioinformatics*, 9

Heinze, D.T., Morsch, M.L., Holbrook, J. (2001). Mining Free-Text Medical Records. *A-Life Medical, Incorporated*, pp.254—258.

Levenshtein V. (1965). Binary codes capable of correcting spurious insertions and deletions of ones. *Problems of Information Transmission*, 1(1): pp. 8—17.

Mykowiecka, A., Marciniak, M. (2006). Domain-driven automatic spelling correction for mammography reports. *Intelligent Information Processing and Web Mining Proceedings of the International IIS: IIPWM'06. Advances in Soft Computing, Heidelberg*

Oravecz, Cs., Dienes, P. (2002). Efficient Stochastic Part-of-Speech Tagging for Hungarian. *Third International Conference on Language Resources and Evaluation*, pp. 710—717.

Patrick J., Sabbagh, M., Jain, S., Zheng, H. (2010). Spelling Correction in Clinical Notes with Emphasis on First Suggestion Accuracy. *2nd Workshop on Building and Evaluating Resources for Biomedical Text Mining*, pp. 2--8.

Pirinen, T.A., Lindén, K. (2010). Finite-State Spell-Checking with Weighted Language and Error Models – Building and Evaluating Spell-Checkers with Wikipedia as Corpus. *SaLTMiL Workshop on Creation and Use of Basic Lexical Resources for Less-Resourced Languages*, *LREC 2010*, pp.13—18.

Prószéky, G., Novák, A. (2005). Computational Morphologies for Small Uralic Languages. *Inquiries into Words, Constraints and Contexts*, pp 150—157.

Rebholz-Schuhmann, D., Kirsch, H., Gaudan, S., Arregui, M., Nenadic, G. (2005). Annotation and Disambiguation of Semantic Types in Biomedical Text: a Cascaded Approach to Named Entity Recognition. *Proceedings of the EACL Workshop on Multi-Dimensional Markup in NLP*.

Stevenson M., Guo, Y., Al Amri, A., Gaizauskas, R. (2009). Disambiguation of biomedical abbreviations. *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, pp. 71.

Constraint Grammar based Correction of Grammatical Errors for North Sámi

Linda Wiechetek

Romssa Universitehta, Norway
linda.wiechetek @ uit.no

Abstract

The article describes a grammar checker prototype for North Sámi, a language with agglutinative and inflective features. The grammar checker has been constructed using the rule-based Constraint Grammar formalism. The focus is on the setup of a prototype and diagnosing and correcting grammatical case errors, mostly those that appear with adpositions. Case errors in writing are typical even for native speakers as case errors can result from spelling mistakes. Typical candidates for spelling mistakes are forms containing the letter *á* and those with double consonants. Alternating double and single consonants is a possible case marker. Case errors in an adpositional phrase are common mistakes. Adpositions are typically homonymous (preposition, postposition, adverb) and ask for a genitive case to the left or right of it. Therefore, finding case errors requires a disambiguation of the adposition itself, a correct dependency mapping between the adposition and its dependent and a diagnosis of the case error, which can require homonymy disambiguation of the dependent itself. A deep linguistic analysis including a module for disambiguation, syntactic analysis and dependency annotation is necessary for correcting case errors in adpositional phrases.

1. Introduction

One of the challenges in grammar checking is the diagnosis and correction of morphosyntactic case. The difficulty lies in unambiguously identifying not only local, but partly global context in which a given form of a word available for e.g. case marking is erroneous. Not only needs one to identify relationships over distance (by means of dependency and valency), but also needs one to disambiguate homonymous items by means of morphosyntactic and semantic constraints. Grammar checkers differ in their approach. There are both statistically based (Atwell, 1987), and machine learning based (Izumi et al., 2003) and rule-based (Naber, 2003)[1] approaches.

While a number of grammar checkers exist for majority languages such as English, Spanish etc. a number of people have also developed grammar checking tools for minority languages. Kevin Scannell has the open-source grammar checking tools *Gramadóir* for Irish (Gaeilge), Afrikaans, Akan, Cornish, Esperanto, French, Hiligaynon, Icelandic, Igbo, Languedocien, Scottish Gaelic, Tagalog, Walloon, and Welsh. [2].

The prototype of the grammar checker for North Sámi, a morphologically complex language is written in Constraint Grammar (Karlsson, 2006), using the CG-3 version (`http://visl.sdu.dk/constraint_grammar.html`) which allows for dependency annotation and a number of other features making the analysis more efficient. There are already a number of grammar checkers written in the constraint grammar formalism, for example grammars for the Scandinavian languages, e.g. Swedish (Arppe, 2000), Danish (Bick, 2006) and Norwegian (Bokmål) (Johannessen et al., 2002) and Norwegian (Nynorsk) is being worked on `http://kaldera.no`, additionally one for Esperanto (Lundberg, 2009) and a grammar-checker module for Basque (Oronoz, 2008) dealing with postposition errors and agreement erros.

2. Errors - definition and classification

What is a grammatical error? A grammar checker usually marks errors that can be resolved by means of the morphosyntactic context. That does not only include purely grammatical errors, but also so-called physical errors, i.e. typos which are not caught by a spellchecker, but result in real wordforms, the following errors are "real-word errors". In some cases morphosyntactic context is not sufficient for the resolution of syntactic errors and semantic and lexical information is necessary. In many cases a grammar checker and a spellchecker work together. The grammar checker generally diagnoses an error and suggests a correct alternative. Depending on the target group of the grammar checker, different kinds of errors are being diagnosed by a grammar checker. While language learners (L2) usually make a lot of grammatical errors, first language (L1) users do not, and their errors are more likely to be mechanical errors (vs. cognitive errors (Miłkowski, 2010)) which result in real-word errors and morphosyntactic errors. Other typical errors are copy-paste errors, where the process of copying part of a sentence in one place and inserting it in another place results in erroneous sentence structure (Johannessen et al., 2002). The target group of the North Sámi grammar checker are first language users.

We distinguish four main errortypes, each of which has many sub-errortags. Those are lexical errors, morphosyntactical errors, syntactical errors, and real-word errors.

type	# rules	# tags
lexical	13	12
morphosyntactic	44	18
syntactic	25	18
real word	89	55
altogether	181	103

Table 1: Ruletypes

Real word errors are a common source of errors for first language users. In North Sámi, consonant and vowel lengths

[1] He is using the opene source language tool approach (`http://languagetool.wikidot.com/papers`).

[2] http://borel.slu.edu/gramadoir/manual/index.html

expressed by double consonants and diacritics (a vs. á) can lead to typos resulting in real word errors. The average homonymy between word forms in North Sámi is 2.6, but some word forms get 10 or more different analyses. Sometimes a real-word error can lead to a morphosyntactic error, e.g. where a missing consonant can result in another morphological case of a word. The line is not always easy to draw between error-types as it can be difficult to identify the "intention" behind the error. Mixing up **a** and **á** as in example (1) makes *vuosttáš* (diminutive of vuostá - cheese) 'little cheese' out of *vuosttaš* 'first'. This can be disambiguated fairly easily in a given context.

(1) a. *Otne lei mis vuosttáš logaldallan.
Today was us cheese.diminutive lecture.
'Today we had our little cheese lecture.'

b. Otne lei mis vuosttaš logaldallan.
Today was us first lecture.
'Today we had our first lecture.'

Typical syntactic errors in North Sámi are incorrect verb forms after auxiliaries, errors regarding subordinate clauses, and use of personal instead of possessive pronouns. Lexical errors can resemble syntactic errors, they involve the use of erroneous lexemes in a certain construction, e.g. the use of the wrong postposition, e.g. *badjel* 'over' instead of *bokte* 'via, by means of' as in example (2).

(2) Dat máksá telefuvnna bokte/*badjel.
He pay telephone via/*over.
'He pays via telephone.'

Morphosyntactic errors refer to the morphological structure of a word. They include compound errors (in North Sámi compounds are written in one word), missing hyphens in constructions where compounds are coordinated, use of wrong morphological case in certain syntactic constructions such as in an adpositional phrase, case-number agreement errors, use of wrong case in coordination, use of wrong tenses in subordinate clauses.

3. Corpus

For developing rules a corpus of 193 sentences (4,349 tokens) has been constructed manually from erroneous sentences found in the Giellatekno corpus of North Sámi (18,142,181 tokens, mostly newspaper text) [3], and online blogs (`http://indigenoustweets.com/blogs/se/`).
Constructing an error corpus takes a lot of time and is rather a matter of coincidence than systematic searching. Others use Wikipedia (Miłkowski, 2007) for automatically constructing an errorcorpus. The North Sámi Wikipedia is written by many L2 speakers, which makes it inadequate as a test corpus for a L1 grammar checker. Additionally it is fairly limited in size (110,000 words).
Case errors are fairly "rule-based", they depend on certain fairly straightforward error patterns.
Real-word errors can be induced automatically by e.g. changing the character *á* to *a* vice versa. Typically, verbs ending in *-it/-at* such as *speadjalastit* 'to mirror, reflect' are a rich source for errors. The participle of the verb *speadjalastit* 'to mirror, reflect' is *speadjalastán*. When *á* is replaced by *a* on the last syllable, it becomes *speadjalastan* which is a compound of *speadjal* 'mirror' and *astat* 'have time', which is a possible but unlikely word.
Changing consonant clusters (single to double consonants vice versa) is another way to induce real-word errors automatically. Good candidates are genitive, nominative, locative, genitive possessive forms of nouns such as *várri* 'mountain' (nominative) vs. *vári* (genitive, accusative).
The testcorpus for postpositional case errors consists of 2000 sentences taken from the Giellatekno corpus for North Sámi. It contains 1000 correct sentences and 1000 sentences where a case error has been inserted.
Given that case errors are fairly straightforward (choice of 6 possible morphological cases) the error has been inserted manually, i.e. the correct case has been changed to an incorrect one. No other parts of the sentence have been changed.

4. How to set up a grammar checker - general architecture

4.1. Existing tools

The North Sámi rule-based grammar checker is written in Constraint Grammar formalism (Karlsson, 2006) and makes use of and enhances existing resources.
The morphological analyzers are implemented with finite-state transducers and compiled with the Xerox compilers `twolc` and `lexc` (Beesley and Karttunen, 2003). The other option is HFST[4].
The morphosyntactic disambiguators analyze the text syntactically, and at the same time disambiguate morphological and syntactic readings by means of context rules ideally leaving only one correct analysis. They are implemented in the CG-framework (Karlsson, 2006) and are based on manually written morphosyntactic rules that select and discard syntactic analyses. They further add grammatical functions and add dependency relations to the analysis. The rules are compiled with vislcg3 [5].

4.2. New modules and adaptions

In addition to the use of existing resources, the following tools have been created and adapted: The noun lexicon has been enriched by means of semantic tags, grammar-checker-specific rules and modification of rules in the disambiguator, a separate grammar-checker grammar including errortag mapping, dependency, valency and semantic role annotation.

4.2.1. Lexicon

The lexicon is enhanced by a number of semantic categories inspired by Bick's 150-200 semantic pro-

[3] `http://giellatekno.uit.no/doc/lang/corp/corpus-sme.html`

[4] `http://www.ling.helsinki.fi/kieliteknologia/tutkimus/hfst/`

[5] `http://beta.visl.sdu.dk/constraint\_grammar.html`

totypes (http://gramtrans.com/deepdict/semantic-prototypes). The categorization is not complete as categories are created as they are needed. Also the Basque grammar checker makes use of semantic categories for the disambiguation of postpositions and considers them necessary.

Currently, there are the following categories: *Masculine, Feminine, Surname, Place, Organisation, Object, Animate, Plant, Human, Group, Time, Text, Route, Measure, Weather, Building, Education, Clothes.*

4.2.2. Disambiguator

The disambiguator needs adaptations in order to be used for grammar checking. As mentioned by Bick (2006) and Johannessen et al. (2002), correct readings sometimes get discarded because of the erroneous context. This is due to contrary philosophies of grammar checking and regular disambiguation. In regular disambiguation, the assumption is made that input text is correct, and based on that assumption words are analyzed syntactically. The grammar checker on the other hand gets both correct and erroneous text as an input. The disambiguator needs to be adapted to those conditions. It needs to output a correct analysis despite errors in the context. The focus does not lie on removing as much ambiguity as possible, but on not discarding correct readings. Therefore adaptions have been made. As Johannessen et al. (2002), we are using a relaxed grammar checker with specific rules. Bick (2006) on the other hand runs the rules twice, some of them before and others only after applying the grammar checker.

Rules that remove correct readings are removed or changed in favor of leaving more ambiguity. Typical modifications consist in negating a typical context in which a certain reading should not be discarded. Rules for specific errortypes such as case errors in adpositional phrases are added. Instead of basing their whole disambiguation on morphosyntactic constraints, lexical and semantic constraints are being used.

4.2.3. Grammar checker module

The grammar-checker module is a separate module and constructed in the same way as the disambiguation grammar. The pedagaogical programs (*Oahpa*) for North Sámi use a basic grammar checker (Antonsen et al., 2009). There are several reasons for starting from scratch: the target group is a different one (L1 users vs. L2 users), the type of text serving as an input. While the input to *Oahpa* consists in very simple sentences, the grammar checker has to deal with complex sentences. In constraint grammar, two types of rules can add tags to words. MAP rules only allow a single error tag to be added per word, where ADD rules permit more than one to be added. The Sámi grammar checker uses ADD rules, while the Basque one uses MAP.

The idea behind that is that a word can have multiple errors which can be added onto each other. However, in some cases, rules do block each other. The architecture within the grammar checker is the following: The first rule determines the grammaticality of a sentence based on the existence of a finite verb (unless there is a given context for leaving it out, e.g. headers, answers in a dialogue etc.). The missing-verb tag blocks the application of certain other rules.

The grammar checker contains both errortags and correct tags. Correct tags have the function of analyzing complex conditions for certain grammatically correct constructions, which form an exception to other error rules.

```
ADD:corr-not-compound (&corr-not-compound)
 TARGET CNOUN IF
 (0C (N Sg Nom) LINK 1 CNOUN LINK p V) ;
```

This rule adds a correct tag to a noun (CNOUN) saying there is no compound error in the case where a dependency relation to a verb can be determined (the parent (p) of the noun is a verb (V)).

Errortag-matching rules refer to the correct tag constraining the mapping of an errortag.

The following rules are ordered in the this way:

1. real-word errors rules
2. dependency mapping rules
3. compound rules
4. lexical error rules
5. syntactical error rules
6. morphosyntactical error rules

Real-word errors, compound errors and lexical errors can constrain the other rules, which is why they are applied first. Dependencies help to identify syntactical and morphosyntactical errors, which is why they are mapped before the respective rules for that. CG3 (Bick, 2006) lets us add dependencies in the same grammar. They are used to construct partial trees for adpositional phrases and argument structures of verbs. Those help recognizing errors where the required dependent could not be matched. The dependency trees are very specific and only partial trees are mapped. Using a complete dependency annotations require full disambiguation with a high F-score, but as mentioned before, many times this is prevented by erroneous context and correct readings are discarded/a full disambiguation is not possible. Therefore a full dependency analysis will not give sufficiently good results. Oronoz (2008) uses dependencies to detect agreement errors, but the dependency analysis of erroneous text give such unreliable results that errortag-mapping is not sucessful.

5. Diagnosis and correction of case errors

A grammar checker needs to diagnose an error (identify its cause) and correct it (suggest an alternative the erronous item is substituted for).

Currently, the errortags include both the error and the correction, e.g.: *&msyn-gen-before-postp* (morphosyntactical error, there should be a genitive case before a postposition), *&msyn-gen-after-prep* (morphosyntactical error, there should be a genitive case after a preposition). The correction is still implicit, the correct form is not being generated yet.

Correcting case is one of the more challenging rules for North Sámi grammar checking as it can often involves long-distance relationships (argument structure) and need large contexts to identify the error.

Language learners typically make a lot of case errors, which is partly due to them infering syntactic constructions from their native language. Case errors that are made by native language users are mostly spelling errors in disguise, i.e. spelling errors resulting in a wrong case (similarly to a real-word errors). In the following example (3), the use of the single consonant **d** instead of **dd** makes a possessive form in nominative case out of the locative form that is required by the verb *jearrat* 'ask'.

(3) Mun jearan
I ask girl.sg.loc/*girl.sg.nom.px.sg3
nieiddas/*nieidas.

'I ask the girl.'

(4) Liikon
Like.1.sg.prs reindeer.meat.ill/reindeer.meat.acc
bohccobirgui/*bohccobierggu.

'I like reindeer meat.'

Another source of case errors are improper use of correct valency possibly due to influence from the majority case structures, e.g. *liikot* 'like' asks for illative case in Sámi, but for accusative case in Norwegian.

Case errors can be influenced locally (adpositions ask for genitive case) or globally (verbs asking for a certain case of their arguments).

Locally influenced case errors are errors in adpositional constructions as in example (5) and in partitive constructions as in example (6).

(5) Sii bidje bálgá mielde
They went path along hill.gen/*hill.nom
dievá/*dievvá badjel.
over.
'They went along the path over the hill'

(6) Muhtun osiin
In some parts
álbmotmeahcis/*álbmotmeahccis leat
nationalpark.Loc/nationalpark.Nom.PxSg3 have
ealggat bilistan nuorramuoraid.
elks destroyed young.trees.
'In some parts of the nationalpark, elks have destroyed the young trees.'

5.0.4. Adpositional phrases

Error detection of case in adpositional phrases is complicated by the extensive homonymy of adpositions themselves. Currently there are 305 adpositions and 1089 possible analyses of those (3.6 possible analyses per postposition).

An incorrect case in front of a postposition is detected in the following way. At first particular rules disambiguate the postposition itself. There are currently 60 rules, mostly rules selecting alternative readings to adpositional readings (e.g. adverbial readings), which are modified as not to discard correct adpositional analyses. *(NEGATE 0 Po LINK -1 Gen)* says that the rule should not apply to potential postpositions with a genitive to its left. Since those rules are assuming correct text as their input and often rely on correct text, the modifications prevent that correct readings in erroneous text are discarded.

51 more rules are disambiguation rules specifically choosing or discarding adpositional readings. Since the morphosyntactic context is not reliable - the main cue for selecting a postposition/preposition is a preceding/following genitive - semantic cues and valency information is used to rule out adpositional or alternative analyses. Especially the postpositions that are homonymous to adverbs and stand alone pose a problem to error tag matching. Nouns are usually rare forms or can be disambiguated by valency information. Verb readings can sometimes be difficult to disambiguate too unless they are rare forms (such as infinite forms that are used in specific contexts only, e.g. the verbgenitive form). Some rare noun forms are possessive suffixed nouns that require a human subject in the same person.

The following rule selects an adverbial reading if the lexeme is *sisa* 'into' and followed by a noun of the category *building* in illative case unless the word to the left of it is in genitive case.

```
SELECT:GramPo Adv IF (0 ("sisa") LINK 1
 BUILDING LINK 0 Ill)(NOT -1 Gen) ;
```

Other rules select a certain reading if the word is part of a multiword expression as *ieš alddis* 'by himself' where the form *alddis* is not a postposition with a possessive suffix ending, but a pronoun, *ieš* 'oneself', in locative case with a possessive suffix ending. This could possibly resolved in a different way.

```
REMOVE:GramPo ("alde") IF (0 PxSg3)
  (-1 ("ieš"));
```

The following rules select an adverbial reading if a certain set of verbs are there and a noun of the category CLOTHES is there.

```
SELECT:GramPo (Adv) IF (0 ("ala") LINK *0
 ("bidjat")  OR ("coggat") OR ("coggalit"))
 (-1 CLOTHES);
```

7 rules of the type SETPARENT/SETCHILD set the dependency relation of genitive nouns/pronouns/numerals to a following postposition that either belongs to the set of postpositions that can only be a postposition or to a disambiguated postposition which is following the genitive. Other rules are dealing with with prepositions.

```
SETPARENT Gen TO (*1C Po BARRIER S-BOUNDARY);
```

This rule depends on good disambiguation. If a secure postposition follows, the genitive is linked via depenedency to the postposition.

The following illustrates the different philosophies of the disambiguation grammars. The regular disambiguation grammar selects an adverb if the noun preceding the adposition is in nominative case, and a postposition if the preceding noun is in genitive case. But in grammar checking, the case of the preceding noun cannot be used as the only criterium for disambiguation of PoS because it can be erroneous. In order to resolve the adverb-adposition ambiguity, other constraints need to be used. In example (7-a), *gorži* ‘waterfall’ is a potentially erronous form. The difficulty here lies in that both *gorži vuolde* and *goržži vuolde* are possible bigrams because the nominative form can potentially be a subject/predicative of the sentence making *vuolde* `under, beneath’ an adverb. The decisive element that allows us to disambiguate between the adverb and adpositional reading of *vuolde* `under, beneath’, is the habitive *mus* ‘I.locativ’ together with the verb *leat* ‘to be’. Without another potential predicative in the sentence, the adpositional reading is discarded in (7-a), and *vuolde* `under, beneath’ is analyzed as an adverb.

(7) a. Mus lea gorži vuolde.
I.loc be.Sg3 waterfall.nom beneath.
‘I have a waterfall beneath (= under me).’

b. *Mus lea goržži vuolde.
I.loc be.Sg3 waterfall.gen under.
‘I have under a waterfall.’

The rule selects the adverbial reading of *vuolde* `under, beneath’ if there is a habitive and the verb *leat* ‘be’ to the left of it.

```
SELECT:GramPo (Adv) IF (0 ("vuolde")
LINK -1 N) (*-1 ("leat"))(*-1 @HAB) ;
```

The final step is the grammar checker error mapping rule itself, which maps an errortag **&msyn-gen-before-postp** to the potential dependent of an adposition (noun, pronoun, numeral, adjective) unless it is in genitive case and a dependent of the adposition.

```
ADD:gen-before-postp (&msyn-gen-before-postp)
TARGET NP-HEAD - ABBR IF (NOT 0 Gen)(1C Po)
(NEGATE 1 N);
```

In some cases, the wrong postposition is selected, e.g. *rastá* ‘through’ instead of *mielde* ‘along’, *badjel* ‘over’ instead of *rastá* ‘through’ etc. While in the Basque grammar checker, all errors in an adpositional phrase are treated as one type (due to other categorizations of case vs. adpositions), in North Sámi those are considered to be lexical errors, while the previously discussed error type is considered to be a morphosyntactic error.

6. Evaluation

For the evaluation, the rules for 5 adpositions represented in table (2) are evaluated.

The postpositions can be used in a local sense, many of them have other uses too. *ala* ‘on’ for example can be used with a number of psychological verbs such as *suhttat* ‘get angry at’, *dorvvastit* ‘rely on’, *luohttit* ‘trust’. *Bokte* ‘via, by means of’ is used as via as in *media bokte* ‘by means of the media’ etc. *Rastá* ‘across’ as in *rastá cearddaid* ‘across ethnicities’. Difficulties are especially there to disambiguate between the verb and the postposition if the case is wrong. but e.g. *boktit* ‘wake’ usually asks for an animate object.

adposition	translation	homonymy
ala/nala	onto	postp, adv, verb (aldat ‘get closer’)
alde/nalde	on	postp, adv
badjel	over	postp, prep
bokte	via	postp, verb (boktit ‘wake’)
rastá	across	postp, prep, adv

Table 2: Adposition homonymy

The testcorpus for evaluation contains 15,968 tokens and consists of 200 sentences for each postposition, 100 with correct case and 100 with incorrect case to evaluate both precision/recall and false alarms. An important task that remains is testing precision on a large corpus (e.g. newspaper corpus).

	err corp	corr corp	complete corp
tokens	16206	16105	18,142,181
errors	1000	0	-
detected errors	825	0	65
false alarms	23	-	14
precision	0.98	-	0.78
recall	0.83	-	-
f-score	0.93	-	-

Table 3: Quantitative evaluation

In the small corpus, there are few false alarms, precision is at 0.98. The recall is at 0.83.

The false alarms are mostly due to errors in disambiguation: adverbial vs. adpositional reading (11), genitive vs. accusative reading (5), verbal vs. nominal reading (1).

The reason for false alarms are exclusively disambiguation problems, typically adverb vs. adposition. But also disambiguation problems of the previous word (the dependent) can cause false alarms, either in terms of wrong part of speech (verb vs. noun) or the wrong case in terms of genitive/accusative homonymy. The disambiguation grammar is responsible for these false alarms. For example *čuovga alde* ‘light on’, *gákti alde* ‘Sámi clothes on’ can either be used in an expression where on means something like "turned on" (light) vs. "wearing" (clothes). Here, the grammar checker grammar is causing the false alarms.

The second measure is recall. The reasons for undetected errors are shown in table 4.

Sometimes more than one errortype apply and the reason can be a combination of several reasons. In many cases the adverb instead of the adposition is erroneously disambiguated with a preceding nominative, illative or locative. Another challenge is removing the verb first person dual forms of *aldat* ‘come closer’ and *boktit* ‘wake’, which still get selected too often. In some cases pre- vs. postposition

type	number
adverb not adposition	51
grammar-checker rule does not hit	49
pre- vs. postposition	40
verb not adposition	15
erronoeus disambiguation of previous word	6

Table 4: Undetected errors

disambiguation needs to be improved. This is most difficult where both a preceding and a following noun are there, and where the noun not belonging to the adpositional phrase is a genitive.

When the noun phrase is complex, e.g. *badjel min ipmárdusrájit* ‘across our understanding-borders,*Vuođđoláhka § 110 a bokte* `by means of constitution law § 110 a’ the grammar checker rules do not always hit. This can be improved by using more detailed constraints.

For testing precision a larger ‘natural’ corpus is needed, where not necessarily many errors can be found. I did a small test, with two of the postpositions ‘ala/nala’ (‘onto’) and ‘rastá’. 65 errors were detected, 51 of those correctly identified errors, and 14 false alarms, giving a precision of 0.78 %.

7. Conclusion

The construction of a grammar checker for North Sámi includes a number of challenges. The basis for the grammatical sentence analysis is no longer a correct sentence, but a potentially erroneous one. Not knowing where the errors might be together with high degree of homonymy of North Sámi word forms make morphosyntactic information partly unreliable for error detection. Since one cannot trust the grammar, one needs to make use of semantics and lexicon that can unambiguously identify the word, e.g. a postposition. A good disambiguation of the adpositions is the key for detecting case errors in adpositional phrases. Detailed adposition-specific rules that refer to the semantic context require some work, but are fairly successful in identifying the correct reading. The qualitative evaluation has revealed holes in the disambiguation of adpositions, but also potential for improvement.

As a next step, I would like to use the results from the qualitative analysis to improve both disambiguation and grammar checking rules, and make a new thorough analysis of precision on the complete Giellatekno corpus.

With regard to the grammar checker as a whole, a number of tasks remain to be done. After resolving case errors that depend on local context, resolving those depending on a global context (argument structure of the verb) by means of valency information can be attempted. I predict the task to be much more difficult than detecting errors in the adpositional phrase context and it remains interesting to see if effective solutions can be found. The resolution of real word errors and agreement errors are another important field of development.

As previous Constraint-based grammar checkers show, e.g. (Johannessen et al., 2002), Constraint Grammar-based grammar checkers can be integrated in Microsoft Office. The endproduct of the grammar checker for North Sámi should be integrated into Microsoft Office, Open Office, MacOSX, and InDesign as the main Sámi newspapers and publishing houses use InDesign.

8. Acknowledgments

Thomas Omma has constructed the error corpus, and written parts of the grammar. I would like to thank him for his linguistic help and advice on grammatical constructions, and Francis Tyers for his eagle eye for grammatical and formatting errors and his positivity.

9. References

L. Antonsen, S. Huhmarniemi, and T. Trosterud. 2009. Constraint grammar in dialogue systems. In *NEALT Proceedings Series 2009*, volume Volume 8, pages 13–21.

A. Arppe. 2000. Developing a grammar checker for swedish. In *Proceedings from the 12th Nordiske datalingvistikkdager*, Trondheim.

E. S. Atwell. 1987. How to detect grammatical errors in a text without parsing it. In *Proc. 3rd EACL*, pages 38––45, Copenhagen.

K. R. Beesley and L. Karttunen. 2003. *Finite State Morphology*. CSLI publications in Computational Linguistics, USA.

E. Bick. 2006. A constraint grammar based spellchecker for danish with a special focus on dyslexics. *A Man of Measure: Festschrift in Honour of Fred Karlsson on his 60th Birthday. Special Supplement to SKY Jounal of Linguistics*, 19:387–396.

E. Izumi, K. Uchimoto, T. Saiga, T. Supnithi, and H. Isahara. 2003. Automatic error detection in the japanese learners english spoken data. In *Companion Volume to Proc. ACL'03*, pages 145—148, Sapporo, Japan.

J. Bondi Johannessen, K. Hagen, and P. Lane. 2002. The performance of a grammar checker with deviant language input. In *Proceedings of the 19th International Conference on Computational Linguistics*, pages 1223–1227, Taipei, Taiwan.

F. Karlsson. 2006. *Constraint Grammar - A Language-Independent System for Parsing Unrestricted Text*. Mouton de Gruyter, Berlin.

S. Petrović Lundberg. 2009. Collecting and processing error samples for a constraint grammar-based language helper for esperanto. Master's thesis, Stockholms Universitet, Department of Linguistics.

M. Miłkowski. 2007. Automated building of error corpora of polish. *Corpus Linguistics, Computer Tools, and Applications – State of the Art. PALC 2007, Peter Lang. Internationaler Verlag der Wissenschaften 2008*, pages 631–639.

M. Miłkowski. 2010. Developing an open-source, rule-based proofreading tool. *Software – Practice and Experience 2010*, 40(7):543–566.

D. Naber. 2003. A rule-based style and grammar checker diploma thesis. Master's thesis, University of Bielefeld.

M. Oronoz. 2008. *Euskarazko errore sintaktikoak detektatzeko eta zuzentzeko baliabideen garapena: datak, postposizio-lokuzioak eta komunztadura*. Ph.D. thesis, Lengoaia eta Sistema Informatikoak Saila. Donostia.

Toward a Rule-Based System for English-Amharic Translation

Michael Gasser

Indiana University
Bloomington, Indiana, USA
gasser@indiana.edu

Abstract

We describe key aspects of an ongoing project to implement a rule-based English-to-Amharic and Amharic-to-English machine translation system within our L^3 framework. L^3 is based on Extensible Dependency Grammar (Debusmann, 2007), a multi-layered dependency grammar formalism that relies on constraint satisfaction for parsing and generation. In L^3, we extend XDG to multiple languages and translation. This requires a mechanism to handle cross-lingual relationships and mismatches in the number of words between source and target languages. In this paper, we focus on these features as well as the advantages that L^3 offers for handling structural divergences between English and Amharic and its capacity to accommodate shallow and deep translation within a single system.

1. Rule-Based Machine Translation

Among other disadvantages, African languages (and the communities of people who speak them) suffer from a lack of available documents in the languages, one aspect of the Linguistic Digital Divide (Paolillo, 2005). Machine translation (MT) and computer-assisted translation (CAT) could play a role in alleviating this problem. The ultimate goal of our project is the application of MT and CAT to the production of publication-quality documents in the major languages of the Horn of Africa.

Despite the impressive recent achievements of statistical machine translation (SMT), rule-based machine translation (RBMT) continues to offer advantages in certain contexts (Barreiro et al., 2011; Bond et al., 2011; Forcada et al., 2011; Mayor et al., 2011; Mel'čuk and Wanner, 2006; Ranta et al., 2010). SMT requires large parallel corpora, which are not available for under-resourced languages. The data sparsity problem is especially serious for morphologically complex languages such as Amharic because such languages have very large numbers of distinct word forms. Finally, for SMT systems, which normally rely on relatively primitive models of constituent order, significant structural differences between the languages can present problems.[1]

The relative advantages of SMT and RBMT also depend on the purpose of the MT system. For extracting the gist from a document, SMT systems already perform adequately for some language pairs and some domains. When the goal is publication-quality documents, however, an RBMT or RBCAT system, designed for a narrow content domain, would perform better (Ranta et al., 2010).

For our purposes, RBMT is clearly the way to start, and we are developing a framework for RBMT and RBCAT systems for under-resourced languages, L^3. L^3 relies on a powerful and flexible grammatical theory that we hope will be able to handle arbitrary syntactic divergences between languages. At the moment, we are far from our long-term goal of a set of tools that would allow developers to rapidly design MT systems for limited domains in a new language pair, something analogous to what Apertium (Forcada et al., 2011) offers. In this paper we discuss some features of L^3 that have emerged out of the development of a small prototype English-Amharic translation system.

In the next section, we introduce Extensible Dependency Grammar (XDG), the grammatical formalism behind L^3. Next we discuss the enhancements to XDG that are required for MT, including two aspects of the system not described in our previous work (Gasser, 2011b), the implementation of both shallow and deep MT and the handling of mismatches in the number of words. Next we show how L^3 deals with structural divergences between languages. We conclude with a discussion of ongoing work.

2. Extensible Dependency Grammar

Dependency grammars are a popular alternative to constituency grammars because of their simplicity, the ease of the integration of syntax and semantics, and their handling of word-order variation and long-distance dependencies. These advantages apply to RBMT as well (see, for example, Bick, 2007; Mel'čuk and Wanner, 2006). XDG (Debusmann et al., 2004; Debusmann, 2007) is a flexible, modular dependency grammar formalism that relies on constraint satisfaction for processing. Although XDG has not been tested with wide coverage grammars and unconstrained input, we believe that its flexibility and its proven capacity to handle complex syntactic constraints outweigh this drawback, especially since our goal is relatively small grammars for restricted input. Debusmann (2007) has developed a partial XDG grammar of English that handles many complex syntactic phenomena, and in earlier work (Gasser, 2010), we have shown how the unusual features of Amharic relative clause syntax can also be dealt with in this framework.

Like other dependency grammar frameworks, XDG is lexical; the basic units are words and the directed, labeled arcs connecting them. In the simplest case, an analysis of a sentence is a directed graph over a set of **nodes**, one for each word in the analyzed sentence, including a distinguished **root node** representing the end-of-sentence punctuation. As in some other dependency frameworks, for example, Meaning-Text Theory (see Mel'čuk and Wanner, 2006 for the application of MTT to MT), XDG permits analyses on

[1] Hybrid RBMT-SMT systems are beginning to address some of these deficiencies.

multiple **dimensions**, each corresponding to some level of grammatical abstraction.

In L^3, each language is represented on two dimensions, Immediate Dominance (ID) and Linear Precedence (LP), and a further dimension, Semantics (SEM), is responsible for language-independent conceptual structure. A key feature of XDG is the **interface dimensions** that relate dimensions such ID and LP to one another. Interface dimensions have no arcs of their own but instead constrain how arcs in the related dimensions correspond to one another.

2.1. Analyses as multigraphs

In the general case, then, an analysis of a sentence is a **multigraph**, consisting of a separate dependency graph for each dimension over a single sequence of word nodes. Figure 1 shows a possible analysis for the English sentence *the doctor cured the patient* on the ID and SEM dimensions, each represented by a plane in the figure. (The LP dimension is omitted here and in subsequent figures for the sake of simplicity.) Each node is represented by a pair of circles or squares joined by dashed lines. The square node is the end-of-sentence root node.[2] Arrows go from heads to dependents (daughters). For simplicity, we refer to the core arguments of semantic predicates as arg1 and arg2, rather than agent, patient, etc. On the SEM dimension, we maintain the convention that only content words participate in the representation. That is, any strictly grammatical words appearing in the ID dimension are effectively "deleted" in the SEM dimension. As shown in the figure, "virtual deletion" is handled in XDG through the use of special del arcs (Debusmann, 2007). In subsequent figures, we omit the del arcs, indicating deleted nodes with unfilled circles.

A grammatical analysis is one that conforms to a set of **constraints**, each applying to a particular dimension. Constraints belong to several categories, the most important of which are **graph** constraints, restricting the structure of the dependency graph; **valency** constraints, governing the labels on the arcs into and out of nodes; **agreement** constraints; **order** constraints; and various **linking** constraints that apply to interface dimensions and govern the manner in which arcs on one dimension are associated with arcs on another dimension.

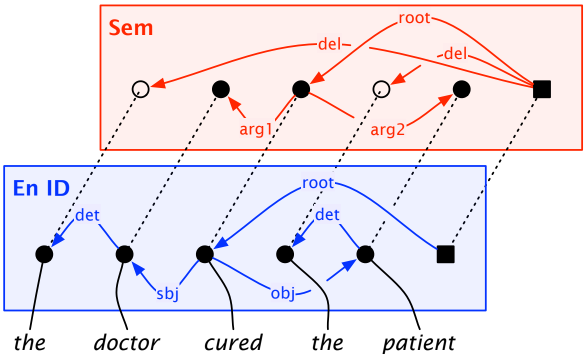

Figure 1: XDG analysis of an English sentence.

2.2. The lexicon

An XDG grammar of a language consists of a set of dimensions, each with its own set of constraints and arc labels, and a lexicon. As XDG is completely lexical, all specific grammatical constraints are stored in word-level units.

The **lexicon** consists of a set of **entries** associated with word forms or lemmas. Each entry contains one or more grammatical constraint attributes. Entry 1 shows a portion of the entry for the English lemma *cure*. The entry includes three valency constraint attributes on the ID dimension. The word requires outgoing subject (sbj) and object (obj) arcs and an incoming root arc.[3] (The "!" represents the requirement of exactly one arc with the given label.)

Entry 1 Portion of English entry for the lemma *cure*

```
- lemma: cure
  ID:
    out: {sbj: !, obj: !}
    in: {root: !}
```

2.3. Parsing and generation

Parsing within XDG begins with a **lexicalization** phase which creates a node for each word in the sentence and searches the lexicon for matching entries. For morphologically complex languages, such as Amharic, it is impractical to store all word forms in the lexicon, and we employ in-house morphological analyzers to pre-process the input words for such languages. Morphological analysis of the input words results in one or more lemmas and sets of grammatical features for each analyzed word. The word forms or lemmas in the input are matched against lexical entries, and a copy of each matching entry (a **node entry**) is added to the nodes. Each node is identified by an index representing its position in the input sentence.

The next phase is **variable and constraint instantiation**. Each of the constraints referenced in the constraint attributes in the node entries is instantiated. The constraints apply to a set of variables, which are created during this phase. For example, each node n has a daughters variable whose value is the set of indices of the daughter nodes of n. Among the constraints that apply to such a variable are graph constraints.

Finally, **constraint satisfaction** is applied to the variables and constraints that have been instantiated. If this succeeds, it returns all possible complete variable assignments, each corresponding to a single analysis of the input sentence, that is, a multigraph across the sentence nodes.

Because an XDG grammar is declarative, it can be used for generation as well as for analysis. The main difference is that for generation the semantic input does not specify the positions for words in the output. This problem is handled in a straightforward fashion through the creation of a position variable for each node; these variables are constrained by explicit order constraints. Another difference relates to the frequent mismatch in the number of nodes between semantics and the ID and LP dimensions (Pelizzoni and Nunes, 2005), a problem we discuss below.

[2] XDG does not handle sentence fragments; we are currently extending the system to have this capacity.

[3] In our simple grammar, there are no dependent clauses, so all finite verbs are the heads of sentences.

3. L^3

L^3 is an extension of XDG to translation. Enhancements to the basic framework include separate lexica for each language and **cross-lingual links** joining lexical entries in different lexica. Within L^3 we treat semantics, consisting of a single SEM dimension, as a language with its own lexicon.

3.1. Multilingual multigraphs

Adaptation of XDG to translation is straightforward once we make the leap to thinking of a sentence and its translation into another language as a single multilingual "sentence" with a single set of multilingual word nodes. This is illustrated for the sentence of Figure 1 in Figure 2, where we have shown the ID dimensions for English and Amharic and the SEM dimension.[4] Note that the order of the words in the Amharic output is subject-object-verb rather than subject-verb-object, as in the English input.

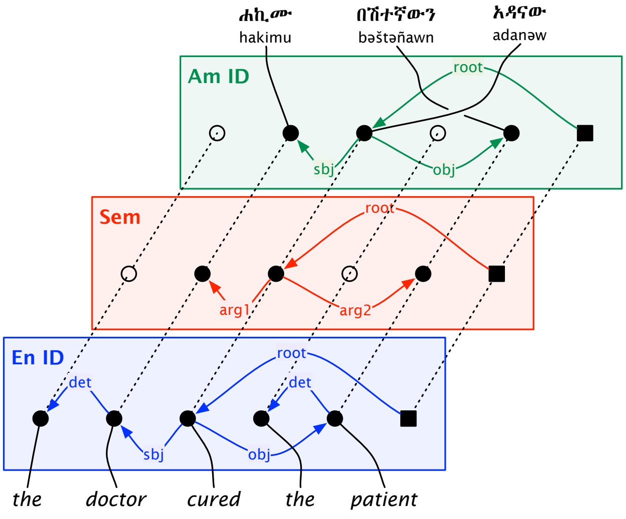

Figure 2: Multigraph for a bilingual "sentence".

3.2. Cross-lingual links

In keeping with the lexical nature of XDG, all cross-lingual knowledge in the system takes the form of links joining lexical entries in different languages. By convention in L^3, these always join the ID dimension of a language to another dimension, either the SEM dimension of semantics, or the ID dimension of another language. These links specify a target lexical entry and optionally a constraint attribute on the interface dimension joining the two arc dimensions. Entry 2 shows the portion of the English entry for the verb *cure* that links it to the entry in the semantics lexicon for the corresponding CURE event type and the corresponding entry in the Amharic lexicon, አዳነ *adanə*. There are attributes for the linking constraint, linking end, on both interface dimensions, IDSEM and IDID.

The linking end constraint is illustrated in general and for the special case of IDSEM for the subject of *cure* in Figure 3. For a node n, linking end specifies a relationship between arc labels $l1$ and $l2$ on dimensions $d1$ and $d2$ respectively. It constrains n to have as the daughter on its $l1$ arc in $D1$ a node which is the daughter of some node on $D2$, not necessarily n, on an arc with label $l2$. As we will see below, linking end, and other similar linking constraints, are the key to representing structural differences on different dimensions (ID and Sem or ID in one language and ID in another) across the same set of word nodes.

Entry 2 Cross-lingual links in the entry for *cure*

```
- lemma: cure
  cross:
    sem:
      lex: CURE
      IDSem:
        linkend: {arg1: [sbj]}
    am:
      lex: adane
      IDID:
        linkend: {sbj: [sbj]}
```

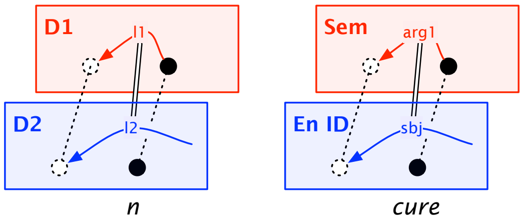

Figure 3: Linking end, in general and for subject of *cure*.

3.3. Shallow and deep translation

The fact that cross-lingual links are possible between English and Amharic directly as well as between English or Amharic and semantics means that the depth of translation is flexible in L^3.

Shallow (or transfer) vs. deep (or interlingua) translation is a distinction going back to the early days of RBMT. Deep translation (e.g., Bond et al., 2011; Mel'čuk and Wanner, 2006) makes use of a language-independent semantic level, rules mapping source-language lexical items and structures to semantic units, and rules mapping semantic units to target-language lexical items and structures. The main advantage is that it is possible to translate from any of the system's source languages into any of its target languages without special-purpose rules for language pairs. The drawback of the deep approach is the difficulty faced in devising a semantics that is abstract enough to cover a large set of languages. The addition of new languages may necessitate changes in the semantics, which may in turn require changes in all of the source-to-semantics and semantics-to-target interfaces. This disadvantage is mitigated to some extent when translation is limited to narrow domains (Ranta et al., 2010).

Shallow approaches do not suffer from these problems; since the rules apply only to a specific language pair, there is no need to search for abstract general representations and no need to update the whole system when new language

[4] Since the Amharic nouns ሐኪም *hakim*'doctor' and በሽተኛ *bəštəña* 'patient' can be either masculine or feminine, the Amharic sentence shown is only one of four possible translations for the English sentence.

pairs are added. In addition, shallow MT is normally more efficient than deep MT because less processing is required at both ends.

Given the advantages of both approaches, it would make sense to integrate them in some way. Human translators seem to use such an eclectic approach, relying on a deep understanding when they need to, but “faking it” and making use of shortcuts when they can or when they lack the knowledge required for a deep understanding (Byrne, 2006). As far as we know, however, all existing MT systems operate either one way or the other.

L^3 integrates both shallow and deep approaches into a single system. It is straightforward to simply tell L^3 to make use of source-to-semantics and semantics-to-target links or only source-to-target cross-lingual links during translation. When they are available, the latter will always be faster, involving fewer variables and fewer constraints to satisfy. For example, consider the translation pair illustrated in Figure 2. For translation of the English sentence into Amharic, 3544 constraints are required for the deep approach, while only 2879 are required for the shallow approach. For translation of the Amharic sentence into English, 5594 constraints are required for the deep approach, 4371 for the shallow. There is a savings in the time required for constraint satisfaction of 28.1% in the former case, 30.3% in the latter.

3.4. Node mismatch

For cases where words in the input sentence do not correspond to explicit semantic units, XDG makes use of del arcs to implicitly delete the nodes on the SEM dimension. When there is a node mismatch in the opposite direction, however, this solution will not work (Pelizzoni and Nunes, 2005). This happens, for example, in generation, when the semantic input has no nodes for words that must appear in the output, such as determiners and auxiliary verbs. It also happens frequently in translation. For example, like many other languages, Amharic is a zero-subject (“pro-drop”) language which may have no explicit subject. Like most other Semitic languages and many Bantu languages, it is also a zero-object language which may have no explicit direct or indirect object when this is coded as an affix on the verb. Thus in the English translation of an Amharic sentence such as አዳናት *adanat* ‘he cured her’, consisting of a verb only, the nodes for the English pronouns *he* and *her* must come from somewhere.

For this purpose, L^3 includes the possibility of **empty nodes**. Empty nodes are created during processing on the basis of **trigger nodes** associated with explicit input words. There are several types of empty nodes for different syntactic contexts. Here we focus on empty nodes for verb arguments that may not be explicit in the source languages. In the lexicon for Amharic and other such languages, finite verb entries include an attribute for a subject empty node, and finite transitive verbs include an attribute for an object empty node. Each of these empty node categories in turn has its own lexical entry (indicated by @SBJ and @OBJ in what follows), specifying a set of constraint attributes just as for any other entry. Thus verb nodes are the triggers for these argument empty nodes. When an input Amharic sentence contains a finite verb, a subject empty node is automatically added to the node list and assigned to the empty subject lexical entry. If the verb is transitive, an object empty node is also created and assigned to the empty object lexical entry.

The key property of the subject and object empty nodes is that there is no way of knowing prior to constraint satisfaction whether they will be needed or not. If the input Amharic sentence has no explicit subject (or object), the empty nodes get realized as explicit nodes in the English output:

አዳናት ⇒ እዳናት @SBJ @OBJ
⇒ {*cured*, *he*, *her*}
⇒ *he cured her*

If, on the other hand, the input Amharic sentence has an explicit subject (or object), the nodes remain empty in the English output. In the following sentence, for example, the noun አስቴርን *astern* ‘Esther (acc.)’ plays the role of object:

አስቴርን አዳናት ⇒ አስቴርን እዳናት @SBJ @OBJ
⇒ {*Esther*, *cured*, *he*, *zero*}
⇒ *he cured Esther*

To handle these cases we have introduced a set of empty node constraints in XDG. Informally, the constraints specify that the verb can have only one daughter along a sbj (obj) arc. If there is an explicit subject (object), the node for this constituent’s head plays the role of daughter along the sbj (obj) arc from the verb, and the corresponding empty node remains empty in the target language. If there is no explicit subject, the empty node plays the role of sbj (obj) daughter in Amharic. Because of Amharic agreement constraints, the empty node is constrained to have particular person, number, gender, and case features. This node is realized in English as an explicit pronoun. Which pronoun it becomes is determined by English agreement constraints: each candidate pronoun has features that must agree with the features of the empty node, and the selection of the appropriate pronoun is accomplished through constraint satisfaction.

3.5. Translation

Given the modifications of the basic XDG framework discussed above, translation proceeds more or less in the same fashion as parsing and generation in XDG. As an example, consider the translation of the sentence *the doctor cured the patient* into Amharic. If the input language is morphologically complex, it is first processed with a morphological analyzer, resulting in one or more combinations of lemmas and grammatical features for each node. This is not the case for our English input sentence, however. Next lexicalization searches for entries in the source language lexicon that match the input words or lemmas. In addition to language-specific attributes, such as valency, agreement, and order, the matched lexical entries may also provide cross-lingual links, either to semantics or to the target language. These links are traversed during lexicalization, and the attributes in the “language” on the other end of the link are copied in the relevant node entry. For example, via cross-lingual links, node 3, corresponding to *cured* in the input, gets features from the lexical entry for the Amharic translation of *cure*, አዳነ *adanə*, including the general valency, order, and agreement constraints of Amharic transitive verbs.

Finally, constraint satisfaction applies, as before. For a morphologically complex target language, as in this example, the result is a set of ordered target-language lemmas along with grammatical features. The final step is morphological generation, which yields target-language surface forms. We use our in-house Amharic morphological generator for this purpose (Gasser, 2011a).

As noted already for the monolingual case with respect to parsing vs. generation, the fact that all of the information in the lexica is declarative means that the same knowledge can be used for translation in both directions. An exception is the case of empty nodes, which are deleted in one direction and inserted in the other. Thus, given separate general mechanisms for handling empty nodes in the two directions, the fact that the system has the knowledge to allow it to translate the English sentence *the doctor cured the patient* into the Amharic sentence ሐኪሙ በሽተኛውን አዳነው *hakimu bəštəñawn adanəw* (among other possibilities) means that it can translate the Amharic sentence to the English sentence.

3.6. Structure mapping

An important concern in the design of MT systems is that they have the capacity to represent the structural divergences between languages. Dorr (1994) provides one classification of the sorts of divergences that can occur.

Here we discuss an English-Amharic example illustrating two of Dorr's divergence types: thematic and categorial. Amharic has a set of impersonal experiential verbs whose subject agreement is always third person singular masculine and whose object suffixes agree with the experiencer argument. If the experiencer also takes the form of an explicit argument of the verb, this argument is unmarked for case, like a subject, but it agrees with the verb's object suffix rather than with the verb's subject affixes. We refer to this argument as a "topic". An example is the verb ደከመ *dəkəmə* 'be tired': ደከማት *dəkəmat* 'she is tired', lit. 'it tired her'; አስቴር ደከማት *aster dəkəmat* 'Esther is tired', lit. 'Esther, it tired her'. For these verbs Amharic and English differ in two ways. English uses an adjective, along with a form of *be*, while Amharic uses a dedicated verb. In English the subject of *be* is the experiencer, while in Amharic the experiencer is cross-indexed as obligatory object agreement on the verb and (if present) as the syntactic topic of the sentence.

In L^3 structural divergences are handled on interface dimensions. Consider first the lexical entry for the Amharic verb ደከመ *dəkəmə* 'be tired'. With respect to the Amharic ID and LP dimensions, this is a transitive verb with the unusual property that it may have a topic argument (the experiencer) that agrees with the verb's object morphology. Because the topic may not be realized and because it corresponds to an explicit argument in the semantics and in languages such as English, the verb acts as a trigger for a **topic empty node**.

In Entry 3, we show some of the information in the English and Amharic lexica that enables translation of these sentences. For Amharic we show a portion of the entry for the verb lemma ደከመ *dəkəmə*. This acts as a trigger for an empty topic node (with lexical entry @TOP), which requires an outgoing arc in the ID dimension with a top label. Also on the ID dimension, there is an agreement constraint attribute specifying that the topic must agree with the object suffix on the verb on the person, number, and gender (png) feature. The Amharic verb has a cross-lingual link to semantics that associates this lemma with the semantic lemma TIRED and stipulates that the arg1 arc from this node on the SEM dimension should go to the node that has an incoming top arc on the Amharic ID dimension.

For English two relevant entries are shown. The first is for the verb *be* with predicate adjectives (as in *be tired*). This entry stipulates that this node must have an outgoing arc in the ID dimension with a padj (predicate adjective) label. This entry also has a cross-lingual link to semantics that deletes the associated node on the SEM dimension. Finally, we show a portion of the lexical entry for the English adjective *tired*. It stipulates that this word must have an incoming arc in the ID dimension with a padj label. There is also a cross-lingual link to semantics that associated this word with the semantic lemma TIRED and stipulates that the arg1 arc from this node on the SEM dimension should go to the node that has an incoming sbj arc on the English ID dimension.

Figure 4 illustrates the translation of the sentence *she is tired* into Amharic. L^3 returns two translations in this case, one with and one without the explicit topic pronoun እሷ *ɨswa* 'she' (indicated in the figure by the gray circle for node 1 in Amharic ID). The same entries shown in Entry 3 would enable translation in the reverse direction.

Entry 3 TIRED in English and Amharic

```
AMHARIC
- lemma: dekeme
  empty: [@TOP]
  ID:
    out: {top: !}
    agree: [[top, obj, png]]
  cross:
    sem:
      lex: TIRED
      IDSem:
        linkend: {arg1: [top]}
ENGLISH
- lemma: be_padj
  ID:
    out: {padj: !}
  cross:
    sem:
      lex: zero
- lemma: tired
  ID:
    in: {padj: !}
  cross:
    sem:
      lex: TIRED
      IDSem:
        linkend: {arg1: [sbj]}
```

4. Project status and ongoing work

Because we are making some basic modifications to XDG, we have re-implemented the framework from the bottom up. Our implementation is in Python; lexica are encoded in YAML format. All of our software, includ-

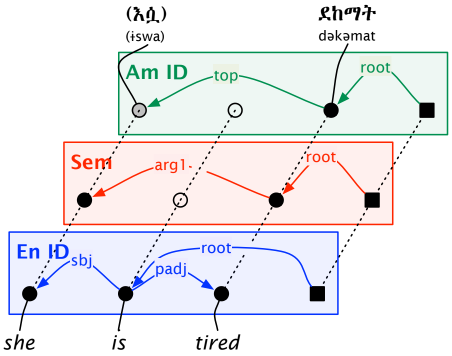

Figure 4: Translation of *she is tired* into Amharic.

ing the XDG implementation, lexica, and morphological analyzer and generator for Amharic, is available under a GNU GPL3 license at `http://www.cs.indiana.edu/~gasser/Research/software.html`.

It is premature to attempt an evaluation of our very rudimentary English-Amharic translation system.[5] Our next step is to augment both the English and Amharic grammars with more structures and to apply it to translation within the restricted domain of arithmetic and to computer-assisted translation within the domain of economics. In parallel with this work, we are developing an API for the grammar framework that will enable users to create simple grammars for additional languages. We are also exploring ways to integrate machine learning into the framework by using existing Amharic-English parallel corpora to aid in lexical disambiguation.

5. Conclusion

In this paper, we have introduced L^3, an evolving framework for RBMT and RBCAT that we are applying to the development of an English-Amharic MT system. We have discussed and illustrated some of the features of L^3: its bidirectionality, its capacity to handle structural divergences between typologically diverse languages such as English and Amharic, and its integration of shallow and deep translation into a single system. Although our Amharic-English MT system is still only a toy, by focusing on features that distinguish the languages, we feel we are on the right track.

References

Barreiro, A., Scott, B., Kasper, W., and Kiefer, B. (2011). Openlogos machine translation: Philosophy, model, resources, and customization. *Machine Translation*, 25(2):107–126.

Bick, E. (2007). Dan2eng: wide-coverage Danish-English machine translation. In *Proceedings of Machine Translation Summit XI*, pages 37–43, Copenhagen.

Bond, F., Oepen, S., Nichols, E., Flickinger, D., Velldal, E., and Haugereid, P. (2011). Deep open-source machine translation. *Machine Translation*, 25(2):87–105.

Byrne, J. (2006). *Technical Translation: Usability Strategies for Translating Technical Documentation*. Springer, Dordrecht, the Netherlands.

Debusmann, R. (2007). *Extensible Dependency Grammar: A Modular Grammar Formalism Based On Multigraph Description*. PhD thesis, Universität des Saarlandes.

Debusmann, R., Duchier, D., and Kruijff, G.-J. M. (2004). Extensible dependency grammar: A new methodology. In *Proceedings of the COLING 2004 Workshop on Recent Advances in Dependency Grammar*, Geneva/SUI.

Dorr, B. (1994). Machine translation divergences: a formal description and proposed solution. *Computational Linguistics*, 20(4):597–633.

Forcada, M. L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Sánchez-Martínez, F., Ramírez-Sánchez, G., and Tyers, F. M. (2011). Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144.

Gasser, M. (2010). A dependency grammar for Amharic. In *Proceedings of the Workshop on Language Resources and Human Language Technologies for Semitic Languages*, Valletta, Malta.

Gasser, M. (2011a). HornMorpho: a system for morphological processing of Amharic, Oromo, and Tigrinya. In *Proceedings of the Conference on Human Language Technology for Development*, Alexandria, Egypt.

Gasser, M. (2011b). Towards synchronous extensible dependency grammar. In *Proceedings of the Second International Workshop on Free/Open-Source Rule-Based Machine Translation*, Barcelona.

Mayor, A., Alegria, I., de Ilarraza, A. D., Labaka, G., Lersundi, M., and Sarasola, K. (2011). Matxin, an open-source rule-based machine translation system for Basque. *Machine Translation*, 25:53–82.

Mel'čuk, I. and Wanner, L. (2006). Syntactic mismatches in machine translation. *Machine Translation*, 20(2):81–138.

Paolillo, J. (2005). Language diversity on the internet: Examining linguistic bias. In UNESCO Institute for Statistics, editor, *Measuring Linguistic Diversity on the Internet*. UIS, Montreal, Quebec, Canada.

Pelizzoni, J. M. and Nunes, M. d. G. V. (2005). N:m mapping in XDG — the case for upgrading groups. In *Proceedings of the Workshop on Constraint Solving and Language Processing*, Roskilde, Denmark.

Ranta, A., Angelov, K., and Hallgren, T. (2010). Tools for multilingual grammar-based translation on the web. In *Proceedings of the Association for Computational Linguistics System Demonstrations*, Beijing.

[5] Evaluation of our Amharic morphological analyzer and generator has been reported on elsewhere (Gasser, 2011a).

Technological tools for dictionary and corpora building for minority languages: example of the French-based Creoles

Paola Carrión Gonzalez(1,2), Emmanuel Cartier(1)
(1)LDI, CNRS UMR 7187, Université Paris 13 PRES Paris Sorbonne Paris Cité
(2)Departamento de Traducción e Interpretación, Facultad de Filosofía Y Letras, Universidad de Alicante
E-mail : pccg1@alu.ua.es, ecartier@ldi.univ-paris13.fr

Abstract

In this paper, we present a project which aims at building and maintaining a lexicographical resource of contemporary French-based creoles, still considered as minority languages, especially those situated in American-Caribbean zones. These objectives are achieved through three main steps: **1)** Compilation of existing lexicographical resources (lexicons and digitized dictionaries, available on the Internet); **2)** Constitution of a corpus in Creole languages with literary, educational and journalistic documents, some of them retrieved automatically with web spiders; **3)** Dictionary maintenance: through automatic morphosyntactic analysis of the corpus and determination of the frequency of unknown words. Those unknown words will help us to improve the database by searching relevant lexical resources that we had not included before. This final task could be done iteratively in order to complete the database and show language variations within the same Creole-speaking community. Practical results of this work will consist in 1/ A lexicographical database, explicitating variations in French-based creoles, as well as helping normalizing the written form of this language; 2/ An annotated corpora that could be used for further linguistic research and NLP applications.

1. Introduction

Minority-languages have always existed and will continue to exist. That is an historical, sociological and political fact. But nowadays, electronic devices, internet communication and computer storage enable to keep track of their existence and, more, to study their evolution. In this perspective, our project aims at setting up a methodology and tools facilitating the study of these languages through NLP technologies. The project will take as example the American-Caribbean creoles, largely spread over the Caribbean area and considered as a cultural vehicle, but not yet an official language except for Haiti. This language has not yet attained normalization, as it lacks sufficient lexicographical resources and a real political strategy.

We will present in this paper the first steps to setup a dictionary of American-Caribbean Creoles, mirroring the effective use of this language in web corpus. It will use the up-to-date Natural language processing tools developed to study the major languages. We will first present the existing lexicographical resources for this language, then will detail the main steps of the project: compilation of existing and available dictionaries, use of web corpora to complete and tune the lexicographical resources.

2. Existing Lexicographical Resources in French-based Creole

In this section, we will present the main existing lexicographical resources in French-based Creoles. Our goal is twofold: first to study how lexicographers have dealt with the specific Creole situation, in terms of macro and microstructure, as well as in terms of number and nature of words included; second, to explicit which dictionaries can be reused for numerical purposes, in terms of availability, copyrights and ease of compilation.

This section is naturally divided into two parts: the first one will detail the main historical dictionaries, in paper format; the second one will explicit electronic resources. It will conclude by identifying the resources that we can use.

2.1 Paper-based dictionaries

The first lexicographical works date from the end of the 19th century and were mainly represented by bilingual lexicons.

Specific Creoles had focused attention such as the Haitian Creole and other varieties of Caribbean Creole languages (in Guadeloupe or Martinique). Albert Valdman, Annegret Bollée, Robert Chaudenson, Hector Poullet and Raphaël Confiant are the precursors of the lexicographical development on Creole languages. They produced several dictionaries that contributed to the lexicographical description. In 1885, Lafcadio Hearn compiled hundreds of proverbs of various types of Creole languages with various cultural information. A few decades later, research teams coming from several universities initiated ambitious projects whose achievements would become the major reference in Creole languages lexicography. First, lexicographers have focused on Haitian and Indian Ocean Creole languages, as more resources were available. Then, Caribbean Creoles, with the works of Hector Poullet, Sylviane Telchid and Raphaël Confiant have been developed.

These various attempts are characterized by the disparity of macro and microstructures. The first difference resides in orthographic conventions: some authors remain closer to French while others recommend a system allowing to break the speech continuum with its parent language, due to the process of decreolization. Other differences in microstructure are more common in lexicography, depending on the description objectives (insertion of spelling variants, available examples, translation of these examples in other languages, parts of speech, origin of entry words, etc.). We must also point out specific Creole languages descriptive elements, which have been specified by (Hazaël-Massieux, 2002): presence of false friends ("fig" = figue), erroneous and uncontrolled diversion (*chokatif - chokativité*), creation by erroneous integration of French words (*abònman* / abonnement).

Macrostructure’s main problems are also noticed by the same author: treatment of diatopic variation (all the spelling variants should be specified); demarcation, sometimes complicated, between French-based Creole and its parent language; and selection of the technical and scientific vocabulary. The lack of monolingual[1] works also constitutes one of the most considerable lack in Creole languages.

The main etymology dictionaries: DECA and DECOI

The main lexicographical resources in French-based Creoles derive from two projects aiming at studying etymology, one dealing with Indian Ocean Creoles (IO => DECOI) and the other American-Caribbean Creoles (AC => DECA). The second project is managed by Annegret Bollée (University of Bamberg in Germany) and Ingrid Neumann-Holzschuch (University of Regensburg), with the collaboration of Dominique Fattier (University of Cergy-Pontoise), inspired from (Chaudenson, 1979).

IO Creoles were first studied because more documentation was available. The project ended in the DECOI, the Etymological Dictionary of Indian Ocean French-based Creoles (1993), managed by Bollée, with the cooperation of Patrice Brasseur, Robert Chaudenson and Jean-Paul Chauveau. Composed of four volumes, it is divided into two parts: the first one devoted to French-originated words and the second to words with other and unknown origins. A large part of the etymological information comes from (Chaudenson, 1974). The last volumes of the dictionary appeared in 2007. The DECA then began. This last work, still in progress, is largely based on Haitian Creole material from Albert Valdman (Creole Institute, University of Indiana); the *Haiti Linguistic Atlas* (HLA), elaborated by (Fattier, 1998) and supervised by Robert Chaudenson, is also one of the sources of the DECA; Félix-Lambert Prudent who prepares a dictionary of Creole from Martinique contributes also with the creation of this dictionary. The purpose of this work is not only to establish etymological indications of Haitian Creole, but also to compare varieties of Creoles, (mainly from Guadeloupe and Reunion regions). One of the most important difficulties is the orthographic variation in Creoles (Allen, 1998). The DECA will also include an electronic version, in TEI format.

2.2 Electronic format dictionaries

The most exhaustive on-line dictionaries are Krengle[2] and Webster's online dictionary[3].

Krengle (approximately 18000 entries) is an Haitian Creole – English dictionary, developed by Eric Kafe, a computational linguist (University of Copenhagen). The English section is connected to Wordnet, so every entry offers definitions and semantic relations such us hyperonyms, hyponyms, synonyms and antonyms, sometimes accompanied with examples. The last update of this resource was made in July 2008. This dictionary can be partly downloaded for free: a list of English - Creole and Creole - English words, with some pairs of sentences in English - Creole.

Webster's online dictionary is an online multilingual Thesaurus offering more than 1200 languages. The Creole part is the result of a revision of the work supervised by the Haitian Creole specialist Noah Porter in 1913. The dictionary can be enriched by users via moderation. This tool has interesting characteristics as it recognizes several varieties of Creole, so that users can point out lexical resemblances and dissemblances in the same family of Creole languages. In addition, it includes synonyms and refers to other lexicographical sites.

Other lexicographical databases are available on the web, offering less information than aforementioned, but easier to retrieve, such as glossaries and lexicons of words of frequent use, which partially supplied information in the database. See (Carrión, 2011) for more details.

3. Project Architecture

This quick overview of existing lexicographical resources show that research and development are still in its infancy, and far from exhaustive, whereas our goal in this project is to build an electronic dictionary covering most of the general-purpose vocabulary, as well as identifying spelling variants and an environment to maintain it through continuous corpora analysis. As a result we have setup an architecture enabling to build up and maintain dictionaries through corpus analysis and an iterative process between dictionary and corpora, as described in figure 1.

This architecture comprises five main steps:

1. Step 1 (S1): this preliminary step aims at building a first electronic compilation from existing lexicographical resources. This implies gathering the existing resources, either digitized resources from paper-based dictionaries, or fully electronic resources. This also implies identifying available resources. This initial step is detailed in section 4;

2. Step 2 (S2) : this second step aims at setting up an infrastructure enabling corpora building and feeding; this means identifying web available resources as well as setting up automatic procedures to retrieve on a regular basis these documents; it is detailed in section5.1.

3. Steps 3 and 4: (S3-S4) morphosyntactic analysis of the corpora to maintain the existing dictionary; automatic analysis of corpora will explicit unknown words, and some of them will have to be included in the initial dictionary; iteration of this procedure will permit to complete the existing dictionary, as well as enabling morphosyntactic annotation of the corpora; this step is detailed in section 5.2.

4. Step 5 (S5) : this step, out of the scope of this paper, will be implemented as soon as the dictionary is sufficiently completed; annotated corpus could then be validated and then queried using linguistic and statistical tools, so as to improve information in the dictionary. It will be evoked in the section 5.3.

[1] There is only a monolingual dictionary in Mauritian Creole, elaborated by Arnaud Carpooran: *Diksioner Morisien*, Koleksion Text Kreol, Ile Maurice, 2009.

[2] http://www.krengle.net

[3] http://www.websters-online-dictionary.org/browse/

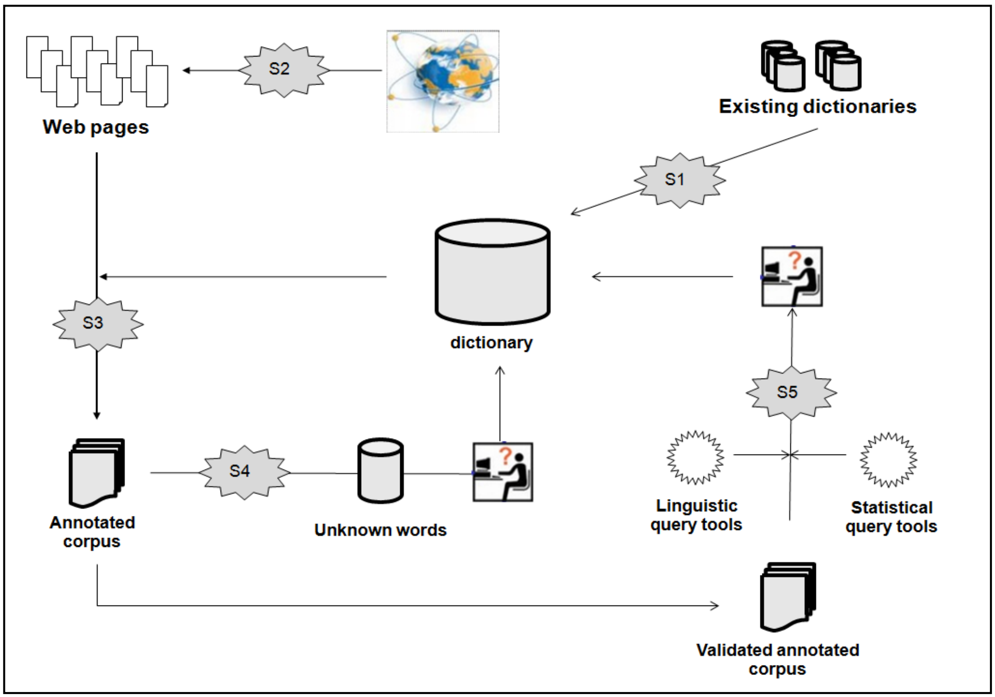

Figure 1 : project architecture

4. Dictionary building: existing resources compilation

Among existing resources, only a few ones are electronically available. Among these, two cases: electronically-based online dictionaries, digitized available dictionaries, originally designed for the paper format. We will detail the main steps used to deal with the data.

4.1. Electronically-based online dictionaries

One of the main problems of on-line lexicographical databases derives from their educational purpose, for most of them. These tools are often reduced to small lexicons which don't offer enough linguistic information (mostly only the entry and its definition or translation). We have retrieved around 2000 entries from these resources:

Lexilogos[4] (161 entries): this site is connected which other Creole languages sites and includes a glossary of Caribbean Creole. Examples and variations are available as the main information of every entry, but no morphological marks on inflected marks are included.

Ecrit créole[5] (90 entries): a Caribbean Creole lexicon which offers the translation or the definition for each entry, but no morphosyntactic or inflected information.

Pédagogie [6] (250 entries): a lexicon from Reunion compiled from Jean Albany's « P'tit glossaire » (Ed. Hi-Land O.I.), Jules Bénard's « Petit Glossaire Créole » (Ed. Alizées) and « Dictionnaire illustré de la Réunion ». Its spelling is based on French language.

Petit lexique créole antillais [7] (107 entries): the definition or the translation of every entry is the only available information.

Antan Lontan[8] (124 entries): created by Marie-Andrée Blameble, this Caribbean Creole lexicon contains only a definition or a translation for each entry. Examples are sometimes available.

Dictionnaire créole[9] (477 entries): this lexicon contains word-forms from Haiti, Guadeloupe, Martinique and Reunion Creoles. It does not offer morphological information, examples or references. However, several meanings are often indicated for one entry.

Choubouloute[10] (60 entries): small list of Creole words from Martinique. Sometimes, spelling variants of the entry are included, but no other linguistic information.

Potomitan [11] (614 entries): Raphaël Confiant's lexicon from Martinique. It only includes the translation of every entry.

Créole Réunionnais[12] (112 entries): small lexicon from Reunion which refers to an on-line dictionary / translator containing altogether 5168 words and Creole expressions.

4.2. Paper-based / digitized dictionaries

Although less numerous, these digitized resources contain much more linguistic information. They are generally bilingual resources that require more complex data processing.

[4] http://www.lexilogos.com/creole_langue_dictionnaires.htm
[5] http://ecrit.creole.free.fr/lexique.html
[6] http://pedagogie2.ac-reunion.fr/clglasaline/Disciplines/Creole/lexiquecreole.htm
[7] http://www.ieeff.org/creole.html
[8] http://antanlontan.chez-alice.fr/motscreo.htm
[9] http://www.dictionnaire-creole.com/
[10] http://www.chouboulute.fr/Lexique-Creole.html
[11] http://www.potomitan.info/dictionnaire/francais.php
[12] http://www.mi-aime-a-ou.com/le_creole_reunionnais.htm

Kwéyòl Dictionary[13] (about 4000 entries): this bilingual lexicon (creole-english / english –creole), focuses on the Saint Lucia Creole. Several meanings are indicated, as well as phrases containing the entry. For each entry is indicated: word part of speech, examples in Creole and their English translation, spelling variants, cross-references, semantic relations (synonyms, antonyms) and etymology. This dictionary microstructure has been used as initial model for the database development.

English Creole Dictionary [14] (7500 entries): this bilingual dictionary (english - creole / creole – english) is bsed on the translation of the Bible in Haitian Creole. The microstructure only contains the translation of each lexical entry; the word part of speech is sparsely specified.

Haitian Creole-English Dictionary[15] (8221 entries): this bilingual dictionary of Haitian Creole -English offers a rich microstructure: word part of speech, examples in Creole and English, spelling variants, semantic relations, etymology of the word, polysemy indications, phrases and sometimes other indications (pejorative, familiar word, euphemism, etc.)

Petit Lexique du Créole Haïtien[16] (408 entries): the poet Emmanuel Védrine includes as main information for each lexical entry the syllable division, the translation or definition in French, examples in Creole and French, the word part of speech, spelling variants, semantic relations and sometimes, word etymology.

As a result, we have downloaded more than 20 000 entries from these on-line or digitized resources, with the following microstructure: part of speech, examples in Creole and their English translation, spelling variants, cross-references, semantic relations (synonyms, antonyms) and etymology, source dictionary.

5. Corpora to improve the dictionary

Web Corpora to help lexicographical studies has been a thorough trend since the advent of internet (see for example Kilgariff, 2003; Baroni, 2009). Some systems have also focused on web corpora for minority languages (see for example Scannel, 2007)

As mentioned in the architecture, our project will build its dictionary not only from existing electronic resources, but also by morphosyntactically analyzing corpora (with the help of the previously setup dictionary) and extracting unknown words as a first step. Clearly, iteration of this process will end up with a more exhaustive dictionary, following the Zipf's law.

In this section, we will detail the steps of this iterative process: corpora downloading, automatic analysis, and dictionary improvement. A last point will detail an on-going development that will enable to maintain the dictionary by scanning on a regular basis the web corpus.

5.1. Corpora Building and Feeding

Corpora has been mostly built from Haitian Creole resources, as other varieties are not yet sufficiently represented on the web. We have scheduled three main corpora: a first corpus, limited to 1 million words, will be setup to tune the morphosyntactical analysis and the dictionary improvement process, with in mind the general-purpose language; a second one uses the whole translated Bible, so as to complete and tune the dictionary with a specific vocabulary; the last experiment builds an evolving corpus, so as to maintain the dictionary and setup an adequate environment for lexicographers.

5.1.1. First and second corpora

The first corpus consists in 15 different sources, either manually or automatically downloaded. This corpus aims at rendering as much as possible the varieties of French-based Creole (see (Carrión, 2011 for details on this corpus). The documents all belong to the "general-purpose language". One part has been manually downloaded from websites, whereas automatic download has been setup for two newspapers (« Voa News 2 » and « Alter Presse 3 », both containing Haïtian creole). This corpus finally consists of 1 208 862 words.

The second corpus, the whole Bible translated in Haitian Creole, has been automatically downloaded from the BibleGateway [17] with httrack, with a filter on the webpages link, each containing the keyword "HCV" (*Haitian Creole Version*). It consists of about 4 million words.

Retrieval and cleaning of web pages

The automatic download has been done with Httrack, an open-source software. After the download, it was necessary to convert html files into a text format; this step comprises three different tasks: identification of the encoding and conversion into UTF-8 if applicable, identification and retrieval of the textual zone of interest into the html page, conversion from html to txt. These tasks have been solved by perl scripts, using some previous done work and state-of-the-art techniques (Cartier, 2007, 2009).

Morphosyntactic analysis of the corpus, word frequency of unknown words

The aim of this step is to improve the dictionary by analyzing the given corpus and identifying unknown words sorted by frequency. A Perl program has been used for this task (see Cartier, 2007 for details). From the annotated first and second corpora, we have generated statistics as follows:

Corpus	Recognized words	Unknown tokens	Unknown words / Unique words
Corpus 1(1 208	631984	576878	345653 (28,6%)

[13] www.linguafranka.net/saintluciancreole/dictionary/dictionaryfrontmatter.pdf

[14] www.ngohaiti.com/disaster/downloads/creoledictionary.pdf (*English-Creole Dictionary / Kreyòl-Angle Diksyonè*, publié par Eastern Digital Ressources, 2005)

[15] www.dunwoodypress.com/148/PDF/HCED_sample.pdf

[16] http://www.potomitan.info/vedrine/lexique.pdf

[17] http://www.biblegateway.com

862 tokens)	(52,3%)	(47,7%)	/ 243 892 (70,55%)
Corpus 2 (4 505 442 tokens)	3722628 (82,6%)	782814 (17,4%)	343657 (7,6%) / 336 896 (98,03%)

Table 1: corpora statistics

The following remarks apply to these figures:

1/ The first outstanding element concerns the quick lexicographic coverage of the dictionary: whereas the first corpus consists of 28,6% unknown words, with about 70% unique words, these figures fall down with the second corpus (where these unknown words are integrated in the dictionary) to 7,6%, with about 98% of unique words; this is a confirmation of Zipf's law (see Manning et al, 1999) : about 20% of words represent about 90% of word occurrences, and about 80% of words represent about 10% of occurrences; the consequence is also that whereas a relatively small corpus enable to cover 90% of lexicographic entries, only a really huge corpus enable to tend to 100% coverage.

2/ Dictionary coverage : at the end of the two processes, the dictionary is composed of 123245 unique words-forms; this is congruent with the dictionary coverage of main language, considering that Creole language is not morphologically rich; nevertheless, this coverage has to be tuned with the fact that our processes do not recognize phrases, whereas they represent at least 50% of the vocabulary (see Sag et al, 2002, for example); it has also to be tuned with the fact that our processes integrate unknown words without linguistic information, as the processes has been automated. Finally, spelling variants have still to be gathered.

Analysis of unknown words

Analysis of unknown words is just a first step of our process. In fact, to really connecting dictionary to corpora, it would be necessary to have automatic procedures to track the meaning evolutions, rather than word forms existence. This step is presented in the next section.

Unknown words retrieved from corpus belong to various categories: words from other language (specifically from English, in our case), misspelled words, proper names, specific notations, real unknown words. Clearly, we have to remove all but the last category. This first filtering results in the figures of the third column, table 1. The resulting list has been included in the dictionary without any linguistic information, except for a small part of it with information taken from web-based dictionary (that could not be retrieved globally, but can be used for individual word search) or paper-based dictionaries (see Carrión, 2011 for the list of these dictionaries). For each of these words, we have decided, in a first step, to include only part of speech, translation in French and English, and varieties if applicable.

5.1.2. A live-corpus and an environment for lexicographers

The experiments done have enabled to complete the dictionary substantially. But our goal is not only to complete the dictionary but also to maintain it, that is check continuously the life of words: emerging, stabilization, meaning changes, disappearance. Towards this goal, we have glued existing environments dealing with corpora handling. Among existing systems, we have retained two systems: SketchEngine (Kilgariff, 2004) and the IMS Workbench (Christ, 1994; Evert, 2011); the first one is certainly the most complete, as it proposes a web crawler and several statistical and linguistic tools to search the corpora; but its main drawback is that it is not freely available. The second environment is also really interesting because it is free and it uses one of the most powerful Linguistic search tools: CQP. But it does not include any tool to retrieve the web nor tools to convert it to the environment internal format.

As a result, we have decided to combine various tools available: a customized web crawler to retrieve corpora; Textbox for conversion to XML and morphosyntactical analysis, mwetoolkit to generate statistics and CWB to search for the corpus, as well as CQPWeb to have a web-based graphical environment for lexicographers. As this project is on-going, we will not detail it in this paper.

6. Conclusion

This paper has presented an on-going project whose goals are: 1/ to explicit a methodology to improve NLP and linguistics development and research for minority-languages, focusing on the American-Caribbean creoles; 2/ to setup procedures and finally an environment for lexicographers to store and maintain lexicographical data from existing resources and web corpora.

It is also important to specify that this project would provide not only a normalizing educational tool, but a translation tool that may be of great help for “mixed” literatures translation and understanding.

According to the general architecture of the project, we have first compiled existing lexicographical resources, either paper or electronically-based; this step permitted to gather about 20 000 lexicographical entries, but exhibited complex-to-solve sparsity problems, as quality, quantity diverge from one resource to another, and macro and micro-structures are far from unified. We have then build a Part-of-Speech tagger from this resource and, using web corpora, have begun to complete the dictionary essentially from unknown words. This step has revealed to be a good procedure, and has to be continued with a live corpus, so as to attain the Zipf's law limit. Finally, we have begun to setup a web-based environment to maintain the dictionary and study lexicographical phenomena through several iterative processes: morphosyntactical analysis and retrieving of unknown words; continuous downloading of web pages; statistical measures over the corpora. This project has generated two crucial elements for the American-Caribbean creoles: a POS annotated corpus and a POS tagger. These data and tools will be soon released as open-source.

In the near future, we have in mind two main tasks: inclusion of the main existing dictionary in the field, the DECA; a theoretical study of the microstructure for the dictionary. We essentially hope that this contribution will help minority-languages to be more considered and studied, through NLP procedures already in use for the major languages, and crucial tools for lexicographers and language practitioners.

7. BIBLIOGRAPHY

Allen, J. (1998) *Lexical variation in Haitian Creole and orthographic issues for Machine Translation (MT) and Optical Character Recognition (OCR) applications.* First Workshop on Embedded Machine Translation systems of the Association for Machine Translation in the Americas (AMTA), Philadelphia, 28 octobre 1998.

Baker P. And Hookoomsing V. Y. (1987) *Diksyoner kreol morisyen: morisyen-English-français*, l'Harmattan

Baroni M. Et Bernardini S. (2004), “BootCaT: Bootstrapping Corpora and Terms from the Web”,in *Proceedings of LREC 2004,*Lisbon, Portugal.

Baroni M. Et Kilgarriff A. (2006), “Large linguistically-processed Web corpora for multiple languages”, in *Proceedings of the 11th EACL Conference*,Trento, Italy.

Baroni M., Kilgarriff, Pomikálek, Rychlý (2006), WebBootCaT: a web tool for instant corpora, *Proc. Euralex*. Torino, Italy.

Baroni, M. , Bernardini S., Ferraresi A. And Zanchetta E.. (2009). The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation*43(3): 209-226

Baroni, M. And Bernardini, S. (Eds.) (2006). Wacky! *Working papers on the Web as Corpus*. Bologna:

Bollee A., Brasseur P., Chaudenson R. Et Chauveau J-P. (1993) *Dictionnaire étymologique des créoles français de l'Océan Indien*, Hamburg: H. Buske

Bollee, A. (1993) *Dictionnaire étymologique des créoles français de l'Océan Indien*, Hamburg: H. Buske

Bollee, A. (2005). Lexicographie créole: problèmes et perspectives, *Revue française de linguistique appliquée* (Vol. X), p. 53-63.

Carpooran, A. (2009). *Diksioner Morisien* (version intégrale), 1017.

Carrion, P. (2011*) Le traitement automatique des langues créoles à base lexicale française*, Master Dissertation (TILDE – Traitement Automatique et Linguistique des Documents Ecrits), Villetaneuse: Université de Paris 13, 2011.

Cartier E. (2007) "TextBox, a Written Corpus Tool for Linguistic Analysis". In FAIRON Cédrick, NAETS Hubert, KILGARRIFF Adam, DE SCHRYVER Gilles-Maurice, (eds), *Building and Exploring Web Corpora (WAC3 - 2007), Cahiers du CENTAL* 4, pp. 33-42. Presses universitaires de Louvain. Louvain-la-Neuve.

Cartier E. (2009) "Corpus for linguistic resources building and maintenance (CLRBM): system architecture and first experiments”, in *5th Corpus Linguistics 2009*, 20-23 juillet 2009, Liverpool

Chaudenson, R (1974) *Le lexique du parler créole de la Réunion*, 2 tomes. Paris: Champion, 1974.

Christ, O. (1994). A modular and flexible architecture for an integrated corpus query system. In *Papers in Computational Lexicography (COMPLEX '94),* pages 22–32, Budapest, Hungary.

Colson, J.-P. (2010a). The Contribution of Web-based Corpus Linguistics to a Global Theory of Phraseology. In: Ptashnyk, S., Hallsteindóttir, E. & N. Bubenhofer (eds.), *Corpora, Web and Databases. Computer-Based Methods in Modern Phraselogy and Lexicography*. Hohengehren, Schneider Verlag, p. 23-35.

Colson, J.-P. (2010b). Automatic extraction of collocations: a new Web-based method. In: S. Bolasco, S., Chiari, I. & L. Giuliano, *Proceedings of JADT 2010,Statistical Analysis of Textual Data*, Sapienza University of Rome, 9-11 June 2010. Milan: LED Edizioni, p. 397-408.

Confiant, R. (2007) *Dictionnaire créole martiniquais-français,* Editions Ibis rouge

Crosbie P., Frank D., Leon E. Et Samuel P. (2001). *Kwéyòl Dictionary*, Castries (Sainte-Lucie), St. Lucia Ministry of Education - SIL International, 2001

Evert, S. And Hardie, A. (2011). *Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium*. Presentation at Corpus Linguistics 2011, University of Birmingham, UK.

Faaß ET AL. (2010). Gertrud Faaß, Ulrich Heid, and Helmut Schmid. Design and application of a Gold Standard for morphological analysis: SMOR in validation. In *Proceedings of the seventh LREC conference* , pages 803 – 810, Valetta, Malta, May 19 – 21 2010. European Language Resources Association (ELRA).

Fattier D. (1998). *Contribution à l’étude de la genèse d’un créole : L’Atlas Linguistique d’Haïti, cartes et commentaires.* Lille, ANRT, collection « thèse à la carte », 6 volumes. 3300p.

Fattier, D. (2007) *Le Projet de Dictionnaire Etymologique des Créoles Français d'Amérique (DECA)*, Université de Cergy-Pontoise, Séminaire du 12 octobre 2007

Ferraresi A. (2007) *Building a very large corpus of English obtained by Web crawling: ukWaC*. Master Thesis, University of Bologna

Ferraresi A., Bernardini S., Picci G. And Baroni M. (2010) “Web Corpora for Bilingual Lexicography: A Pilot Study of English/French Collocation Extraction and Translation”. In Xiao, R. (ed.*) Using Corpora in Contrastive and Translation Studies*. Newcastle: Cambridge Scholars Publishing.

Ferraresi A., Zanchetta E., Baroni M. And Bernardini S. (2008) Introducing and evaluating ukWaC, a very large web-derived corpus of English. In S. Evert, A. Kilgarriff and S. Sharoff (eds.) *Proceedings of the 4th Web as Corpus Workshop (WAC-4) – Can we beat Google?*, Marrakech, 1 June 2008.

Frank D., Crosbie P., Leon E. Et Samuel P. (2001) *Kwéyòl Dictionary*, Castries (Sainte-Lucie)

Hazaël-Massieux, M.-C. (2002). *Prolégomènes à une néologie créole,* en RFLA, 2002

Hearn, L.(1885). *Little dictionary of Creole proverbs, selected from 6 Creole dialects*, translated into French and into English, with notes, complete index to subjects and some brief remarks upon the Creole idioms of Louisiana

Kilgarriff A. (2001), "Web as corpus", in *Proceedings of the Corpus Linguistics 2001* Conference, Lancaster University : 342–344.

Kilgarriff A. Et Grefenstette G. (2003), "Introduction to the Special Issue on the Web as Corpus", in *Computational Linguistics*, no 3, vol. 29.

Kilgarriff A., Rychly P., Smrz P., Tugwell D. (2004) The Sketch Engine. *Proc EURALEX 2004*, Lorient, France; Pp 105-116, (http://www.sketchengine.co.uk)

Manning C. D., Schütze H. (1999) *Foundations of Statistical Natural Language Processing*, MIT Press (1999), p. 24

Mondesir, J. E. (1992) *Dictionary of St. Lucian Creole*, Mouton de Gruyter, Berlin, Allemagne

Poullet H. Et Telchid S. (1990) *Le créole sans peine (guadeloupéen),* Assimil

Poullet H. Et Telchid S. (1999) *Le créole guadeloupéen de poche*, Assimil

Poullet H., Telchid S. Et Montbrand D. (1984) *Dictionnaire des expressions du créole guadeloupéen*, Hatier-Antilles

Rychly P. (2008). A Lexicographer-Friendly Association Score. Proc. 2nd Workshop on *Recent Advances in Slavonic Natural Languages Processing*, RASLAN 2008. Eds Sojka P., Horák A. prvni. Brno : Masaryk University.

Scannel, K. (2007), The Crúbadán Project: Corpus building for under-resourced languages, Cahiers du Cental 4 (2007), pp5-15, 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.

Sharoff S. (2006a), "Creating General-Purpose Corpora Using Automated Search Engine Queries", in Baroni M. & Bernardini S. (Eds), *WaCky! Working Papers on the Web as Corpus*, GEDIT, Bologna.

Sharoff S. (2006b), "Open-source corpora: Using the net to fish for linguistic data", in *International Journal of Corpus Linguistics,* no 4, vol. 11.

Targete J., Urciolo R. G. (1993) *Haitian Creole – English Dictionary with Basic English – Haitian Creole Append*, dp Dunwoody Press, Kensington, Maryland, U.S.A.

Telchid S., Poullet R.Et Anciaux F. (2009*). Le Déterville: dictionnaire français-créole,* PLB éditions

Tourneux H., Barbotin M. Et Tancons M.-H. (2009) *Dictionnaire pratique du créole de Guadeloupe: Marie-Galante: suivi d'un index français-créole*, éd. Karthala

Valdman A., Pooser C. Et Rosevelt J.-B. (1996) *A Learner's Dictionary of Haitian Creole*, Creole Institute, Indiana University, Bloomington, USA

Valdman A., Yoder S.., Roberts C.. Et Yoseph Y. (1981) *English-French dictionary,* Indiana University, Creole institute

Valdman, A. (2007) *Haitian Creole-English Bilingual Dictionary*, Indiana University, Creole Institute

Vedrine, E. W. (2005) *Petit Lexique du Créole Haïtien*, Orèsjozèf Publications, Boston, Massachusetts (USA), 2nd. ed

Describing Morphologically-rich Languages using Metagrammars: a Look at Verbs in Ikota

Denys Duchier[1], Brunelle Magnana Ekoukou[2], Yannick Parmentier[1],
Simon Petitjean[1], Emmanuel Schang[2]

(1) LIFO, Université d'Orléans - 6, rue Léonard de Vinci 45067 Orléans Cedex 2 – France
(2) LLL, Université d'Orléans - 10, rue de Tours 45067 Orléans Cedex 2 – France

prenom.nom@univ-orleans.fr

Abstract

In this paper, we show how the concept of metagrammar originally introduced by Candito (1996) to design large Tree-Adjoining Grammars describing the syntax of French and Italian, can be used to describe the morphology of Ikota, a Bantu language spoken in Gabon. Here, we make use of the expressivity of the XMG (eXtensible MetaGrammar) formalism to describe the morphological variations of verbs in Ikota. This XMG specification captures generalizations over these morphological variations. In order to produce the inflected forms, one can compile the XMG specification, and save the resulting electronic lexicon in an XML file, thus favorising its reuse in dedicated applications.

1. Introduction

Bantu languages form a large family of languages in Africa. In this family, Chichewa and Swahili are the most well-studied, and are used as benchmarks for assessing the expressivity and relevance of morphological theories (Mchombo, 1998; Stump, 1992; Stump, 1998; Stump, 2001) and their implementation (Roark and Sproat, 2007).

Ikota (B25) is a lesser-known language of Gabon and the Democratic Republic of Congo. Language of the Bakota people, with an estimated 25000 speakers in Gabon (Idiata, 2007), Ikota is threatened with extinction mainly because of its abandon for French (the official language of Gabon).

It manifests many grammatical features shared by the Bantu languages (Piron, 1990; Magnana Ekoukou, 2010):

- Ikota is a *tonal language* with two registers (High and Low):

(1) a. ìkàká ”family”
b. ìkákà ”palm”

(2) a. nkúlá ”year”
b. nkúlà ”pygmee”

- Ikota has ten *noun classes*,[1] see Table 1.

- Ikota has a *widespread agreement in the NP*:

(3)					
b-àyítò	**bá**-nɛ́nì	**b**-á	Ø-mbókà	**bà**-tɛ́	**b**-à-ʤá
2-women	2-fat	2-of	9-village	2-DEM	2-Prst-eat

”These fat women of the village are eating”

- Yet, unlike Swahili for instance, Ikota does not have a slot for object agreement.

In this paper, we will consider verbal morphology.

Table 1: Ikota's noun classes

Noun class	prefix	allomorphs
CL 1	mò-, Ø-	mw-, ǹ-
CL 2	bà-	b-
CL 3	mò-, Ø-	mw-, ǹ-
CL 4	mè-	
CL 5	ì-, ʤ-	dy-
CL 6	mà-	m-
CL 7	è-	
CL 8	bè-	
CL 9	Ø-	
CL 14	ò-, bò-	bw-

Production of a lexicon of inflected forms. Our purpose is twofold: first to provide a formal description of the morphology of verbs in Ikota; second, to automatically derive from this description a lexicon of inflected forms. To do so, we propose to adopt the concept of a metagrammar, which was introduced by (Candito, 1996), and used to describe the syntax of Indo-European languages, such as French, English or Italian. Lexicalized wide-coverage tree-grammars for natural languages are very large and extremely resource intensive to develop and maintain. For this reason, they are often automatically produced by software from a highly modular formal description called a metagrammar. The metagrammar is much easier to develop and to maintain.

We propose to adopt a similar strategy to capture morphological generalizations over verbs in Ikota. The outline of the paper is the following. In Section 2., we give a detailed presentation of the morphology of verbs in Ikota. Then, in Section 3., we introduce eXtensible MetaGrammar (XMG), a formal language, used to describe and combine reusable descriptive fragments. In Section 4., we show how to use the XMG framework to describe the morphology of verbs in Ikota. Concretely, we present a metagrammar of verbs in Ikota, which we have also coded in the XMG language, and which can be automatically processed to produce a lexicon

[1] The number of the class in the table corresponds to Meinhof's numbering.

of fully inflected verb forms in Ikota. Finally, in Section 5., we conclude and present future work.

2. Verbs in Ikota

Verbs are constituted by a lexical root (VR) and several affixes distributed on each side of the VR. For the sake of clarity, we will focus here on the basic verbal forms, leaving aside Mood and Voice markers.

Let us now describe infinitival form and the three verbal classes of Ikota.

Verbs in Ikota are distributed in three classes depending on the form of Aspect and Active markers. Infinitive in Ikota is a hybrid word class. It is composed of a noun class prefix (class 14) and a verbal element (VR+Prog+Active).

(4) a. bòʤákà “to eat”
b. bòwɛ́ʧɛ̀ “to give”
c. bòbɔ́nɔ́kɔ̀ “to choose”

Examples (4) illustrate the three verb classes.

Indeed, it seems that the suffix (Prog+Active) has a subjacent form VkV. In the Makokou variant of Ikota, /k/ is realized by [ʧ] when the vowel is [ɛ]. In Standard Ikota, the form is ɛ́kɛ̀.

At a subjacent level, the structure of the infinitival suffix boils down to AKA, with three distinct surface realizations ákà, ɛ́ʧɛ̀, ɔ́kɔ̀.

Examples below illustrate the conjugation of bòʤákà “to eat”, a typical example of the *aka* verb class:

(5) m-à-ʤ-á ǹlɛ́sì
1sg-Prst-eat-Act rice
“I’m eating rice” (Present)

(6) a. m-à-ʤ-á-ná yàná
1sg-Past-eat-Act-Prox yesterday
“I ate yesterday” (Past (yesterday))

b. m-à-ʤ-á-sá kúlá mwáyèkànàmwɛ́
1sg-Past-eat-Act-Prox year last
“I ate last year” (Distant Past)

c. m-é-ʤ-á ǹlɛ́sì
1sg-Past-eat-Act rice
“I ate rice” (Recent Past)

(7) a. m-é-ʤ-àk-à ǹlɛ́sì
1sg-Fut-eat-Asp-Act rice
“I’ll eat rice” (Medium Future)

b. m-é-ʤ-àk-à-ná yàná
1sg-Fut-eat-Asp-Act-Prox tomorrow
“I’ll eat tomorrow” (Future (tomorrow))

c. m-é-ʤ-àk-à-sá kúlá
1sg-Fut-eat-Asp-Act-Prox year
mwáyàkàmwɛ́
next
“I’ll eat next year” (Distant Future)

d. m-ábí-ʤ-àk-à òsátè
1sg-Fut-eat-Asp-Act soon
“I’ll eat soon” (Imminent Future)

As can be deduced from the examples above, Ikota’s verbal affixes ordering can be defined as position classes. From the left to the right:

- the class of Subject agreement prefixes occupies the leftmost, word-initial position.
- Tense prefixes (or what can roughly identified as related to Tense) appears at the left of VR.
- the (aspectual) progressive marker is on the immediate right of VR.
- Active suffix occupies the slot to the left of Proximal. It has two values: Active and Passive (to wit: -Active). Applicative and Causative are kept for further studies.
- the Proximal/Distal suffixes occupy the rightmost position.

Table 3 gives an outline of the VR and its affixes and table 2 exemplifies this schema with bòʤákà “to eat”.

Table 3: Verb formation

Subj-	Tense-	VR	-(Aspect)	-Active	-(Proximal)

3. eXtensible MetaGrammar

eXtensible MetaGrammar (XMG) refers both to a formal language (*a kind of* programming language) and a piece of software, called a compiler, that processes descriptions written in the XMG language (Crabbé and Duchier, 2004). XMG is normally used to describe lexicalized tree grammars. In other words, an XMG specification is a declarative description of the tree-structures composing a grammar. This description relies on four main concepts: (1) **abstraction**: the ability to associate a content with a name, (2) **contribution**: the ability to accumulate information in any level of linguistic description, (3) **conjunction**: the ability to combine pieces of information, (4) **disjunction**: the ability to non-deterministically select pieces of information.

Formally, one can define an XMG specification as follows:

$$\begin{array}{rcl} \mathit{Rule} & := & \mathit{Name} \rightarrow \mathit{Content} \\ \mathit{Content} & := & \mathit{Contribution} \ \mid \ \mathit{Name} \ \mid \\ & & \mathit{Content} \vee \mathit{Content} \ \mid \ \mathit{Content} \wedge \mathit{Content} \end{array}$$

An abstraction is expressed as a rewrite rule that associates *Content* with a *Name*. Such content is either the *Contribution* of a fragment of linguistic description (e.g. a tree fragment contributed to the description of syntax) or an existing abstraction, or a conjunction or disjunction of contents.

One abstraction must be specifically identified as the axiom of the metagrammar. The XMG compiler starts from this axiom and uses the rewrite rules to produce a full derivation. When a disjunction is encountered, it is interpreted

Table 2: Verbal forms of bòʤákà ”to eat”

Subj.	Tense	VR	Aspect	Active	Prox.	Value
m-	à-	ʤ		-á		present
m-	à-	ʤ		-á	-ná	past, yesterday
m-	à-	ʤ		-á	-sá	distant past
m-	é-	ʤ		-á		recent past
m-	é-	ʤ	-àk	-à		medium future
m-	é-	ʤ	-àk	-à	-ná	future, tomorrow
m-	é-	ʤ	-àk	-à	-sá	distant future
m-	ábí-	ʤ	-àk	-à		imminent future

as offering alternative ways to proceed: the compiler successively explores each alternative. In this fashion, the execution of a metagrammar typically produces many derivations. Along one derivation, contributions are simply accumulated conjunctively. At the end of a derivation, the accumulated contributions are interpreted as a specification and given to a solver to produce solution structures. The collection of all structures produced in this manner forms the resulting grammar. It can be inspected using a graphical tool, or exported in an XML format.

The XMG compiler is freely available under a GPL-compliant license, and comes with reasonable documentation.[2] It has been used to design various large tree-based grammars for French (Crabbé, 2005; Gardent, 2008), English (Alahverdzhieva, 2008) and German (Kallmeyer et al., 2008).

XMG was expressedly designed for writing wide-coverage high-level modular tree-grammmars covering both syntactic expression and semantic content. While XMG was never intended for expressing morphology, our current project demonstrates that it can successfully be repurposed for the task, at least in the case of the agglutinative Ikota language.

4. Metagrammar of Ikota verbal morphology

Our formalization of Ikota verbal morphology borrows the notion of *topological domain* from the tradition of German descriptive syntax (Bech, 1955). A topological domain consists of a linear sequence of fields. Each field may host contributed material, and there may be restrictions on how many items a particular field may/must host. For our purposes, the topological domain of a verb will be as described in Table 3, and each field will hold at most 1 item, where an item is the *lexical phonology*[3] of a morpheme.

Elementary blocks. The metagrammar is expressed in terms of elementary blocks. A block makes simultaneous contributions to 2 distinct dimensions of linguistic description: (1) lexical phonology: contributions to fields of the topological domain, (2) inflection: contributions of morphosyntactic features. For example:

2 ← é

tense = `past`
proxi = `near`

contributes é to field number 2 of the topological domain, and features tense = `past` and proxi = `near` to the inflection. Feature contributions from different blocks are unified: in this way, the inflection dimension also acts as a coordination layer during execution of the metagrammar. As Table 2 illustrates clearly, ikota morphology is not cleanly compositional: instead, the semantic contributions of morphemes are determined by mutually constrained coordination through the inflection layer.

Morphosyntactic features. We use p and n for *person* and *number*; tense with possible values `past`, `present`, and `future`; proxi for the *proximal marker* (`none`, `imminent`, `day`, `near`, `far`); vclass for the verbal class (`g1`, `g2`, `g3`); and two polar features: active for *voice* and prog for the *progressive aspect*: prog=`-` marks an eventuality yet unrealized.

Lexical phonetic signs. Careful consideration of Ikota data suggests that regularities across verbal classes can be better captured by the introduction of a *lexical* vowel `A` which is then realized, at the surface level, by a for `vclass=g1`, ɛ for `vclass=g2`, and ɔ for `vclass=g3`, and lexical consonant `K` which is realized by tʃ for `vclass=g2`, and k otherwise.

Rules. Figure 1 shows a fragment of our preliminary metagrammar of Ikota verbal morphology. Each rule defines how an abstraction can be rewritten. For example *Tense* can be rewritten as any one block from a disjunction of 5 blocks. To produce the lexicon of inflected forms described by our metagrammar, the XMG compiler computes all possible non-deterministic rewritings of the *Verb* abstraction.

Example derivation. Let's consider how óʤàkàná (*tomorrow, you will eat*) is derived by our formal system starting from the *Verb* abstraction. First *Verb* is replaced by *Subj* ∧ *Tense* ∧ *VR* ∧ *Aspect* ∧ *Active* ∧ *Proximal*. Then each element of this logical conjunction (order is irrelevant) is, in turn, expanded. For example, *Subj* is then replaced by one block from the corresponding disjunction: the XMG compiler tries all possibilities; eventually it chooses the 2nd block. Figure 2 shows the initial step, a middle step, and the final step of the derivation. The lexical phonology of

[2] See `http://spourcesup.cru.fr/xmg`

[3] We adopt here the *two-level* perspective of lexical and surface phonology (Koskenniemi, 1983)

Figure 1: Metagrammar of Ikota verbal morphology

$$
\begin{array}{rcl}
\textit{Subj} & \rightarrow & \begin{array}{|c|}\hline 1 \leftarrow \text{m} \\\hline \text{p} = 1 \\ \text{n} = \mathtt{sg} \\\hline\end{array} \vee \begin{array}{|c|}\hline 1 \leftarrow \text{ò} \\\hline \text{p} = 2 \\ \text{n} = \mathtt{sg} \\\hline\end{array} \vee \ldots \\[1ex]
\textit{Tense} & \rightarrow & \begin{array}{|c|}\hline 2 \leftarrow \text{é} \\\hline \text{tense} = \mathtt{past} \\ \text{proxi} = \mathtt{near} \\\hline\end{array} \vee \begin{array}{|c|}\hline 2 \leftarrow \text{é} \\\hline \text{tense} = \mathtt{future} \\\hline\end{array} \vee \begin{array}{|c|}\hline 2 \leftarrow \text{à} \\\hline \text{tense} = \mathtt{present} \\\hline\end{array} \vee \begin{array}{|c|}\hline 2 \leftarrow \text{à} \\\hline \text{tense} = \mathtt{past} \\ \text{proxi} = \neg\mathtt{near} \\\hline\end{array} \vee \begin{array}{|c|}\hline 2 \leftarrow \text{ábí} \\\hline \text{tense} = \mathtt{future} \\ \text{proxi} = \mathtt{imminent} \\\hline\end{array} \\[1ex]
\textit{Active} & \rightarrow & \begin{array}{|c|}\hline 5 \leftarrow \mathtt{À} \\\hline \text{active} = + \\ \text{prog} = - \\\hline\end{array} \vee \begin{array}{|c|}\hline 5 \leftarrow \mathtt{Á} \\\hline \text{active} = + \\ \text{prog} = + \\\hline\end{array} \vee \begin{array}{|c|}\hline 4 \leftarrow \text{ébwɛ̀} \\\hline \text{active} = - \\\hline\end{array} \\[1ex]
\textit{Aspect} & \rightarrow & \begin{array}{|c|}\hline 4 \leftarrow \mathtt{ÀK} \\\hline \text{tense} = \mathtt{future} \\ \text{prog} = - \\\hline\end{array} \vee \begin{array}{|c|}\hline \\\hline \text{tense} = \neg\mathtt{future} \\ \text{prog} = + \\\hline\end{array} \\[1ex]
\textit{Proximal} & \rightarrow & \begin{array}{|c|}\hline 6 \leftarrow \mathtt{nÁ} \\\hline \text{proxi} = \mathtt{day} \\\hline\end{array} \vee \begin{array}{|c|}\hline 6 \leftarrow \mathtt{sÁ} \\\hline \text{proxi} = \mathtt{far} \\\hline\end{array} \vee \begin{array}{|c|}\hline \\\hline \text{proxi} = \mathtt{none} \vee \mathtt{near} \\\hline\end{array} \vee \begin{array}{|c|}\hline \\\hline \text{proxi} = \mathtt{imminent} \\ \text{tense} = \mathtt{future} \\\hline\end{array} \\[1ex]
\textit{To-Eat} & \rightarrow & \begin{array}{|c|}\hline 3 \leftarrow \text{ʤ} \\\hline \text{vclass} = \mathtt{g1} \\\hline\end{array} \\[1ex]
\textit{To-Give} & \rightarrow & \begin{array}{|c|}\hline 3 \leftarrow \text{w} \\\hline \text{vclass} = \mathtt{g2} \\\hline\end{array} \\[1ex]
\textit{VR} & \rightarrow & \textit{To-Eat} \vee \textit{To-Give} \\
\textit{Verb} & \rightarrow & \textit{Subj} \wedge \textit{Tense} \wedge \textit{VR} \wedge \textit{Aspect} \wedge \textit{Active} \wedge \textit{Proximal}
\end{array}
$$

Figure 2: A successful derivation

$$
\begin{array}{rcl}
\textit{Verb} & \rightarrow & \textit{Subj} \wedge \textit{Tense} \wedge \textit{VR} \wedge \textit{Aspect} \wedge \textit{Active} \wedge \textit{Proximal} \\[1ex]
 & \rightarrow & \begin{array}{|c|}\hline 1 \leftarrow \text{ò} \\\hline \text{p} = 2 \\ \text{n} = \mathtt{sg} \\\hline\end{array} \wedge \begin{array}{|c|}\hline 2 \leftarrow \text{é} \\\hline \text{tense} = \mathtt{future} \\\hline\end{array} \wedge \begin{array}{|c|}\hline 3 \leftarrow \text{ʤ} \\\hline \text{vclass} = \mathtt{g1} \\\hline\end{array} \wedge \begin{array}{|c|}\hline 4 \leftarrow \mathtt{ÀK} \\\hline \text{tense} = \mathtt{future} \\ \text{prog} = - \\\hline\end{array} \wedge \begin{array}{|c|}\hline 5 \leftarrow \mathtt{À} \\\hline \text{active} = + \\ \text{prog} = - \\\hline\end{array} \wedge \begin{array}{|c|}\hline 6 \leftarrow \mathtt{nÁ} \\\hline \text{proxi} = \mathtt{day} \\\hline\end{array} \\[1ex]
 & \rightarrow & \begin{array}{|cccc|}\hline \multicolumn{4}{|c|}{1 \leftarrow \text{ò} \quad 2 \leftarrow \text{é} \quad 3 \leftarrow \text{ʤ} \quad 4 \leftarrow \mathtt{ÀK} \quad 5 \leftarrow \mathtt{À} \quad 6 \leftarrow \mathtt{nÁ}} \\\hline \text{p} = 2 & \text{prog} = - & \text{tense} = \mathtt{future} & \text{vclass} = \mathtt{g1} \\ \text{n} = \mathtt{sg} & \text{active} = + & \text{proxi} = \mathtt{day} & \\\hline\end{array}
\end{array}
$$

the resulting lexicon entry is obtained by concatenating, in the linear order of the topological domain, the material contributed to the various fields; here: ò+é+ʤ+ÀK+À+nÁ.

Figure 3 shows an example of a failed derivation, i.e. one which does not lead to the production of a lexicon entry. The failure is due to clashing values for feature tense (future and ¬future) and also for feature prog (+ and -).

Surface phonology. At present, our metagrammar models only the lexical level of phonology. The surface level can subsequently be derived by postprocessing. For our example, since vclass=g1, the lexical A becomes a on the surface, and K becomes k. Thus we obtain: ò+é+ʤ+à+à+ná, and finally (through vowel deletion) óʤàkàná.

XMG's constraint-based approach makes it ideally suited to a seamless integration of e.g. *two-level phonology* since the latter is precisely a constraint between lexical and surface phonology (Koskenniemi, 1983). This extension of XMG is a planned milestone of an ongoing thesis.

Caveats. Our formalization of Ikota morphology is very preliminary. As we progress, questions arise for which we do not yet have sufficient data. For example, as can be readily deduced from Figure 1, our current metagrammar (deliberately) omits the "passive future" pending further evidential data from native speakers.

Also, it is too early for us to suggest even a tentative account of Ikota's tonal system and its implications on e.g. the prosodic contours of verb forms. As a consequence, in the interest of accurate descriptive morphology, we have been forced to adopt some tricks, in the formal description, as a practical recourse rather than as a theoretical proposal: such is the case of the tonal alternation in the active voice.

Figure 3: A failed derivation: clashes on tense and on prog

$$\begin{array}{lll}
\textit{Verb} & \rightarrow & \textit{Subj} \wedge \textit{Tense} \wedge \textit{VR} \wedge \textit{Aspect} \wedge \textit{Active} \wedge \textit{Proximal} \\
 & \rightarrow & \begin{array}{|c|}\hline 1 \leftarrow \text{ò} \\ \hline \text{p} = 2 \\ \text{n} = \texttt{sg} \\ \hline\end{array} \wedge \begin{array}{|c|}\hline 2 \leftarrow \text{é} \\ \hline \text{tense} = \texttt{future} \\ \hline\end{array} \wedge \begin{array}{|c|}\hline 3 \leftarrow \text{ʤ} \\ \hline \text{vclass} = \texttt{g1} \\ \hline\end{array} \wedge \begin{array}{|c|}\hline \\ \hline \text{tense} = \neg\texttt{future} \\ \text{prog} = + \\ \hline\end{array} \wedge \begin{array}{|c|}\hline 5 \leftarrow \texttt{À} \\ \hline \text{active} = + \\ \text{prog} = - \\ \hline\end{array} \wedge \begin{array}{|c|}\hline 6 \leftarrow \texttt{nÁ} \\ \hline \text{proxi} = \texttt{day} \\ \hline\end{array} \\
 & \rightarrow & \textbf{failure!}
\end{array}$$

5. Conclusion and future work

In this article, we proposed a formal, albeit preliminary, declarative description of verbal morphology in Ikota, an arguably minority African language. In so doing, we illustrated how the metagrammatical approach can usefully contribute to African language technology.

Additionally, from this formal description, using the XMG compiler, we are able to automatically produce a lexicon of fully inflected verb forms with morphosyntactic features. This lexicon can be saved in XML format, thus providing an easily reusable normalization resource for this less-resourced language.

From a methodological point of view, the use of XMG for expressing our ideas has made it easy to quickly test them by generating the predicted verb forms and their features and then validating the results against the available data.

A further advantage of adopting the metagrammar approach is that, using the same tool, we will be able to also describe the syntax of the language using e.g. tree-adjoining grammars (the topic of an ongoing PhD thesis).

6. References

Katya Alahverdzhieva. 2008. XTAG using XMG. Master Thesis, Nancy Université.

Gunnar Bech. 1955. *Studien über das deutsche Verbum infinitum*. Det Kongelige Danske videnskabernes selskab. Historisk-Filosofiske Meddelelser, bd. 35, nr.2 (1955) and bd. 36, nr.6 (1957). Munksgaard, Kopenhagen. 2nd unrevised edition published 1983 by Max Niemeyer Verlag, Tübingen (Linguistische Arbeiten 139).

Marie Candito. 1996. A Principle-Based Hierarchical Representation of LTAGs. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING'96)*, volume 1, pages 194–199, Copenhagen, Denmark.

Benoît Crabbé and Denys Duchier. 2004. Metagrammar redux. In Henning Christiansen, Peter Rossen Skadhauge, and Jørgen Villadsen, editors, *Constraint Solving and Language Processing, First International Workshop (CSLP 2004), Revised Selected and Invited Papers*, volume 3438 of *Lecture Notes in Computer Science*, pages 32–47, Roskilde, Denmark. Springer.

Benoît Crabbé. 2005. *Représentation informatique de grammaires fortement lexicalisées : Application à la grammaire d'arbres adjoints*. Ph.D. thesis, Université Nancy 2.

Claire Gardent. 2008. Integrating a Unification-Based Semantics in a Large Scale Lexicalised Tree Adjoining Grammar for French. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 249–256, Manchester, UK, August. Coling 2008 Organizing Committee.

Daniel Franck Idiata. 2007. *Les langues du Gabon: données en vue de l'élaboration d'un atlas linguistique*. L'Harmattan.

Laura Kallmeyer, Timm Lichte, Wolfgang Maier, Yannick Parmentier, and Johannes Dellert. 2008. Developing a TT-MCTAG for German with an RCG-based Parser. In *The sixth international conference on Language Resources and Evaluation (LREC 08)*, pages 782–789, Marrakech, Morocco.

Kimmo Koskenniemi. 1983. *Two-Level Morphology: a general computational model for word-form recognition and production*. Ph.D. thesis, University of Helsinki.

Brunelle Magnana Ekoukou. 2010. Morphologie nominale de l'ikota (B25): inventaire des classes nominales. Mémoire de Master 2, Université d'Orléans.

Sam A. Mchombo. 1998. Chichewa: A Morphological Sketch. In Andrew Spencer and Arnold Zwicky, editors, *The Handbook of Morphology*, pages 500–520. Blackwell, Oxford, UK & Cambridge, MA.

Pascale Piron. 1990. Éléments de description du kota, langue bantoue du gabon. mémoire de licence spéciale africaine, Université Libre de Bruxelles.

Brian Roark and Richard W. Sproat. 2007. *Computational approaches to morphology and syntax*. Number 4. Oxford University Press, USA.

Gregory T. Stump. 1992. On the theoretical status of position class restrictions on inflectional affixes. In G. Booij and J. van Marle, editors, *Yearbook of Morphology 1991*, pages 211–241. Kluwer.

Gregory T. Stump. 1998. Inflection. In A. Spencer and A. M. Zwicky, editors, *The Handbook of Morphology*, pages 13–43. Blackwell, Oxford & Malden, MA.

Gregory T. Stump. 2001. *Inflectional Morphology: a Theory of Paradigm Structure*, volume 93. Cambridge University Press.

A Corpus of Santome

Tjerk Hagemeijer, Iris Hendrickx, Haldane Amaro, Abigail Tiny
Centro de Linguística da Universidade de Lisboa
Av. Prof. Gama Pinto 2, 1649-003 Lisbon, Portugal
t.hagemeijer@clul.ul.pt; iris@clul.ul.pt; amaro25@hotmail.com; abigail.tiny@hotmail.com

Abstract

We present the process of constructing a corpus of spoken and written material for Santome, a Portuguese-related creole language spoken on the island of S. Tomé in the Gulf of Guinea (Africa). Since the language lacks an official status, we faced the typical difficulties, such as language variation, lack of standard spelling, lack of basic language instruments, and only a limited data set. The corpus comprises data from the second half of the 19th century until the present. For the corpus compilation we followed corpus linguistics standards and used UTF-8 character encoding and XML to encode meta information. We discuss how we normalized all material to one spelling, how we dealt with cases of language variation, and what type of meta data is used. We also present a POS-tag set developed for the Santome language that will be used to annotate the data with linguistic information.

Keywords: Santome, Creole, S. Tomé and Príncipe, Corpus, Standardization, Annotation

1. Introduction

Santome (São-Tomense, Forro) is a Portuguese-related creole language spoken on the island of S. Tomé in the Gulf of Guinea (Africa). The language has no official status, but according to the 2001 census 72,4% of the population over 5 years old spoke Santome (as a first or second language), on a total population of 137,599 (RPGH – 2001, 2003). For Portuguese, the official and most widely spoken language on the island, this percentage was 98,9%, which shows that there is a high degree of bilingualism. Nevertheless, Santomeans, in particular younger generations, have been shifting towards Portuguese. The result is a gradual loss of diglossia and language attrition of the creole.

Together with Ngola (Angolar), Fa d'Ambô (Annobonense) and Lung'Ie (Principense), which have much smaller numbers of native speakers, Santome is the direct descendant of the proto-creole of the Gulf of Guinea that came into being in the 16th century, on the island of S. Tomé, as the result of language contact between Portuguese, the lexifier language, and several Benue-Congo languages, in particular Edo (Edoid) and Kikongo (Bantu) (Ferraz, 1979; Hagemeijer, 2011). The first concise language studies with samples of Santome date back to the second half of the 19th century (e.g. Schuchardt, 1882; Negreiros, 1895). Since the monograph of Ferraz (1979), the number of studies on this creole has increased significantly. Yet, it still lacks basic language instruments, such as a reference grammar and a dictionary.

The electronic Santome corpus is being built as part of the project *The origins and development of creole societies in the Gulf of Guinea: An interdisciplinary study*. (cf. section 8). Within this project, the corpus under construction will be used primarily for linguistic purposes, in particular data extraction and comparison with the other three Gulf of Guinea creoles mentioned above, in order to reconstruct properties of the proto-creole of the Gulf of Guinea. In addition, the corpus has potential for tasks related to language planning in S. Tomé and Príncipe, such as the development of dictionaries and text materials. To our best knowledge, no electronic corpora of the type presented here have yet been built for Portuguese-related creole languages. Despite a few exceptions mentioned in the next paragraph, this also applies to creole languages with other lexifiers. In part, this may be related to the fact that many creole languages are small minority languages lacking an official orthography or even a writing tradition altogether. Consequently, corpus building can be costly and labor intensive: it requires fieldwork trips for data collection, transcribing, revising, standardizing, and so on.

We would like to mention a few corpora we found for other creole languages. The Corpus of Written British Creole (Sebba, Kedge & Dray 1999) counts around 12,000 words of written English-based Caribbean Creole. The corpus consists of samples from different text genres and is manually annotated with tags that signal lexical, discourse, structure, and grammatical differences between Standard English and the creole. The corpus is available for research purposes. A corpus of 200.000 words of Mauritian creole, a French-based creole language, is available online and searchable via a concordance interface as part of the website of the ALLEX project[1]. There is also a corpus of Tok Pisin (English-related with a Melanesian substrate) consisting of 1047 folktales that were translated to English and published in book form (Slone, 2001). Furthermore, the new digital era offers new

[1] ALLEX project http://www.edd.uio.no/allex/corpus/africanlang.html.

possibilities to gather corpus data for certain creole languages. An example is the COJEC corpus of Jamaican creole, a collection of emails and forum messages of about 40,000 words, written by Jamaican students (Hinrichs, 2006)[2].

2. Corpus

The Santome corpus consists of a compilation of oral and written sources. Since the second half of the 19th century, the language has been written in published and non-published sources, but as it is not an official language, not much material has been produced. The 19th century language samples consist of a few poems by Francisco Stockler (collected from different sources), and language fragments collected in Adolfo Coelho (1880-1886), Schuchardt (1882) and Negreiros (1895). The published sources further include a few newspaper articles from the 1920s and a small number of books and magazines written after the country's independence (1975). The books and cultural magazines typically intersperse Santome and Portuguese texts and some of the Santome texts come with a Portuguese translation. The unpublished sources comprise a number of pamphlets from the 1940s or early 1950s obtained from private sources and many unidentified texts (mostly song texts) collected in the Historical Archive of S. Tomé and Príncipe. Apart from a few known sources that we were unable to locate so far, we believe that we have gathered a significant amount of the existing written materials. Many texts that have been produced can be placed in the domain of folklore (folk tales, proverbs, riddles, etc.). The fact that the language does not reach into other functional spheres typically associated with prestige, such as journalism and education, is one of the reasons why the language is not thriving and may quickly become more severely endangered. Finally, it should be noted that the number of authors that produced the texts is relatively restricted. The author of the newspaper articles from the 1920s is also the one who wrote the pamphlets, many of the song texts were written by a small number of song writers and most of the proverbs were collected by a single author. We also encountered one source of the new media, a blog written in Santome. Blogs are an interesting text genre with an informal writing style and can be seen an online diary expressing the personal opinions of the blog's author. The written subcorpus has currently 99,658 words.

The spoken corpus comprises transcriptions of recordings of predominantly folk tales told by story tellers and conversations and songs that were recorded in 1997 and 2001 with native speakers of the language from different locations in S. Tomé. This subcorpus has currently 52 transcribed recordings of 20 different speakers who produced a total of 84,951 words. The spoken recordings have been freely transcribed in the sense that we have tried to match written text as much as possible. Many of the typical oral phenomena, like fragmented words, extra- linguistic sounds, hesitations, repetitions and linguistic repairs, were not transcribed because we aimed to keep the texts fluently. This type of free transcription can be seen as an additional normalization step of the spoken material. Additional details on the written and spoken subcorpus can be found in section 4.

Since we do not have copyrights for all the materials used in the corpus, we cannot make the corpus freely available at this point. We plan to remedy this problem by making the corpus available for concordances in an online interface, CQPweb (Hardie, forthc.)[3], that allows users to search for concordances of word forms, sequences of words and POS categories. The platform will also allow users to create frequency lists and to restrict the search query to specific text types.

3. Language standardization

Since Santome has no official status, the (Romance-based) orthographies have been highly variable and often quite inconsistent, ranging from etymological orthographies to phonological writing systems, a well known problem for creole languages in general (e.g. Sebba, 1996). A word like [kwa] 'thing', for instance, has been written in the following ways: *cua, cuá, qua, quá, kua, kuá, kwa, kwá*. Many texts show an unnecessary proliferation of accents and irregular morpheme separation of function words (aspect markers and negation markers, for instance) and lexical items. The explanation for the popularity of etymological orthographies, i.e Portuguese-oriented orthographies, can be assigned to the fact that probably over 90% of the creole's lexicon is drawn from Portuguese, the official language. However, Portuguese etymons underwent significant phonological changes when they were historically incorporated in the creole (Ferraz, 1979) and a considerable number of etymologies are unknown or traceable to the African source languages. This has led us to adopt ALUSTP, the 2009 phonology-oriented writing proposal that was ratified in 2010 by the Ministry of Education of Culture of S. Tomé and Príncipe (Pontífice *et al.*, 2009).

(Original)	**(Adapted)**
Inen piscadô nón	*Inen pixkadô non*
di tudu bóca plé	*di tudu boka ple*
di téla cé non glavi ximentxi	*di tela se non glavi ximentxi*
cá chê ni ké d'inen	*ka xê ni ke dinen*
cu amuelê cu buá vonté	*ku amwêlê ku bwa vonte*
chê bá nótxi	*xê ba nôtxi*
chê bá Tlachia	*xê ba Tlaxa*
basta p'inen bála blé d'omali	*baxta pa inen ba ala ble d'omali*
bá bucá vadô panhã cé	*ba buka vadô panha se*

Table 1. Excerpt from Quinta da Graça (1989) in the original and adapted version.

[2] The emails of COJEC are published in the appendix of the book.

[3] CQPweb: http://cqpweb.lancs.ac.uk/

The main principle of this proposal is a one-to-one phoneme-grapheme correspondence. We decided to standardize all material in the corpus to this spelling. The excerpt from a poem written by Quintas da Graça (1989) in Table 1 above illustrates the original writing system and the system used for the corpus.In the original text, [ʃ] is represented by the following graphemes: 's' (*piskadô*), 'x' (*ximentxi*) and 'ch' (*chê*). In the adapted version 'x' occurs in all the contexts. The sound [k] is represented by 'k' (*ké*) and 'c' (*bóca*) in the original text and becomes 'k' in the orthography we use. Santome exhibits a contrast between open-mid vowels ([ɛ], [ɔ]) and close-mid vowels ([e], [o]), which are respectively marked with acute and circumflex accents in the original version. In the adapted version, we maintain the circumflex accent for close-mid vowels and use no accent for open-mid vowels. In the case of vowel [a], accents are redundant altogether, because there is no contrasting pair. An example of morpheme separation follows from the form *p'inen* and *bála*, which become *pa inen* (lit. for they) and *ba ala* (lit. 'go there').

It follows from these examples that adapting all the different original orthographies represents a heavy workload. All the texts were scanned with OCR software or copied manually and then adapted to the proposed standard in a text editor. The original texts are all typewritten (the majority on typewriters) but sometimes in bad state of conservation. Instances of language variation (e.g. *djêlu* ~ *jêlu* 'money'; *idligu* ~ *igligu* 'smoke') were maintained as much as possible, in particular in the spoken corpus. With respect to variation, the written corpus is of course less reliable, because it is not always crystal-clear what variant underlies a given written form. By including variation, the corpus will also be useful to analyze quantitative and regional variation, which can then be used in language planning. The corpus and the variation found therein is also being used in a forthcoming Santome dictionary with over 4,000 lexical entries (Araújo & Hagemeijer, in preparation).

4. Meta data

The format of the corpus follows the general norms for corpus linguistics (e.g. Wynne, 2005) and uses UTF-8 character encoding and XML annotation for the meta data. We decided to encode the meta data about the corpus texts like author and date in a simple XML format that is compatible with the P5 guidelines of Text Encoding Initiative (TEI consortium, 2007). Next, a brief explanation of the XML meta data tags is provided.

- language: In addition to Santome, the project will build a corpus for the other three Gulf of Guinea creoles (cf. introduction). Note, however, that the corpora of these three languages will be much smaller and mostly restricted to spoken data due to the absence of a writing tradition
- corpus: Spoken or written
- title: The title of the text (if any)
- author: The author of the text (if known)
- age: the age of the recorded speaker (spoken data)
- place of recording: geographical location of the recording (spoken data)
- date: The date of publication (if any), which can be exact or approximate. Unless we found evidence to the contrary, we assumed that publication dates are close to the date of writing.
- source: We use the following list of sources: book, newspaper article, (cultural) magazines, pamphlets, online, unknown.
- genre: We use the following list of genres for the written corpus: prose, poetry, proverbs, riddles, song texts, mixed, other. For the spoken corpus, there are three genres, namely prose (folk tales and other stories), music and conversations.
- notes: Tag reserved for any type of additional information, such as the name of publisher and the place of publication.

While some of the tags speak for themselves, a few notes are in place here, particularly with respect to the typology of genre. In light of the predominantly folklore-related materials that were obtained, we did not follow text typology recommendations used for large corpora. Since the main goal of the corpus within the project concerns linguistic analysis, the different genres can serve different purposes. Most importantly, prose should be set apart from the other genres. The narratives in the corpus including folk tales and (personal) stories, as well as the blog, are the best means for investigating specific linguistic topics that require larger portions of text (e.g. clause-linking or anaphoric relations). In proverbs, riddles or poetry, on the other hand, one might find archaic lexicon or structures that are less likely to be found in prose. Another criterion underlying the classification in genres relates to the amount of data that was available for each genre. A more fine-grained division would have lead to genres with smaller amounts of material. In Table 2 we present how the number of files and words is divided over the genres.

written subcorpus		
genre	**files**	**words**
mixed	**10**	**22.652**
music (song texts)	**169**	**21.081**
poetry	**11**	**4.442**
prose	**59**	**40.364**
proverbs	**3**	**9.081**
other	**4**	**1.936**
subtotal	**257**	**99.658**
spoken subcorpus		
conversation	7	20.945
prose	43	62.844
music	2	802
subtotal	**52**	**84.591**
TOTAL	**309**	**184.249**

Table 2. Distribution of files and words across the different genres in the Santome corpus

The high number of files in the category "music" derives from the fact that we are dealing with unpublished sources, often a song text on a sheet of paper. Many of the proverbs, on the other hand, were published in a single volume (Daio, 2002). Finally, the "mixed" genre includes publications – in particular cultural magazines – with different types of texts that belong to one of the other five genres. In these cases the main header receives the label "mixed", but we applied subheaders in line with the TEI guidelines[4] to distinguish between genres in the text, for instance <div genre="music"> … </div>. This strategy was also adopted for other changes in the header data, for instance a change of authors within a collection of poetry. For the spoken part, we only have material of three different text genres. The largest part is made up of told stories (prose).

5. POS annotation

Once a few minor issues related to the uniformization of the data and the headers are settled, we plan to start the enrichment of the corpus with linguistic annotation, namely part-of-speech (POS) tagging. The following tag set has already been prepared based on a small subset of the data and on our knowledge of the language. It still needs testing on a larger data set. The tag set is based on the guidelines by Leech & Wilson (1996) and on the CINTIL tag set that was developed for Portuguese CINTIL corpus (Barreto *et al*., 2006). The adaptation of the grammatical categories was crucial, because Santome is typologically very different from Portuguese and shows greater resemblance to certain West-African languages, such as Edo, its main substrate language (Hagemeijer, 2011; Hagemeijer & Ogie, 2011) or languages from the Kwa cluster (e.g. Aboh, 2004).

Santome is a strongly isolating language without any inflectional morphology and only two productive derivational morphemes. Reduplication and compounding, however, are productive morphological strategies. For reduplicated categories we propose RED: followed by the label of the category that is being reduplicated. Numerals, for instance, can be fully or partially reduplicated (RED:NUM).

(1) *tlêxi-tlêxi* 'in groups of three'
(2) *tlê-tlêxi* 'all three'

In addition to more standard tags, we propose a number of tags that are highly language specific. Ideophones are a special word category consisting of modifiers with specific phonological properties that normally occur with a unique lexical item (nouns, verbs, adjectives).

(3) *kabêsa wôlôwôlô*	'foolish person' (lit. head+id.)
(4) *sola potopoto*	'cry intensely' (lit. cry+id.)
(5) *vlêmê bababa*	'intensely red' (lit. red+id)

Tag	**Category**	**Examples**
ADJ	Adjectives	*glavi* 'pretty', *vlêmê* 'red'
ADV	Adverbs	*oze* 'today', *yôxi* 'yes'
ART	Articles	*ũa* 'a(n)'
CJ	Conjunctions	*maji* 'but', *punda* 'because'
CN	Common Nouns	*mosu* 'boy', *ope* 'foot, leg'
COMP	Complementizers	*kuma* 'that'
DGT	Digits	*0, 1, 42, 12345, 67890*
DEM	Demonstratives	*se* 'this, that', *xi* 'that'
EXC	Exclamatives	*kê* 'what'
FOC	Focus markers	*so, soku*
FW	Foreign words	mostly Portuguese words
ID	Ideophones	*sũũũ* (*pya sũũũ* 'stare at', lit. look+ID)
INT	Interrogatives	*kuma* 'how', *andji* 'where'
ITJ	Interjection	*kaka!* (surprise)
MOD	Modality Markers	*sela* 'must'
NEG	Negation markers	*na, fa, fô*
NUM	Numerals	*dôsu* 'two', *tlêxi* 'three'
PP	Participles	*bixidu* 'dressed',vadu 'split'
PM	Presentational marker	*avia* 'there was'
PNM	Part of Name	*Zon* 'John'
PNT	Punctuation Marks	*., ?, (, …*
POSS	Possessives	*mu* 'my', *bô* 'your'
PREP	Prepositions	*antê* 'until', *ku* 'with'
PREP:NOM	Nominal prepositions	*basu* 'under(neath)', *wê* 'in front of'
PRS	Personals	*n* 'I', *ê* 's/he, it'
PRT	Particles	*an* (interrogative particle)
QNT	Quantifiers	*kada* 'every', *tudu* 'all'
RED:xx	Reduplicated Categories	*kume-kume* 'keep eating' (RED:V)
REFL	Reflexives	*mu, bô, dê, non, …*
RV	Residual Value	abbreviations, acronyms, etc.
SPV	Special Verbs	*loja* 'to encircle, around', *pê* 'to put, in'
STT	Social Titles	*sun* 'Mr.', *san* 'Mrs.'
TAM	Tense-Aspect-Mood markers	*ka, xka, tava, ta.*
V	Verbs	*fla* 'to speak', *mêsê* 'to want'

Table 3. POS-Tag set for the Santome corpus annotation.

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

The language further exhibits grammaticalized preverbal Tense-Mood-Aspect morphemes (TAM), a number of grammaticalized modality markers (MOD) that typically occur clause-initially and discourse-related particles (PRT). Some of these morphemes are illustrated in the following sentence.

(6) *Ola* *nansê* ***ka*** ***ska*** *nda* *ku* *amigu,*
when 2PL TAM TAM walk with friend

sela *nansê* *toma* *kwidadu* ***ê****.*
MOD 2PL take care PRT
‘When you are hanging out with friends, you must be careful.

Another domain that requires special attention is adpositions. Like many West-African languages (e.g. Kwa languages), the prepositional function in Santome can be expressed by prepositions (7), nouns (8) and (defective) verbs (9), for which we will respectively use the following tags: PREP, PREP:NOM and SPV. The tag SPV makes it possible to distinguish between the use of certain verbs as main verbs (V) and verbs in the second position in serial verb constructions.

(7) *Ê* *xê* ***ni*** *ke.*
‘S/he leave PREP.LOC house
‘S/he left home.’

(8) *Ê* *sa* ***wê*** *ke*
3SG be eye house
‘’S/he’s in front of the house

(9) *Ê* *saya* *kanwa* ***pê*** *ple.*
3SG pull canoe put beach
‘S/he pulled the canoe on the beach.’

For the POS annotation we will use an automatic POS-tagger trained on a manually annotated sample both to support the annotation process and to tag the rest of the corpus automatically. For the annotation, the tagger can speed up the POS tagging as manual verification is faster than annotating every word manually. We aim to have a sample of 50K words of the corpus manually verified and the other part will be tagged automatically.

6. Final remarks

We presented the construction and annotation of a corpus of Santome. The process of resource creation for a creole language like Santome has to deal with problematic issues like lexical language variety both in spoken and written material, a small body of written material with spelling variance, lack of standardized resources, such as a standard spelling, dictionary or grammatical reference. We aimed to address these issues by 1) collecting all written material that we could find into one uniformly encoded corpus 2) adding meta data information 3) standardizing the spelling of the written material to one systematic spelling 4) transcribing spoken material in the same spelling format 5) development of a POS-tagset for Santome. We expect the corpus to be a useful resource to establish the degree of relatedness between the four Gulf of Guinea creoles and a tool in language maintenance and revitalization, partly through the development of other language resources. Despite the fact that creole languages constitute different genetic units and not a single language family, it is often highlighted that they share certain linguistic (typological) properties. Therefore we believe that a more widespread corpus-based approach to these languages will endow comparative research on creoles with tools that allow for investigating these claims based on larger amounts of data.

7. Acknowledgements

The Santome corpus is funded by the Portuguese Foundation of Science and Technology (FCT) as part of the project *The origins and development of creole societies in the Gulf of Guinea: An interdisciplinary study* (PTDC/CLE-LIN/111494/2009) and the FCT program Ciência 2007/2008 (Iris Hendrickx).

8. References

Aboh, E. (2004). *The morphosyntax of complement-head sequences: Clause structure and word order patterns in Kwa*. Oxford: Oxford University Press.

Aráujo, G. & Hagemeijer, T. (in preparation). *Dicionário santome-português / português-santome*. São Paulo: Hedra.

Barreto F., Branco, A., Ferreira, E., Mendes, A., Bacelar do Nascimento, M. F. P., Nunes, F. and Silva, J. (2006). Open resources and tools for the shallow processing of Portuguese. In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC2006), Genoa, Italy.

Coelho, A. (1880-1886). Os dialectos românicos ou neo-latinos na África, Ásia e América. In Jorge Morais Barbosa (ed.) [1967], *Crioulos*. Lisboa: Academia Internacional de Cultura Portuguesa.

Daio, O. (2002). *Semplu*. S. Tomé: Edições Gesmédia.

Ferraz, L. (1979). *The creole of São Tomé.* Johannesburg: Witwatersrand University Press.

Hagemeijer, T. (2011). The Gulf of Guinea creoles: genetic and typological relations». *Journal of Pidgin and Creole Languages*, 26:1, pp. 111-154.

Hagemeijer, T. & Ogie, O. (2011). Edo influence on Santome: evidence from verb serialization and beyond. In Claire Lefebvre (ed.), *Creoles, their substrates, and language typology*. Amsterdam, Philadelphia: John Benjamins, pp. 37-60

Hardie, A (forthcoming) "CQPweb - combining power, flexibility and usability in a corpus analysis tool". Online available: http://www.lancs.ac.uk/staff/hardiea/cqpweb-paper.pdf

Hinrichs, L. (2006). *Codeswitching on the Web: English and Jamaican Creole in e-mail communication*. (Pragmatics and Beyond New Series 147). Amsterdam: John Benjamins.

Geoffrey Leech and Andrew Wilson (1996). EAGLES. Recommendations for the Morphosyntactic Annotation of Corpora. Technical Report. Expert Advisory Group on Language Engineering Standards. EAGLES Document EAG-TCWG-MAC/R.

Negreiros, A. (1895). *Historia Ethnographica da ilha de S. Tomé*. Lisbon.

Pontífice, J. *et al.* (2009). *Alfabeto Unificada para as Línguas Nativas de S. Tomé e Príncipe (ALUSTP)*. São Tomé.

Quintas da Graça, A. (1989). *Paga Ngunu*. S. Tomé: Empresa de Artes Gráficas.

RGPH – 2001. (2003). *Características educacionais da população – Instituto Nacional de Estatística*. S. Tomé e Príncipe.

Schuchardt, H. (1882). Ueber das Negerportugiesische von S. Thomé. *Sitzungsberichte Wien* 101. 889-917.

Sebba, M. (1996). Informal orthographies, informal ideologies spelling and code switching in British Creole. Cadernos de Linguagem e Sociedade, Vol. 2, No 1.

Sebba, M., Kedge, S.; Dray, S. (1999). The corpus of written British Creole: A user's guide. http://www.ling.lancs.ac.uk/staff/mark/cwbc/cwbcman.htm ((Date of access: Feb 27, 2012)

Slone, T.H. (2001). One Thousand One Papua New Guinean Nights: Folktales from Wantok Newspapers: Volume 1, Tales from 1972-1985 and Volume 2, Tales from 1986-1997 (Papua New Guinea Folklore Series) , Masalai Press, Oakland, California.

TEI Consortium (2007). TEI P5: Guidelines for Electronic Text Encoding and Interchange. www.tei-c.org/Guidelines/P5/ (Date of access: Feb 25, 2012).

Wynne, M. (2005). *Developing Linguistic Corpora*: a Guide to Good Practice. Oxford: Oxbow Books.

The Tagged Icelandic Corpus (MÍM)

Sigrún Helgadóttir[*], **Ásta Svavarsdóttir**[*], **Eiríkur Rögnvaldsson**[†],
Kristín Bjarnadóttir[*], **Hrafn Loftsson**[‡]

[*]The Árni Magnússon Institute for Icelandic Studies, [†]University of Iceland, [‡]Reykjavík University
Reykjavík, Iceland
sigruhel@hi.is, asta@hi.is, eirikur@hi.is, kristinb@hi.is. hrafn@ru.is

Abstract

In this paper, we describe the development of a morphosyntactically tagged corpus of Icelandic, the *MÍM* corpus. The corpus consists of about 25 million tokens of contemporary Icelandic texts collected from varied sources during the years 2006–2010. The corpus is intended for use in Language Technology projects and for linguistic research. We describe briefly other Icelandic corpora and how they differ from the *MÍM* corpus. We describe the text selection and collection for *MÍM*, both for written and spoken text, and how metadata was created. Furthermore, copyright issues are discussed and how permission clearance was obtained for texts from different sources. Text cleaning and annotation phases are also described. The corpus is available for search through a web interface and for download in TEI-conformant XML format. Examples are given of the use of the corpus and some spin-offs of the corpus project are described. We believe that the care with which we secured copyright clearance for the texts will make the corpus a valuable resource for Icelandic Language Technology projects. We hope that our work will inspire those wishing to develop similar resources for less-resourced languages.

Keywords: corpus, tagging, Icelandic

1. Introduction

This paper describes the Tagged Icelandic Corpus (the *MÍM* corpus) and how it was created. The project has been developed at The Árni Magnússon Institute for Icelandic Studies (AMI)[1]. The *MÍM* corpus is a synchronic corpus that will contain about 25 million running words. The texts are taken from different genres of contemporary Icelandic, i.e. texts produced in 2000–2010. All the texts have already been collected, part of the corpus has been tagged and is available for search (about 17.7 million tokens in October 2011).[2] The texts have already been used for various Language Technology (LT) projects. The *MÍM* corpus will be available in its entirety, both for search and download, in the summer of 2012.

Work on the corpus building started in 2004. It was one of the main projects of an LT Program launched by the Minister of Education, Science and Culture in 2000 (Rögnvaldsson et al., 2009). From the beginning, the *MÍM* corpus was mainly intended for use in LT, and the product of the work should be a balanced collection of contemporary texts, morphosyntactically tagged and lemmatised and supplied with metadata in TEI-conformant XML format (Burnard and Bauman, 2008). However, it soon became apparent that it would also be necessary to supply a web-based search interface to the corpus, for the benefit of researchers, teachers, students and lexicographers.

The paper is structured as follows. In Section 2., we describe briefly other Icelandic corpora. In Section 3., we give an account of the *MÍM* corpus and how it was created. The availability and use of the corpus is described in Section 4., and related projects are mentioned in Section 5. Finally, we conclude with a summary in Section 6.

2. Icelandic Corpora

At the turn of the century Icelandic LT virtually did not exist (Rögnvaldsson et al., 2009). In a report, written for the Minister of Education, Science and Culture in 1999 (Ólafsson et al., 1999), the lack of corpora for the development of LT tools is given a particular mention. The compilation of a balanced morphosyntactically tagged corpus of 25 million words was therefore one of the projects supported by the special LT Program launched in 2000.

However, a small corpus, annotated with morphosyntactic tags and lemmata, existed at the Institute of Lexicography (now a part of the AMI). This corpus had been compiled for the making of the Icelandic Frequency Dictionary (*IFD*), *Íslensk orðtíðnibók*, published in 1991 (Pind et al., 1991). The *IFD* corpus[3] consists of just over half a million running words, containing 100 fragments of texts, approximately 5,000 running words each. The corpus has a heavy literary bias as about 80% of the texts are fiction.

The tagset of the *IFD* is more or less based on the traditional Icelandic analysis of word classes and grammatical categories, with some exceptions where that classification has been rationalized. The underlying tagset contains about 700 tags, of which 639 tags actually appear in the corpus. The tags are character strings where each character has a particular function, denoting a (specific value of a) grammatical category. The tagging and lemmatisation of the *IFD* corpus was manually corrected and hence the corpus can be used as a gold standard for training part-of-speech (PoS) taggers.

Íslenskur orðasjóður[4] is an Icelandic corpus of more than 250 million running words collected from all domains ending in *.is* during the autumn of 2005, together with an auto-

[1] http://www.arnastofnun.is/

[2] http://mim.hi.is/

[3] http://www.malfong.is

[4] http://wortschatz.uni-leipzig.de/ws_ice/

matically generated monolingual lexicon, comprising frequency statistics, samples of usage, cooccurring words and a graphical representation of the word's semantic neighbourhood (Hallsteinsdóttir et al., 2007). The web texts were cleaned substantially before inclusion in the corpus. Since the corpus is neither balanced nor morphosyntactically tagged, its usefulness for certain types of linguistic research and LT projects is limited. Despite some limitations, this corpus is the only very large corpus of Icelandic in existence and it has proven to be useful in several projects. Of these, it is worth mentioning a project to create a Database of Semantic Relations (Nikulásdóttir and Whelpton, 2010), and projects to develop context sensitive spelling correction for Icelandic and the correction of OCR texts obtained from old print (ongoing unfinished projects).

The *Icelandic Parsed Historical Corpus* (*IcePaHC*)[5] is a diachronic treebank that was released in version 0.9 in August 2011 and contains about one million running words from every century between the 12^{th} and the 21^{st} centuries inclusive (Rögnvaldsson et al., 2011). The texts are annotated for phrase structure, PoS-tagged and lemmatised. The corpus is designed to serve both as an LT tool and a syntactic research tool. The corpus is completely free and open since most of the texts are no longer under copyright.

3. Creating the *MÍM* Corpus

In this section, we describe the creation of the *MÍM* corpus. We describe text collection, procedures for securing consent from copyright holders to use their material, text sources for written and spoken texts, methods for cleaning and annotation of the texts, and, finally, the creation of the metadata.

3.1. Text collection

Since the *MÍM* corpus is the first large balanced and tagged corpus with Icelandic text, one of the main criteria for its compilation was that it should contain a "balanced" or a "representative" text collection. However, researchers do not always agree on what is meant by these concepts. "Representativeness" has been defined as either "representing the population of texts or representing the structure of readership" (Przepiórkowski et al., 2010). Either of these criteria is difficult to establish. Following the population of texts would for instance mean that, for the period in question, most of the texts should have been sampled from the web. Following the structure of readership would require a survey of readership to be undertaken which was not practical at the time. A very pragmatic approach to the text collection was, however, adopted. An attempt was made to collect texts from different genres and from different sources. Only texts that were available in digital form were acquired. The texts were to have been written in the 21^{st} century, i.e. during the years 2000–2010, and be original writings in Icelandic. The texts were also to be morphosyntactically tagged and supplied with metadata.

In planning the text collection, the British National Corpus (*BNC*)[6] project (Aston and Burnard, 1998) was used as a model. However, with the advent of the Internet and the World Wide Web, the publishing scene has changed dramatically since the early nineties when the *BNC* was created. All the texts in the *BNC* corpus came from printed sources, apart from the spoken component. Since the budget of the *MÍM* project did not allow for the typing of text, the main restriction on the text collection was that the texts should be electronically available. Great care was taken in securing permission from copyright owners to use their text. The second restriction is thus that if a permission was not obtained for a particular text it was not included in the corpus.

Source	%
Printed newspapers	27.9
Printed books	22.3
Printed periodicals	8.7
Blog	7.6
Text from www.visindavefur.is	6.8
Text from government websites	6.4
Text from websites of organizations	6.2
Legal texts and adjudications	4.1
Texts written-to-be-spoken	2.9
School essays	2.6
Spoken language	2.2
Online newspapers and periodicals	1.5
Miscellany	0.8
Total	100.0

Table 1: Texts in *MÍM* by source

Table 1 shows the contribution of texts from the various text sources (media in *BNC* terminology) to the corpus material. Over one third of the texts were harvested directly from the World Wide Web. The spoken component, which comprises about 2.2% of the corpus texts, was made available by other projects.

3.2. Permissions clearance

Since the *MÍM* corpus was originally intended mainly for use in LT projects, it was considered of utmost importance to secure copyright clearance for the texts to be used. It was anticipated that most of the texts would be protected by copyright (final figure is about 88.5%). Early on in the project, cooperation was secured from the *Writer's Union of Iceland*[7], the *Association of Non-fiction and Educational Writers in Iceland*[8] and the *Icelandic Publishers' Association*[9]. All these associations recommended to their members that they should cooperate with the project. The most important of these, and the most difficult to secure, was the recommendation of the publishers' association, since publishers are normally the keepers of digital copies of published material.

Permission was sought from all owners of copyrighted texts included in the the *MÍM* corpus. Official texts (e.g. law, judicial texts, regulations and directives) are not copyrighted

[5] http://www.linguist.is/icelandic_treebank/

[6] http://www.natcorp.ox.ac.uk/

[7] http://rsi.is/

[8] http://hagthenkir.is/

[9] http://bokautgafa.is/

(11.5%). All copyright owners signed a special declaration and agreed that their material may be used free of licensing charges. In turn, AMI agrees that only 80% of each published text is included and that copies of the *MÍM* corpus are only made available under the terms of a standard license agreement. The crucial point in the license agreement is that the licensee can use his results freely, but may not publish in print or electronic form or exploit commercially any extracts from the corpus, other than those permitted under the fair dealings provision of copyright law. Data induced from the corpus, for example by a statistical PoS tagger, is considered results and may be used in commercial products. The license granted to the licensee is non-transferable.

With the help of a solicitor, legal documents were drawn up: A declaration for copyright holders to sign and a user license for prospective users of the corpus. Copyright holders were contacted either by e-mail or ordinary mail and received a copy of the declaration to sign, a copy of the user license, and a leaflet describing the *MÍM* project. Copyright holders were usually contacted twice. If there was no response after the second contact their text was discarded.

3.3. Written texts

It was decided that about 20–25% of the texts should be taken from **printed books**. Again, a very pragmatic approach had to be adopted. Publishers that were willing to cooperate were contacted. Books were selected from their catalogues and the authors contacted. If a positive answer was not obtained within a reasonable time limit another book was substituted and the procedure repeated. When the copyright owner had given his or her consent the publisher was contacted to obtain a digital copy of the book. It was soon found that the publishers only had digital copies available of books that had been published during the last few years. It was therefore not possible to include the texts of all books that permission was obtained for. Texts from books comprise about 22% of the corpus material and are taken from 117 books (47 novels, 12 biographies and memoirs and 58 books containing non-fiction).

The largest portion of text, about 28%, is taken from newspapers, mostly from **printed newspapers** (less than 1% from two **online newspapers**). The printed newspapers are *Morgunblaðið* (20%) and *Fréttablaðið* (8%). It is relatively easy to obtain permission to use text from newspapers since it is sufficient to get a signature from the editor. The texts from *Morgunblaðið* were obtained directly from their database, classified by content. The text was sampled so as to reflect seasonal variation in the topics under discussion. The text files from *Morgunblaðið* contained some metadata that could be removed automatically. The text from *Fréttablaðið* was obtained in PDF files. The text was extracted from the PDF files and had to be rearranged to a certain extent. The text from the two online newspapers was harvested directly from the web as clean text.

Text from **printed periodicals** (8.7%) was obtained from various sources. Most of the texts came from two publishers who each publishes a number of periodicals. Permission was obtained from the publishers and all the texts were delivered on a CD as either Word files or PDF files. A number of specialized periodicals were also sampled. They cover subjects like farming, aviation, immigrants, linguistics, medicine, natural sciences, computing, literature, history, fishing, education, and mathematics and sciences. Each editor had to be approached, and in some instances it was necessary to approach the author of each article in these periodicals. The texts were delivered as Word files, PDF files or harvested directly from the web.

Blog texts comprise about 7.6% of the corpus and they were harvested directly from the web. Each blogger was approached by e-mail and asked to consent to his text being used in the corpus. The blog texts in the corpus will be anonymous, only classified by type of writer, i.e. as texts written by politicians, theologians and what was called “general bloggers”.

The University of Iceland operates a website where the public can post questions on any subject (**www.visindavefur.is**). The answers are written by university academic staff and they cover most subjects taught at the university. The editors very kindly made answers from 38 writers available to the *MÍM* project and they also secured their permission to use the texts. The material comprises about 6.8% of the corpus texts and covers diverse subjects like meteorology, nursing, philosophy and anthropology.

About 11.5% of the texts in the corpus are official texts and therefore not covered by copyright. These are speeches from the Icelandic Parliament (Alþingi), (about 1% of the corpus texts, part of the texts **written-to-be-spoken** in Table 1), **legal texts and adjudications** (4.1%), and texts from the **websites of government ministries** (6.4%). All these texts, apart from the parliamentary speeches that were obtained from the database of Alþingi, were harvested directly from the respective websites.

Text was obtained from the **websites** of 14 **organizations** (6.2%). Permission was secured from the directors of these organizations and the text harvested directly from their websites. These websites represent diverse organizations like The Icelandic Road Administration, Save the Children in Iceland, and The Icelandic Tourist Board.

Texts classified in Table 1 as texts **written-to-be-spoken** comprise 2.9% of the corpus and are divided between the parliamentary speeches, radio and TV news scripts, and speeches harvested from various websites. These speeches are sermons delivered by ministers of the church in Reykjavík, addresses delivered at meetings, and radio scripts. The parliamentary speeches are not protected by copyright, but each of the other authors had to be contacted individually.

School essays (2.6%) are both essays from university students and papers written as a part of final examinations in Icelandic in a grammar school in Reykjavík. University students were contacted by e-mail, and they sent their essays back by e-mail, either as Word files or PDF files. The examination papers were obtained from the school office. Each student was contacted individually. Papers were not included in the corpus unless the writers had given their consent.

Only a small portion of the text was harvested from **online newspapers and periodicals** (1.5%). Permission was obtained from the editors.

In the category **miscellany** there are various small text excerpts, e.g. from teletext, leaflets, program notes from the Icelandic Symphony Orchestra, and text from electronic mailing lists.

3.4. Spoken texts

The budget of the project did not allow for extensive collection and transcription of spoken language. Through collaboration with other projects, it was, however, possible to secure some spoken language data. It consists of about 500,000 running words of transcribed text which is about 2.2% of the corpus. The spoken data was obtained through four different projects (Thráinsson et al., 2007) and it includes transcriptions of about 54 hours of natural speech, recorded in different settings in the period 2000–2006. The collection contains monologues, interviews and spontaneous conversations between adults of both sexes and with different backgrounds. The monologues are speeches from unprepared sessions in the Icelandic Parliament, recorded in 2004-2005. The interviews come from a sociolinguistic project and include several sessions, each with an interviewer and three interviewees. The conversations were recorded in informal settings, such as the homes or work places of one or more of the participants. 2–5 persons took part in each conversation. All the recordings have been carefully transcribed in a predefined format.

Permission was sought from each speaker to use the recordings anonymously for the purpose of language research. In the transcriptions, all names have been substituted by pseudonyms, and other personal data has been removed, since the permission is conditional upon not revealing personal information. The transcribed text from all the recordings will be made a part of the *MÍM* corpus. Moreover, the transcriptions aligned with the sound files will form a separate corpus which will be made searchable on a special website. This corpus will be protected by username and password. One part of the spoken language corpus, which contains transcribed recorded debates from the Icelandic Parliament, can be used more freely as restrictions regarding public data are not as strict as in the case of private dialogues.

3.5. Cleaning the text

As already mentioned in Section 3.3., texts obtained for the *MÍM* corpus came in various formats. The main formats were PDF files, Word files, XML files, text drawn from databases, and text harvested directly from the web. Texts from PDF files were extracted by a special program developed by a member of the *MÍM* team. As a last resort, we used optical character recognition software that is used for extracting text from scanned paper documents (ABBYY FineReader: `http://finereader.abbyy.com/`). Some texts came in Word documents which are easy to convert to text. The parliamentary speeches were delivered as XML files from a database at Alþingi. Text and metadata were extracted automatically with a program developed by a member of the *MÍM* team. Text from the *Morgunblaðið* database is easy to handle and contains metadata that can be extracted automatically and then removed before the text is morphosyntactically tagged and included in the corpus. Text harvested from the web is usually quite clean.

The importance of the cleaning phase should be emphasized. The quality of the text will influence later phases of the corpus building, i.e. sentence segmentation and tokenisation, which in turn influence the quality of the morphosyntactic tagging.

Texts from printed books and periodicals that are delivered either as PDF files or Word files usually contain hyphenation. Those texts were run through a program that joined the two parts of a word that had been split between lines. Various other measures had to be taken, either with automatic or semi-automatic means. We removed manually long quotations in a foreign language, long quotations from Old Icelandic texts and from new texts that we did not have permission to use, as well as footnotes, tables of content, indexes, reference lists, poems, tables and pictures. All texts were run through a cleaning program that standardizes quotation marks, both single and double, and hyphens.

The text files obtained for the corpus were either encoded using UTF-8 or ISO-8859-1 character encoding. It was decided that all texts in *MÍM* should be converted to UTF-8 encoding. However, in the tagging process (Section 3.6.) one of the taggers used requires text in ISO-8859-1 character encoding and the current version of the software used for searching the corpus also requires text in ISO-8859-1 character encoding. As a consequence all characters that are not a part of the ISO-8859-1 character set had to be substituted with simplified versions. Although long texts in foreign languages and Old Icelandic were removed there still remain names and short quotations. As an example of characters that had to be replaced the character *œ* was substituted with the character *ö* from the modern Icelandic alphabet and the the Greek character η was replaced with the character sequence *eta*.

3.6. Annotating the text

The annotation phase consists of sentence segmentation, tokenisation, morphosyntactic tagging and lemmatisation. After morphosyntactic tagging and lemmatisation, the texts, together with the relevant metadata, are transferred into TEI-conformant XML format with special programs developed by the *MÍM* team.

The procedure and software used for sentence segmentation, tokenisation, morphosyntactic tagging and lemmatisation has been explained by (Loftsson et al., 2010) in their work on the *GOLD* corpus (see Section 5.).

The tagset used was developed for the *IFD* corpus (see Section 2.). The automatic morphosyntactic tagging accuracy has been estimated as 88.1-95.1%, depending on text type (Loftsson et al., 2010).

3.7. Metadata

All texts in the corpus are accompanied by metadata. For published texts, the metadata comprises bibliographic data like title, name of author(s), age and gender of author(s), name of editor(s) (if applicable), publisher, date and place of publishing. For other texts, metadata is recorded to identify the text. For spoken data, various information on the recorded sessions and the speakers is registered. Most of the metadata had to be manually created, but metadata on files from the newspaper *Morgunblaðið* and on parliamentary speeches was created automatically. The metadata is shown for each text example retrieved through the search interface and is a part of the downloadable texts in TEI-conformant XML format. Individual texts can be selected for search through the search interface and also classified by source which reflects approximately the classification in Table 1. The texts will also be searchable by the target age group (adults, teenagers, children).

4. Availability and use of *MÍM*

4.1. Availability

As mentioned in Section 1., the corpus was originally made to be used in LT projects. However, it soon became obvious that a web-based search interface to the corpus was necessary to enable researchers, teachers, students and lexicographers to search the tagged corpus. The Norwegian search interface *Glossa* (Johannessen et al., 2008), which in turn uses the *IMS Corpus Workbench*[10] as a search engine, is being adapted to be used with the *MÍM* corpus.

An experimental search interface is already operating where about 17.7 million words of the corpus texts are available for search (see Section 1.). In the summer of 2012, all the corpus texts will be searchable. The corpus will also be available in TEI-conformant XML format in the summer of 2012, through download from a special webpage where prospective users register and agree to the licensing terms.

As a part of the project META-NORD[11], the Icelandic META-NORD team has established a special website (`http://www.malfong.is/`) where Icelandic Language Resources can be identified and located. Information on the *MÍM* corpus will be available there, as well as links to webpages for search and download of the corpus material.

Most of the published texts have been made accessible for search in their entirety (without annotation) in the Text collection of the AMI[12], where the outcome of the search is presented in KWIC format.

4.2. Uses of the corpus

The search interface is already being used in teaching Icelandic at the University of Iceland. The texts have been made available to the same projects as *Íslenskur orðasjóður* has been used for, as mentioned in Section 2.

The texts in the corpus are being used to augment the vocabulary in the *Database of Modern Icelandic Inflection* (*DMII*) (Bjarnadóttir, 2012). This database is available for search[13] and for download[14] for use in LT projects.

In the future, automatic lookup in the corpus will be possible, both from the Nordic *ISLEX*[15] database (Sigurðardóttir et al., 2008) and from the *DMII*. The user would then be given a chance to retrieve text examples from the corpus containing the word(s) he has looked up in the respective database. There is also a possibility of offering information on the frequency of particular word forms found in electronic databases based on the frequency in the *MÍM* corpus.

5. Related projects

The *MÍM* project has been carried out over a number of years. Various other projects have been worked on at the same time by the *MÍM* project group. Four will be mentioned here.

The first is a corpus of about 1 million running words which has been sampled from *MÍM*. This corpus which we call *GOLD* (Loftsson et al., 2010) is intended as a reliable standard for the development of LT tools. Tagging and lemmatisation of this subcorpus will be manually corrected.[16] This corpus will augment the *IFD* corpus (see Section 2.) which has been used for training statistical taggers and developing LT tools. The *GOLD* corpus is nearly twice the size of the *IFD* corpus and the texts are more varied. The *GOLD* corpus will be made available through the official site for Icelandic LT Resources, (`http://www.malfong.is`), for search, for download, and as training and test sets for the training of statistical taggers.

The second project is a separate corpus of about 500,000 words of spoken language, described in Section 3.4. This corpus is intended for theoretical and practical purposes relating to the spoken language.

The third is a project where about 1.7 million words of old Icelandic texts in normalized spelling have been tagged with morphosyntactic tags and lemmatised (Rögnvaldsson and Helgadóttir, 2011). Accuracy of the tagging was estimated as 92.7%. These texts are available (`http://www.malfong.is`) for search and download for use in linguistic research and LT projects.

The fourth is an experimental project, carried out in the summer of 2011, to add semantic analysis to the morphosyntactic tagging in the *MÍM* corpus, using the semantic analysis and classification of the vocabulary of *Íslenskt orðanet*[17] (Jónsson, 2010). *Íslenskt orðanet* is a database tracing semantic relations based on a large collection of word combinations. As a result, various links between lexical, grammatical and semantic features in the text examples of the corpus were established and users equipped with new and varied search choices.

6. Conclusion

As pointed out in the introduction, the *MÍM* corpus was built to serve two purposes. Firstly, it can be used in LT projects and, secondly, for language research. The part of the corpus open for search has already proved to be useful. The texts cannot be downloaded yet, but they have been made available to researchers, e.g. to a project where a Database of Semantic Relations is being created and in a project to develop context sensitive spelling correction for

[10] `http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/`

[11] `http://www.meta-nord.eu/`

[12] `http://www.arnastofnun.is/page/arnastofnun_gagnasafn_textasafn`

[13] `http://bin.arnastofnun.is/`

[14] `http://ordid.is/gogn/`

[15] `http://www.islex.hi.is/`

[16] As of February 2011 about 90% of the morphosyntactic tags have been manually corrected.

[17] `http://www.ordanet.is/`

Icelandic. Various spin-offs of the corpus project that will serve the LT community have been identified. The *MÍM* corpus is unique in the context of Icelandic LT, as it is the only large tagged corpus in Icelandic. Since permission for the use of texts in the corpus was secured from all copyright holders, and since researchers can obtain the texts and use them in LT projects despite some restrictions, the availability of the *MÍM* corpus will be better than is usually the case of corpora.

It is our wish that this work will inspire those wishing to develop a similar resource for less-resourced languages.

7. Acknowledgements

This project has been financed by the Language Technology Program of the Icelandic Ministry of Education, Science and Culture which ran during the years 2000–2004, the Icelandic Research fund through the project Viable Language Technology Beyond English (no 090662011), the University of Iceland Research Fund, the Student Innovation Fund and META-NORD (funded by the ICT PSP Programme, as part of the Competitiveness and Innovation Framework Programme under grant agreement no 270899). The authors would like to thank the many young students and researchers who have worked on various aspects of the project from time to time. Thanks are also due to the developers of *Glossa*, especially Anders Nøklestad. We would also like to thank the anonymous reviewers for helpful comments. Last, but not least, thanks are due to the numerous authors, writers, editors and others who have very kindly given permission for the use of their text and the publishers who made digital copies available. Without their cooperation this project could not have been completed.

8. References

G. Aston and L. Burnard. 1998. *The BNC handbook: exploring the British National Corpus with SARA*. Edinburgh University Press, Edinburgh.

K. Bjarnadóttir. 2012. The Database of Modern Icelandic Inflection. In *Proceedings of "Language Technology for Normalization of Less-Resourced Languages", workshop at the* 8^{th} *International Conference on Language Resources and Evaluation, LREC 2012*, Istanbul, Turkey.

L. Burnard and S. Bauman. 2008. Guidelines for Electronic Text Encoding and Interchange P5 edition. Text Encoding Initiative. `http://www.tei-c.org/Guidelines/P5/`.

E. Hallsteinsdóttir, T. Eckart, D. Biemann, and M. Richter. 2007. Íslenskur orðasjóður – Building a Large Icelandic Corpus. In *Proceedings of the* 16^{th} *Nordic Conference of Computational Linguistics (NoDaLiDa 2007)*, Tartu, Estonia.

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

J. H. Jónsson. 2010. Lemmatisation of Multi-word Lexical Units: Motivation and Benefits. In H. Bergenholtz, S. Nielsen, and S. Tarp, editors, *Lexicography at a Crossroads. Dictionaries and Encyclopedias Today, Lexicographical Tools Tomorrow*, pages 165–194. Bern: Peter Lang.

H. Loftsson, J. H. Yngvason, S. Helgadóttir, and E. Rögnvaldsson. 2010. Developing a PoS-tagged corpus using existing tools. In *Proceedings of "Creation and use of basic lexical resources for less-resourced languages", workshop at the* 7^{th} *International Conference on Language Resources and Evaluation, LREC 2010*, Valetta, Malta.

A. B. Nikulásdóttir and M. Whelpton. 2010. Extraction of Semantic Relations as a Basis for a Future Semantic Database for Icelandic. In *Proceedings of "Creation and use of basic lexical resources for less-resourced languages", workshop at the* 7^{th} *International Conference on Language Resources and Evaluation, LREC 2010*, Valetta, Malta.

R. Ólafsson, E. Rögnvaldsson, and Þ. Sigurðsson. 1999. Tungutækni [Language Technology]. Skýrsla starfshóps [Committee Report]. Menntamálaráðuneytið [Ministry of Education, Science and Culture]. Reykjavik, Iceland.

J. Pind, F. Magnússon, and S. Briem. 1991. *Íslensk orðtíðnibók [The Icelandic Frequency Dictionary]*. The Institute of Lexicography, University of Iceland, Reykjavik, Iceland.

A. Przepiórkowski, R. L Górski, M. Łaziński, and P. Pęzik. 2010. Recent Developments in the National Corpus of Polish. In *Proceedings of LREC 2010*, Valetta, Malta.

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

E. Rögnvaldsson, H. Loftsson, K. Bjarnadóttir, S. Helgadóttir, A. B. Nikulásdóttir, M. Whelpton, and A. K. Ingason. 2009. Icelandic Language Resources and Technology: Status and Prospects. In R. Domeij, K. Koskenniemi, S. Krauwer, B. Maegaard, E. Rögnvaldsson, and K. de Smedt, editors, *Proceedings of the NODALIDA 2009 Workshop Nordic Perspectives on the CLARIN Infrastructure of Language Resources*. Odense, Denmark.

E. Rögnvaldsson, A. K. Ingason, E. F. Sigurðsson, and J. Wallenberg. 2011. Creating a Dual-Purpose Treebank. *Journal for Language Technology and Computational Linguistics*, 26(2):141–152.

A. Sigurðardóttir, A. H. Hannesdóttir, H. Jansson, H. Jónsdóttir, L. Trap-Jensen, and Þ. Úlfarsdóttir. 2008. ISLEX – an Icelandic-Scandinavian Multilingual Online Dictionary. In *Proceedings of the XIII Euralex International Congress*, Barcelona.

H. Thráinsson, Á. Angantýsson, Á. Svavarsdóttir, T. Eythórsson, and J. G. Jónsson. 2007. The Icelandic (Pilot) Project in ScanDiaSyn. *Nordlyd*, 34(1):87–124.

Semi-automated extraction of morphological grammars for Nguni with special reference to Southern Ndebele

Laurette Pretorius, Sonja Bosch
University of South Africa
PO Box 392, UNISA, 0003, Pretoria, South Africa
E-mail: pretol@unisa.ac.za, boschse@unisa.ac.za

Abstract

A finite-state morphological grammar for Southern Ndebele, a seriously under-resourced language, has been semi-automatically obtained from a general Nguni morphological analyser, which was bootstrapped from a mature hand-written morphological analyser for Zulu. The results for Southern Ndebele morphological analysis, using the Nguni analyser, are surprisingly good, showing that the Nguni languages (Zulu, Xhosa, Swati and Southern Ndebele) display significant cross-linguistic similarities that can be exploited to accelerate documentation, resource-building and software development. The project embraces recognized best practices for the encoding of resources to ensure sustainability, access, and easy adaptability to future formats, lingware packages and development platforms.

1. Introduction

The normalisation and technological development of under-resourced languages not only involves the creation of basic resources, tools and technologies for processing these languages electronically, but also requires that such work be sustainable. Sustainability means that resources will remain accessible and available into the future even if the formats, software systems and development platforms on which they were developed become obsolete. To be sustainable they must even transcend communities of practice, domains of applications and the passage of time (Simons & Bird, 2008). Maxwell's (2012) approach to the sustainability of (morphological) grammars is of significance. He makes the point that grammatical descriptions must be reproducible (testable) and archivable, and shows that by making a grammar, including machine-processible rules, archivable, it also becomes reproducible. By using literate programming a descriptive and a formal grammar are interwoven. Furthermore, the formal grammar is a structured XML version of the descriptive grammar and is also parsing technology agnostic. In other words, it is archivable, (re)producible, human and machine-readable.

The problem that we address in this article is in some sense the converse one – we perform a semi-automated corpus-based extraction of a novel formal morphological grammar for Southern Ndebele from an existing morphological parser for Nguni. Our notion of an automatically extracted morphological grammar (as defined in section 4) is suitable for both human and machine processing. On the one hand, it employs a well-established grammar formalism that facilitates human readability and enables linguists to understand, evaluate and enhance the current descriptive status of the language. On the other hand, the machine processability makes it suitable as a resource from which new parsers can be (semi-)automatically constructed using other sustainable formalisms as well. In cases where groupings of under-resourced languages exist, developments towards sustainability of language resources for the group on the basis of collective information is of specific significance for the least resourced members of the group, both in terms of sustainable human-readable and machine-readable resources.

We examine the case of Southern Ndebele (ISO 639-3: `ndl`)[1], a resource-scarce language, for which we have semi-automatically obtained a morphological grammar from a general Nguni morphological analyser, which was bootstrapped from a mature hand-written Zulu morphological analyser called ZulMorph (Pretorius & Bosch, 2010). Due to the lack of Southern Ndebele language resources and data for inclusion in the bootstrapped Nguni analyser, Southern Ndebele morphological analysis relies heavily on the morphological structure of the other languages in the Nguni group.

The surprisingly good results obtained for Southern Ndebele morphological analysis suggest that a significant fragment of Southern Ndebele morphological grammar is covered by the Nguni analyser via the cross-linguistic similarities within the group. In this paper we describe a procedure for extracting this fragment and making it explicit with the purpose of creating a purely Southern Ndebele language resource that could be used in various ways, including the following: (a) to complement and extend the existing Southern Ndebele (paper-based) documentation towards enhanced Southern Ndebele language description; (b) to subject the fragment of the Southern Ndebele morphological grammar to human elicitation for quality assurance, correctness and enhancement purposes (the standard formal grammar notation is amenable to both human and machine processing); (c) to create a machine-readable formal morphological grammar that could serve as basis for the automated construction of morphological parsers in the future towards sustained Southern Ndebele language technological development.

The article is structured as follows: a background on Nguni languages; a brief explanation of the bootstrapping process with regard to Nguni languages; an illustration of the automated extraction of a morphological grammar for Southern Ndebele from the Nguni analyser; evaluation and results; and finally a conclusion and future work.

[1] Southern Ndebele should be differentiated from Northern Ndebele (ISO 639-3: (`nde`), spoken in Zimbabwe.

2. Nguni languages

The Nguni group of languages (S30 according to Guthrie's classification (Nurse & Philippson, 2003:649)) belongs to the South Eastern zone of the Bantu language family, and includes the following official languages of South Africa: Xhosa (S41), Zulu (S42), Swati (S43) and Southern Ndebele (S407).

The Nguni languages have a rich and complex morphology with a noun classification system that categorises nouns into a number of noun classes, as determined by prefixal morphemes also known as noun prefixes. Noun prefixes also link nouns to other words in the sentence by means of a system of concordial agreement, which is the pivotal constituent of the whole sentence structure of these languages, and governs grammatical correlation in most parts of speech.

2.1 Southern Ndebele

Southern Ndebele was the last South African Bantu language to receive official recognition, in fact during 1985 in the previous homeland of KwaNdebele which led to Southern Ndebele being introduced as official subject in schools for the first time. The language is spoken predominantly in the Mpumalanga and Gauteng provinces by a population of approximately 640,000.

Although Southern Ndebele has been an official language for 27 years, no comprehensive grammar, suitable for use by teachers and students, exists. Two master's dissertations on certain aspects of Southern Ndebele grammar were written by Potgieter (1950) (a description of the Ndzundza dialect) and Jiyane (1994). The first dictionary (isiNdebele/English, 2006) appeared more than 20 years into the official status of the language. Although an open-source spell checker for Southern Ndebele, based on word lists[2], a spellchecker and hyphenator[3], and corpora of 1.0 million untagged tokens[4] exist, no dictionary in electronic format, nor a detailed formally documented linguistic description was available for the bootstrapping process.

3. Nguni finite-state morphological analysis

Finite-state technology remains a preferred approach for modelling the morphology of natural languages. While machine learning approaches have grown in use, results for Nguni remain at the proof-of-concept level, due to among others morphological complexity and severe lack of appropriate language data (see for example Spiegler et al., 2008).

3.1 ZulMorph

The development of a broad-coverage finite-state morphological analyser prototype for Zulu (ZulMorph) is based on the Xerox Finite-state Tools (Beesley & Karttunen, 2003) and is reported on in detail in several publications, e.g. Pretorius and Bosch (2010). The Xerox software tool **lexc** is used to enumerate the required and essential natural-language lexicon and to model the morphotactic structure of Zulu words in this lexicon. Subsequently **lexc** source files are produced and compiled into a finite-state network which renders morphotactically well-formed, but rather abstract morphophonemic or lexical strings. The morphophonological (phonological and orthographical) alternations are modelled with rules written in the Xerox **xfst** format. Here the changes (orthographic/spelling) that take place between lexical and surface words when morphemes are combined to form new words/word forms, are described. These lexical strings are referred to again in section 4.3 where the extraction of the morphophonological rules is discussed.

Finally, the **lexc** and **xfst** finite-state networks are composed into a single network, namely a so-called lexical transducer that includes all the morphological information about the language being analysed, and constitutes the computational morphological analyser of the language, in this case the Zulu morphological analyser ZulMorph. It is customary to refer to the analysis language as the upper language of the transducer and to the surface form language as the lower language. Table 1 shows a summary of the core components of ZulMorph:

Morphotactics
Affixes for all parts-of-speech (e.g. SC, OC, CL PREF, V SUF, N SUF, TAM morphemes etc.) **Pronouns** (e.g. absolute, demonstrative, quantitative) **Demonstrative copulatives** **Word roots** (e.g. nouns, verbs, relatives, adjectives, ideophones, conjunctions) **Rules** for legal **combinations** and **orders** of morphemes (e.g. *ba-ya-si-khomb-is-a* and not **si-ba-ya-khomb-a-is*)
Morphophonological alternations
Rules that determine the **form** of each morpheme (e.g. *ku-hamb-w-a* > *ku-hanj-w-a*, *u-mu-lilo* > *u-m-lilo*)

Table 1: Core components of ZulMorph

3.2 A bootstrapped Nguni analyser

Due to the lack of language resources, bootstrapping of applications for new languages, based on existing applications for closely related languages, has gone a long way to reduce development time and efforts of building morphological analysers for lesser resourced languages – spoken by relatively few people – thereby ensuring technological development for such languages as well. Antonsen et al. (2010) report on the notable gain in reusing grammatical resources when porting language technology to new languages in the Uralic language family. The use of ZulMorph to bootstrap broad-coverage finite-state morphological analysers for Xhosa, Swati and

[2] http://translate.org.za/content/view/1610/54/

[3] http://www.nwu.ac.za/content/nwu-potchefstroom-campus-context-11

[4] http://web.up.ac.za/default.asp?ipkCategoryID=1883&sub=1&parentid=482&subid=1866&ipklookid=9

Southern Ndebele is discussed extensively in Bosch et al. (2008). The results of a preliminary evaluation based on parallel test corpora of approximately 7,000 types each for the four languages, indicate that the "high degree of shared typological properties and formal similarities among the Nguni varieties warrants a modular bootstrapping approach" (Bosch et al. 2008:66). The bootstrapping process is done in various stages by reusing the core components of the Zulu analyser for the three additional Nguni languages, shown in Table 1.

The bootstrapping approach functions as semi-automatic support to human linguistic expertise that allows linguists to focus their attention on just those aspects in which the languages differ. Adapting ZulMorph to provide for affix variations in the related languages, e.g. the form of morphemes in the 'closed' classes, proved to be a trivial implementation matter. However, certain areas in the grammars of individual languages that differ substantially from those applicable to Zulu required custom modelling and were built into the analyser as additional components e.g. the copula construction and the formation of the extended noun stem of Southern Ndebele. In the latter case the noun stem may suffix morphemes signifying the diminutive, augmentative etc. In the Southern Ndebele corpus two Southern Ndebele specific constructions occur which need to be included in the morpheme sequencing: a noun stem may suffix a demonstrative pronoun `[Dem7][Pos1]` or a possessive concord followed by a pronominal stem `[PossConc3] [PronStem7]`, e.g.

isilwanesi ('this animal')

`i[NPrePre7]si[BPre7]lwana.7-8[NStem-NR]`[5]

`lesi[Dem7][Pos1]`

umsilaso ('its tail')

`u[NPrePre3]mu[BPre3]sila.3-4[NStem]`

`wa[PossConc3]so[PronStem7]`

The Nguni lexical transducer has approximately 270 000 states and 833 000 transitions and occupies 14.3 MB of memory.

Simon and Bird (2008) identify six necessary and sufficient conditions for the sustained use of such resources. In particular, a language resource must be extant, discoverable, available, interpretable, portable, and relevant. The sustainability characteristics for ZulMorph and the Nguni analyser are given in Table 2.

Sustainability characteristics of ZulMorph and Nguni analyser
Extant Yes: Xerox finite-state tools implementations; appropriately backed-up off-site; mature prototypes in an advanced state of completion.
Discoverable Not yet: has not been released yet.
Available Limited: data analysis done on request, e.g. for National Centre for HLT, South Africa[6] .
Interpretable Yes: strictly based on the finite-state formalism and tools as described in (Beesley and Karttunen, 2003); adheres to relevant encoding standards; appropriately documented.
Portable, best practices Yes: shown to be compatible with equivalent open source initiatives such as foma (Hilden, 2009) and HFST (Lindén et al., 2011). Finite-state computational morphology is well established and can be expected to survive into the future. Finite-state research agendas already make provision for certain known limitations (Wintner, 2007).
Relevant Yes: constitute essential enabling technologies for next stages in the natural language processing pipeline of the agglutinating morphologically complex Nguni languages.

Table 2: Sustainability characteristics of ZulMorph and the Nguni analyser

4. Extraction of Southern Ndebele morphological grammar

The morphological grammar for Southern Ndebele consists of two main components, viz. rules that govern the morphotactics and rules that model the morphophonological alternations of Southern Ndebele. The morphotactics component (section 4.2) is a set of rules of the form $N \rightarrow N+$ or $N \rightarrow \Sigma$ where + is the Kleene plus operator, N is the (finite) set of morphological labels/tags and Σ is the (finite) set of actual morphemes. A distinguished symbol S ε N is the start symbol. The set of ordered morphophonological alternation rules are encoded by means of **xfst** (conditional) replacement rules. The rule `A -> B || L _ R` means that any string in language `A` is replaced by any string in language `B` only if the left context of the string in `A` is in the language `L` and the right context is in the language `R`. `A`, `B`, `L` and `R` are regular languages. These are addressed in section 4.3

The extraction (reverse engineering) of a morphological grammar for Zulu would be based on the full finite-state description of the complete Zulu morphology, as implemented in ZulMorph by means of **lexc** and **xfst**. This approach is essentially different from the extraction of a Southern Ndebele morphological grammar, which is corpus-based since the Nguni analyser does not contain much by way of explicit Southern Ndebele information. The grounding of the morphological grammar in Southern Ndebele therefore takes place via appropriate attested corpora. This approach results in a partial morphological grammar from the Nguni morphological analysis of the words in the corpus. Future work will focus on bigger corpora in order to increase the coverage of the Southern Ndebele morphological grammar.

4.1 General corpus-based approach

For the purposes of demonstrating the validity of the

[5] The notation NR indicates a Southern Ndebele specific morpheme.

[6] http://www.dac.gov.za/newsletter/khariambe_3_4.html

approach, a small representative Southern Ndebele corpus was used. After standard pre-processing and tokenisation were performed, the word list of 180 types (unique words) was subjected to morphological analysis with the Nguni analyser. In order to further constrain the proof-of-concept to tractable scope for the purpose of this article, only analyses that are based on noun stems were considered. Typical analyses that form the basis of the extraction process explained in sections 4.2 and 4.3 are as follows:

wesilwana ('of the animal')

```
wa[PossConc1]i[NPrePre7]si[BPre7]lwana.7-8
[NStem-NR]
```

wabhudanga ('he dreamt')

```
wa[PTSC1]bhudang[VRoot-NR]a[VerbTerm]
```

The coverage of the morphological grammar (both morphological structure and word root lexicons) can be increased by (a) systematically including other parts of speech, for example verbs, and (b) using larger corpora.

4.2 Rules for morpheme sequencing and the lexicon

The morphological grammar rules are automatically extracted from the morphological analyses by means of a pattern-matching procedure. As is customary for finite-state approaches all possible analyses for any given form are produced. For the purposes of obtaining the morphological grammar no (context dependent) disambiguation is necessary since all analyses are assumed to be valid and are therefore relevant for the extraction of morphological grammar rules.

In general, the following main parts of speech are recognised in the Nguni languages: noun, pronoun, demonstrative, qualificative, verb, copulative, adverb, ideophone, interjection, conjunction and interrogative (cf. Poulos and Msimang, 1998:26). In the corpus, we focus on a selection of these parts of speech, which contain the tag `NStem`, and we discuss examples of morpheme sequencing rules obtained from the analyses of these words.

Noun

The noun in the Nguni languages is constructed of two main parts, namely a noun prefix and a noun stem with the annotation `[NStem]` in ZulMorph. The noun prefix is the carrier of class information[7] and is usually divided into a so-called preprefix and a basic prefix. The noun stem may suffix morphemes signifying the diminutive, augmentative etc. In the Ndebele corpus two Ndebele specific constructions occur which required an enhancement of the morpheme sequencing: a noun stem may suffix a demonstrative pronoun or a possessive concord followed by a pronominal stem, e.g.

```
Noun -> NPrePre BPre NStem AugSuf
Noun -> NPrePre BPre NStem-NR PossConc PronStem
Noun -> NPrePre BPre NStem-NR Dem
```

Copulative

The copulative is a non-verbal predicate in Nguni and can be formed with a variety of words or stems, e.g. nouns, pronouns, adverbial forms etc. The following examples demonstrate the morpheme sequencing in copulatives formed from noun stems:

```
Copulative -> CopPre BPre NStem
Copulative -> SubjSC PreLoc-s LocPre (NPrePre)
BPre NStem
```

Qualificative

The qualificative part of speech is a collective term that covers different types of qualifying or descriptive words such as the adjective, relative, possessive and enumerative, as illustrated below:

```
Qualificative -> PossConc NPrePre NStem DimSuf
Qualificative -> RelConc AdvPre NPrePre BPre NStem
```

Adverb

The adverb in Nguni languages is quite a mixed bag, involving numerous types of grammatical constructions, mainly derived from other parts of speech such as nouns, pronouns, demonstratives, qualificatives etc. Since we focus on words containing noun stems in this paper, the adverbs under discussion are formed by prefixes and/or suffixes added to a noun. Constructions include adverbs formed by using prefixes *nga-* (instrumental), *na-* (associative) etc.; the locative prefixes *ku-*, *e-*; and prefix *e-* in combination with the locative suffix *–ini.* , e.g.

```
Adverb -> AdvPre NPrePre BPre NStem DimSuf
Adverb -> LocPre NPrePre BPre NStem DimSuf
Adverb -> LocPre NPrePre BPre NStem LocSuf
```

The latter exemplifies a long distance dependency between a locative prefix and a locative suffix. In particular, it is a circumfix or a co-ordinated pair consisting of a locative prefix which requires a locative suffix. This dependency cannot be completely captured by grammar rules since it requires idiosyncratic information about the specific noun stem.

4.3 Rules for morphophonologial alternations

Morphophonological alternations are the rules that determine the form of each morpheme. The rules for morphophonological alternations extracted from words based on noun stems in the Southern Ndebele corpus, are discussed by means of examples. In each case the morphemes that have undergone change are underlined in the surface word; then the lexical form is given, followed by the alternation rule in human readable form and an **xfst** representation.

We briefly explain the use of the lexical forms in identifying the specific rules that that were applicable to Southern Ndebele. Using the **lexc** morphotactics finite-state network on its own yields lexical forms as lower language strings.

These lexical forms contain special multicharacter symbols that are used in the **xfst** finite-state network to ensure that the rules fire correctly. We mention only a few. `^BR` and `^ER` denote the beginning and end of a word root.

[7] For conciseness of the grammar the class information is removed, since the focus is on the morpheme sequencing.

This is necessary for preserving the word root and to manage alternations that take place at word root boundaries; `^U`, `^MU`, `^I`, `^N`, `^SI`, etc. are placeholders for the noun prefixes to ensure that rules that apply only to such prefixes do not fire in cases where `u`, `mu`, `i`, `n`, `si`, etc. appear as other morphemes or parts of morphemes. They are finally removed by means of auxiliary rules. The `%` symbol is used in **xfst** to literalise special **xfst** symbols.

The use of the lexical form is illustrated by means of an example. In the first example it denotes the morpheme sequence `eisirhodloini.` When compared to the surface form *esirhodlweni* it is clear that the rules that fired must have been those that replace `ei` with *e* and `oini` with *weni*. By inspection the **xfst** rules

```
define VowelCombs  a e -> e , a i -> e ,
a o -> o , a u -> o , e a -> e , e i -> e ,
e u -> e , u a -> a , u o -> o;
```

and

```
define oiniRule o %^ER i n i -> w e n i;
```

may be identified as relevant and appropriate for Southern Ndebele.

Consonantalisation

esirhodlweni (‘in the court yard’)

`e^LP^I^SI^BRrhodlo^ERini`

Rule: o + ini > weni

```
define oiniRule o %^ER i n i -> w e n i;
```

Vowel coalescence

nenja (‘and the dog’)

`na^I^N^BRja^ER`

Rule: a + i > e

esinombala (‘that has the colour’)

`esina^U^MU^BRbala^ND^ER`

Rule: a + u > o

See the `VowelCombs` rule above.

Vowel elision

emthini

`e^LP^U^MU^BRthi^ERini`

Rules: e + u > e (where e is a locative prefix); mu > m (where mu is followed by more than one syllable).

```
define muRule
%^MU -> [m | 0 ] || _ %^BR m
.o. %^MU -> m || _ [%^BR Syllable Syllable %^ER
| %^BR Syllable %^ER [Vowel | Syllable] | %^BR
Syllable Syllable]
.o. %^MU -> m || _ %^BR Vowel
.o. %^MU -> m u;
```

Palatalisation

emlonyeni

`e^LP^U^MU^BRlomo^ERini`

Rule: mo + ini > nyeni

```
define locRule m o %^ER i n i -> n y e n i
```

5. Results and discussion

We obtained the morphological grammar rules (in which the non-terminal symbols are self-explanatory labels/tags) and morphemes (terminal symbols) that are applicable to the words in the corpus. The morpheme sequencing rules below have been condensed somewhat, but still reflect the automatic extraction. These rules should still be subjected to human elicitation. The | is the union operator, (and) denote optionality and [and] are used to delimit the scope of the union operator.

S → Adverb|Copulative|Noun|Qualificative
Adverb → AdvPre NPrePre BPre NStem (DimSuf)
Adverb → AdvPre NPrePre NStem (DimSuf)
Adverb → NegPre PTSC AdvPre NPrePre BPre NStem
Adverb → [PTSC|SC|SitSC|SubjSC] AdvPre NPrePre BPre NStem
Adverb → LocPre NPrePre BPre NStem (DimSuf)
Adverb → LocPre NPrePre BPre NStem LocSuf
Adverb → LocPre NPrePre NStem DimSuf
Copulative → SC PreLoc-s LocPre (NPrePre) BPre NStem
Copulative → SubjSC PreLoc-s LocPre (NPrePre) BPre NStem
Copulative → CopPre NPrePre NStem (DimSuf)
Copulative → NegPre PTSC CopPre BPre NStem DimSuf
Copulative → NegPre SC CopPre BPre NStem (DimSuf)
Copulative → ([PTSC|SC|SitSC|SubjSC]) CopPre BPre NStem (DimSuf)
Copulative → [PTSC|SC|SubjSC] CopPre NPrePre BPre NStem
Noun → (NPrePre) (BPre) NStem (DimSuf)
Noun → NPrePre BPre NStem ([AugSuf|DimSuf])
Qualificative → NPrePre BPre NStem PossConc PronStem
Qualificative → PossConc NPrePre BPre NStem (DimSuf)
Qualificative → PossConc NPrePre NStem DimSuf
Qualificative → RelConc AdvPre NPrePre BPre NStem
Qualificative → RelConc CopPre BPre NStem (DimSuf)
Qualificative → RelConcPT AdvPre NPrePre BPre NStem
Qualificative → RelConcPT CopPre BPre NStem DimSuf

The *N → Σ* grammar rules are summarised in Tables 3-5. For example, *LocSuf → ini* is the last row in Table 4.

Cl	**N-Pre-Pre**	**Bpre**	**Poss-Conc**	**Rel-Conc/ PT**	**SC/ Subj SC/ SitSC**	**PT SC**	**Pron-Stem**
1	*u*	*mu*	*wa*	*o*	*u/a/e*		
2	*a*	*ba*	*ba*				
1a	*u*						
2a							
3	*u*		*wa*	*o*	*u*		
4				*e*	*i*	*ya*	
5	*i*	*li*	*la*	*eli*	*li*		
6	*a*	*ma/me*	*a*	*a*	*a/e*	*a*	
7	*i*	*si*		*esi*	*si*		*so*
8			*za*		*zi*		
9	*i*	*n*		*e*	*i*	*ya*	
10	*i*	*zin*	*za*		*zi*		
14	*u*	*bu*	*ba*				
15			*kwa*				
1pp				*esi*	*si*		
2ps				*o*	*u*		

Table 3: Prefixes that depend on class, number and person

Prefix	**Morpheme**
AdvPre	*na, nga*
CopPre	*ngu, wu, bu, ku, li, si, zi*
LocPre	*e, ku, o*

NegPre	*a*
PreLoc-s	*s*
AugSuf	*kazi*
DimSuf	*ana*
LocSuf	*ini*

Table 4: Other affixes

Zulu	Xhosa	Southern Ndebele
bala.3-4	*cabanga.11-10*	*bhudango.5-6*
dlebe.9-10	*hle.11-10*	*bizo.5-6*
khathi.7-8	*hlolo.1-2*	*dlebe.9-10*
lomo.3-4	*hlolokazi.1-2*	*kukurumbu.9-10*
suku.10-11	*nto.9-10*	*pungutja.5-6*
thongo.14	*phapha.5-6*	*rhodlo.7-8*
vila.5-6	*qadi.5-6*	*tjhada.5-6*

Table 5: An extract of noun stems with class information

The alternation rules are manually extracted from the 180 rule Nguni **xfst** script, as explained in section 4.3.

Observations

The experiment based on 180 words, focussed on noun stems, already covers a wide spectrum of the Southern Ndebele morphological grammar. Moreover, increasing the corpus will improve the rules (morphology), the word root lexicons and the affixes for all parts of speech.

Human elicitation was responsible for the removal of Class 11 concordial elements in Table 3 since this noun class does not feature in Southern Ndebele. The adverb in Southern Ndebele is not described by Jiyane (1994), therefore the adverbial prefixes and locative prefixes identified from the corpus and listed in Table 4, also call for human elicitation.

In Table 5, noun stem cross-linguistic similarities are illustrated. A total of 88 possible noun stems are identified in the analysis of the small representative Southern Ndebele corpus. Of these 80.6% (71 noun stems) are Zulu and (11.4%) 10 noun stems are Xhosa. The 7 noun stems (8%) listed under Southern Ndebele, are noun stems with relevant class information that are not shared with either Zulu or Xhosa. Here too, human elicitation will confirm the appropriateness of the Zulu and Xhosa noun stems together with their class information, in a Southern Ndebele context.

6. Conclusion and future work

The proof-of-concept corpus-based morphological grammar extraction procedure yielded a novel prototype language resource for Southern Ndebele. The procedure scales well. This resource is human-readable, adds to the description of the language and is also machine-readable, allowing parser development, and supports sustainability. Future work includes the extension of the grammar extraction procedure to all parts of speech; the application to larger corpora; a comprehensive evaluation of the approach; the extension of the proof-of-concept to a possible evaluation procedure for existing morphological parsers for the other Nguni languages; and the representation of the extracted formal grammar in XML as a *de facto* standard for sustainability.

7. Acknowledgement

The constructive comments of Ken Beesley and two anonymous reviewers are appreciated.

8. References

Antonsen, L., Trosterud, T., Wiechetek, L. (2010). Reusing grammatical resources for new languages. In *Proceedings of LREC 2010*, pp. 2782—2789.

Beesley, K.R., Karttunen, L. (2003). *Finite state morphology*. Stanford, CA: CSLI Publications.

Bosch, S., Pretorius, L., Fleisch, A. (2008). Experimental Bootstrapping of Morphological Analysers for Nguni Languages. *Nordic Journal of African Studies*, 17(2), pp. 66--88.

Hulden, M. (2009). Foma: a finite-state compiler and library. In *Proceedings of the EACL 2009 Demonstrations Session*, pp. 29--32.

isiNdebele/English Dictionary. (2006). Johannesburg: Phumelela Books.

Jiyane, D.M. (1994). Aspects of isiNdebele grammar. MA Dissertation. University of Pretoria.

Lindén, K., Silfverberg, M., Axelson, E., Hardwick, S., Pirinen, T.A. (2011). HFST - Framework for compiling and applying morphologies in systems and frameworks for computational morphology. In *Communications in Computer and Information Science*, (100).

Maxwell, M. (2012). Electronic grammar and reproducible research. *Language Documentation and Conservation.* Preprint. To appear.

Nurse, D., Philippson, G. (2003). The Bantu languages. London: Routledge.

Poulos, G., Msimang, T. (1996). *A linguistic analysis of Zulu.* Pretoria: Via Afrika Limited.

Potgieter, E.F. (1950). Inleiding tot die klank- en vormleer van isiNdzundza, 'n dialek van Suid-Transvaalse Ngoeni-Ndebele, soos gepraat in die distrikte Rayton en Pretoria. MA Dissertation. University of South Africa.

Pretorius, L., Bosch, S.E. (2010). Finite-state morphology of the Nguni language cluster: modelling and implementation Issues. In A. Yli-Jyrä, Kornai, A., Sakarovitch, J. & Watson, B. (Eds.), *Finite-State Methods and Natural Language Processing 8th International Workshop, FSMNLP 2009. Lecture Notes in Computer Science*, Vol. 6062, pp. 123--130.

Simons, G.F., Bird, S. (2008). Toward a global infrastructure for the sustainability of language resources. In *Proceedings of the 22nd Pacific Asia Conference on Language, Information and Computation.*

Spiegler, S., Gol´enia, B., Shalonova, K., Flach, P., Tucker, R. (2008). Learning the morphology of Zulu with different degrees of supervision. In *Spoken Language Technology Workshop*, SLT. IEEE, pp. 9--12.

Wintner, S. (2007). Strengths and weaknesses of finite-state technology: a case study in morphological grammar development. *Natural Language Engineering*, 14(4), pp. 457--469.

Tagging and Verifying an Amharic News Corpus

Björn Gambäck

Norwegian University of Science and Technology
Trondheim, Norway
gamback@idi.ntnu.no

Abstract

The paper describes work on verifying, correcting and retagging a corpus of Amharic news texts. A total of 8715 Amharic news articles had previously been collected from a web site, and part of the corpus (1065 articles; 210,000 words) then morphologically analysed and manually part-of-speech tagged. The tagged corpus has been used as the basis for testing the application to Amharic of machine learning techniques and tools developed for other languages. This process made it possible to spot several errors and inconsistencies in the corpus which has been iteratively refined, cleaned, normalised, split into folds, and partially re-tagged by both automatic and manual means.

1. Introduction

There is a major shortage of language processing tools and resources for (almost all) African languages. This paper focuses on Amharic, the primary language of Ethiopia, and on the correction and processing of a tagged Amharic text corpus, taking as the starting-point a set of Amharic news articles collected at Stockholm University from an Ethiopian web news archive, and then morphologically analysed and manually part-of-speech tagged at Addis Ababa University.

Amharic is the second largest Semitic language: speakers of Arabic count in hundreds of millions, of Amharic in tens of millions, and of Hebrew and Tigrinya in millions. Ethiopia is divided into nine regions, each having its own official language, but Amharic is used as the *lingua franca* at the national level, so the number of second language speakers is fairly high. Giving a figure for the size of the Amharic speaker body is not easy, but based on the latest Ethiopian census carried out in 2007 (CSA, 2010), estimates of the current population of Ethiopia (CIA, 2012), and approximations of the percentage of Ethiopians speaking Amharic given the previous census in 1994 (Hudson, 1999), it is reasonable to assume that about 30 million persons speak it as first language and more than 10 million as second language.

Amharic uses a unique script (shared with Tigrinya), which in contrast to Arabic and Hebrew is written from left to right. The script is commonly known as “Ethiopic script”, “Ge’ez” or *fidel* (lit. “alphabet” in Amharic) and is basically syllabic with most of its 275 characters representing consonant-vowel combinations. The language is quite diversified both when spoken and written, and has, for example, no standard spelling for compounds and loan-words. It has a complex morphology, where nouns (and adjectives) are inflected for gender, number, definiteness, and case. Definite markers and conjunctions are suffixed to the nouns, while prepositions are prefixed. Like other Semitic languages, the verbal morphology is rich and based on triconsonantal roots.

Despite the large number of speakers, there have been few efforts to create language processing tools for Amharic. A deterrent to progress was lack of standardisation: an international standard for Ethiopic script was agreed on only in 1998 and incorporated into Unicode in 2000. Several representation formats for the script were used before that, thwarting language processing and electronic publication in Amharic. As an effect, almost all work on Amharic language processing has taken place after the millenium shift.

Another major deterrent to progress in Amharic language processing has been the lack of large-scale resources such as corpora and tools. Thus, for example, the best result reported by an Amharic part-of-speech tagger before the availability of the corpus discussed in this paper was by Adafre (2005). That work suffered from only having access to a 1,000 word training corpus, resulting in a word error rate of over 25%. As we shall see (in Section 5.), that can be improved to figures below 10% using a 200k word corpus; a number which still is high though, when compared to better-resourced language, for which WER of 2–4% is common.

Language processing for Amharic has in fact taken two major steps forward in recent years, both through the creation of a reasonably-sized tagged corpus and through the appearance of HornMorpho (Gasser, 2009), the most complete morphological processing tool for Amharic (and Tigrinya) to date. HornMorpho uses a finite-state approach to allow for both analysis and generation of nouns (and adjectives which are regarded as nouns) and verbs. Gasser (2011) attempts a small evaluation of HornMorpho’s performance on 200 randomly selected words each for the two “wordclasses”, reporting that about 95.5% of the noun/adjectives, and 99% of the verbs received an analysis (i.e., all legal combinations of roots and grammar structures for them could be found).

The Amharic news corpus in the present paper had previously been extracted from the web and manually part-of-speech tagged, as described in Section 2. The core of the paper is the normalisation and clean-up measures that had to be performed after the manual tagging (Section 3.) and the re-tagging and splitting of the corpus into folds (Section 4.). The corpus has been tested by the application to Amharic of several machine learning techniques for part-of-speech tagging, discussed in Section 5. This process enabled the spotting of errors and inconsistencies in the corpus, which was subsequently refined both automatically and manually.

2. Creating the Corpus

Corpora are commonly being distinguished by being untagged or tagged (that is, marked up with tags such as part-of-speech or sentence structure, etc.), as well as by being balanced or domain-specific. While a balanced corpus would provide a wide selection of different types of texts, the corpus discussed in this paper is made up of texts from the news domain only: all texts used in the corpus come from the web archives of Walta Information Center (`www.waltainfo.com`), a private Addis Ababa-based news and information service providing daily Ethiopia-related news coverage in Amharic and English.

All Amharic news items (8715 in total) from the start of the service in March 2001 to December 2004 were downloaded from the Walta web archive using a web-crawler and stored in an XML structure by staff at the Department of Computer and System Sciences, Stockholm University (Argaw and Asker, 2005).

Due to the above-mentioned lack of standard representation, a variety of ways had been used to encode the *fidel* in the Amharic texts. In order to have a unified representation of the corpus and to simplify further analysis, the Stockholm University staff transliterated all texts into SERA, "System for Ethiopic Representation in ASCII" (Yacob, 1997), a convention for transcription of Ethiopic into 7-bit ASCII, and then back to a common Unicode compatible font (Ethiopia Jiret). Both the Ethiopic and the SERA transliterated form are stored in the XML structure of the corpus, under different fields. The full, untagged corpus contains about 1.7 million words after preprocessing (Argaw and Asker, 2005).

A portion of the untagged corpus was selected for manual tagging. This part of the corpus consisted of 1065 news texts from September 11, 2001 to May 8, 2002 (i.e., the first eight months of the Ethiopian year 1994; the Ethiopian calendar runs approximately seven years and eight months behind the Gregorian). In total 207,315 words in the Unicode version (*fidel*), and 207,291 words in the SERA-transcribed version.

The texts were tagged by staff at ELRC, the Ethiopian Languages Research Center at Addis Ababa University, using a tag set developed at ELRC and described by Demeke and Getachew (2006). The tagset is made up of thirty classes based on type of word only: the tags contain no information on grammatical categories (such as number, gender, tense, and aspect). Table 1 contains short explanations of the different classes in the column labeled "Description". The tags used in the manual tagging are the ones in the column called 'ELRC' in the table (which is an adapted version of the word-class table given by Demeke and Getachew). The annotation was carried out by nine trained linguists, who wrote the proposed tags with pen on hard copies. The hand-written tags were later typed out and digitalized by non-linguists.

The tagged corpus is available at `nlp.amharic.org`.

3. Cleaning the Corpus

Unfortunately, the corpus available on the net contains quite a few errors and tagging inconsistencies: several portions contain inconsistent manual tagging, in addition to the inter-annotator disagreement which can be expected in any manual tagging endeavor.

3.1. Possible Error Sources

Obviously, the tagging procedure introduced several possible error sources, with nine persons doing the tagging and others inserting the hand-written tags into the electronic version of the corpus. Many errors are also due to lack of resources: lack of trained personnel and time/funding for the tagging meant that each section of the corpus was only tagged by one person, while lack of computational resources meant that typists (non-linguists) rather than the linguists themselves entered the tags into the digital version of the corpus —- leaving room for misreadings and misinterpretations of the hand-written tags by the typists. Finally, the transliteration of the texts written in Ethiopic script into the SERA (ASCII) version actually also added some errors, thus, for example, a few non-ASCII characters still remain in the ASCII part of the published corpus.

In its present state, the corpus is still useful, but a main aim of the present work was to improve its quality and to "clean" it. To this end, many non-tagged items have been tagged. In contrast, some items in the corpus contained double tags, which have been removed.

3.2. Multi-Word Expressions

The on-line tagged corpus contains several headlines of the news texts tagged as one multi-word unit, assigned the end-of-sentence punctuation tag (`<PUNC>`), while some headlines have no tag at all. For those cases all the words of the headlines were re-tagged separately.

In contrast, the corpus also contains several "true" multi-word expressions (MWE) that have been assigned a single tag by the annotators. However, since the words still can be written separately, that has introduced a source of error, with the non-final words of the MWEs having no tag directly assigned to them. Reflecting the segmentation of the original Amharic text, all white-spaces were removed from the SERA-transcribed version, merging multi-word units with a single tag into one-word units. The alternative would have been to assign new tags to each of the words in a collocation.

Both approaches have pros and cons. On one hand, tagging each part of an MWE separately increases the size of the corpus (measured in words), and potentially reduces ambiguity by creating fewer lexical units. On the other hand, if the human annotator had assigned a single tag to the entire MWE, it makes sense to also attach a single tag to it, reflecting the choice of the linguist: It is important to correct pure mistakes in the human annotation, but not to interfere with *conscious* decisions taken by the annotators. Furthermore, attaching tags to different parts of a collocation is problematic due to the lack of spelling standards for compounds in Amharic: given that a unit AB can be written both as A B and as AB in the source texts, it was deemed best to aim for consistency and remove any white-spaces inside the collocations (i.e., in effect enforcing the AB option).

3.3. Inconsistencies

Items such as `'`, `/`, etc., had been assigned a range of different tags, but have now been consistently re-tagged as punctuation (`<PUNC>`). Consistent tags have also been added to word-initial and word-final hyphens, by changing them into

Class	Tag description	ELRC	BASIC	SISAY
	Nouns			
1	Any basic or derived noun not matching classes 2–5	N	N	N
2	Verbal/infinitival noun, formed from any verb form	VN	N	N
3	Noun attached with a preposition	NP	N	N
4	Noun attached with a conjunction	NC	N	N
5	Noun with a preposition and a conjunction	NPC	N	N
	Pronouns			
6	Any pronoun not matching classes 7–9	PRON	PRON	**N**
7	Pronoun attached with a preposition	PRONP	PRON	**N**
8	Pronoun attached with a conjunction	PRONC	PRON	**N**
9	Pronoun with a preposition and a conjunction	PRONPC	PRON	**N**
	Verbs			
10	Any verb not matching classes 11–15	V	V	V
11	Auxiliary verb	AUX	**V**	AUX
12	Relative verb	VREL	V	V
13	Verb attached with a preposition	VP	V	V
14	Verb attached with a conjunction	VC	V	V
15	Verb with a preposition and a conjunction	VPC	V	V
	Adjectives			
16	Any adjective not matching classes 17–19	ADJ	ADJ	AJ
17	Adjective attached with a preposition	ADJP	ADJ	AJ
18	Adjective attached with a conjunction	ADJC	ADJ	AJ
19	Adjective with a preposition and a conjunction	ADJPC	ADJ	AJ
	Numerals			
20	Cardinal numeral not matching classes 22–24	NUMCR	NUM	NU
21	Ordinal numeral not matching classes 22–24	NUMOR	NUM	NU
22	Numeral attached with a preposition	NUMP	NUM	NU
23	Numeral attached with a conjunction	NUMC	NUM	NU
24	Numeral with a preposition and a conjunction	NUMPC	NUM	NU
	Others			
25	Preposition	PREP	PREP	AP
26	Conjunction	CONJ	CONJ	**AP**
27	Adverb	ADV	ADV	AV
28	Interjection	INT	INT	I
29	Punctuation	PUNC	PUNC	PU
30	Unclassified	UNC	UNC	R

Table 1: The three tagsets with tag descriptions (bold-face tags mark genuine differences)

two tokens, and making the following replacements:

$$-\textit{word}\ \texttt{<TAG>} \;\Rightarrow\; -\ \texttt{<PUNC>}\ \textit{word}\ \texttt{<TAG>} \quad (1)$$

$$\textit{word}-\ \texttt{<TAG>} \;\Rightarrow\; \textit{word}\ \texttt{<TAG}'\texttt{>}\ -\ \texttt{<PUNC>} \quad (2)$$

For the word-final case (2) this sometimes meant that a new tag (<TAG′>) had to be introduced for a word, namely in those cases when the originally assigned tag was <PUNC> (i.e., the tag relating to the -), rather than the tag relating to *word*. In the same fashion, a couple of cases of - tagged as <N> or <NP> were changed to punctuation, <PUNC>.

Making these changes is consistent with about half of the previous manual tagging of the corpus.

3.4. Tagging Errors and Misspellings

Furthermore, some direct tagging errors and misspellings have been corrected. Those include tag misspellings (or typos), such as <PUNC instead of <PUNC>. The published corpus also contains 14 cases similar to

`leityoPya <NP>rEdiyo <N>` (3)

that is, with >rEdiyo written without a space. Those have all been corrected, as well as: two occurrences of meaningless] in the text, three occurrences of superfluous $ characters, three occurrences of \ characters, and one occurrence of a ... sequence.

3.5. Known Remaining Problems in the Corpus

Both time expressions and numbers in the corpus suffer from not having been consistently tagged at all, but just removing some of them can hardly be the way to handle that, so those had to be left as they were. In addition, many words have been transcribed into SERA in several versions, with only the cases differing. However, this is also difficult to account for since the SERA notation in general lets upper and lower cases of the English alphabet represent different symbols in *fidel* (Ethiopic script). Thus those were also left unchanged.

After corrections the SERA-transcribed version of the corpus contains 200,863 tagged words (compared to 207,291 words with 200,533 tags in the original corpus). 33,408 are unique wordforms, with 39,921 possible word-tag combinations (32,556 resp. 40,510 in the original corpus). 86% of the wordforms (28,731) have only one possible tag assignment, that is, are unambiguous. 3,533 have two possible tags, 744 have three, and 400 have four or more, including two words (*beteleym* and *yahl*) that have been given eleven possible tag assignments each. In the original corpus, 82% (26,844 wordforms) were unambiguous.

3.6. Alternative Cleaning Methods

In addition to the cleaning procedure described here, the corpus has also been partially cleaned in two other independent efforts. Hence Tachbelie (2010) primarily targeted inconsistencies in the annotations of collocations. She assigned separate tags to all the words in an MWE, rather than merging the words in the MWE and assigning it a single tag, as in the approach taken in Section 3.2. After cleaning, Tachbelie's version of the corpus contained 205,355 tokens.

Gebre (2010) carried out a more thorough cleaning of the corpus, in many ways following a strategy similar to the one of the present paper, but with the addition of treating collocations similarly to Tachbelie (2010). After cleaning, his version of the corpus contained 206,929 tokens. Quite importantly, the number of ambiguous tags was drastically reduced: in the original corpus only 38% of the tokens were unambiguous, but after removing tagging inconsistencies, this increased to 74% (compared to 82% resp. 86% above).

4. Re-Tagging and Splitting the Corpus

The manual tagging of the corpus utilized a 30-class tag set described by Demeke and Getachew (2006). The new, corrected corpus has been marked up with three different tagsets, all shown in Table 1.

4.1. Three Tagsets

Firstly, the full, original 30-tag set developed at the Ethiopian Languages Research Center. This version of the corpus will hereinafter be referred to as 'ELRC'. It differs from the published corpus in way of the corrections described in the previous section.

Secondly, the corpus was mapped to 11 basic tags also given by Demeke and Getachew (2006). This set consists of ten word classes: Noun, Pronoun, Verb, Adjective, Preposition, Conjunction, Adverb, Numeral, Interjection, and Punctuation, plus one tag for problematic words (unclear: <UNC>). This tagset will be called 'BASIC' below.

The main differences between the twose tagsets pertain to the treatment of prepositions and conjunctions: in 'ELRC' there are specific classes for, e.g., <PRONP>, <PRONC>, and <PRONPC>, i.e., for pronouns attached with preposition, conjunction, and both proclitic preposition and enclitic conjunction (similar classes occur for nouns, verbs, adjectives, and numerals). In addition, numerals are divided into cardinals and ordinals, verbal nouns are separated from other nouns, while auxiliaries and relative verbs are distinguished from other verbs. Hence the 'ELRC' tagset is made up of thirty subclasses of the eleven 'BASIC' classes.

Baseline	ELRC	BASIC	SISAY
Most frequent tag overall	35.50	58.26	59.61
Most frequent tag for word	79.64	83.05	83.10
Most likely tag	82.64	90.07	90,19

Table 2: Baselines for the three tagsets

Thirdly, for comparison reasons, the 'ELRC' tagset was also mapped to a tagset introduced by Adafre (2005), which will be referred to as 'SISAY' and is shown in the right-most column of Table 1. It consists of 10 different classes, including one for Residual (R) which was assumed to be equivalent to <UNC>. In addition, both <CONJ> and <PREP> were mapped to Adposition (AP) in Adafre's classification, and both <N> and <PRON> to N. The other mappings were straight-forward, except that the 'BASIC' tagset groups all verbs together, while Adafre (2005) kept Auxiliary (AUX) as its own class. The tags in bold-face in Table 1 indicate the major differences between this tagset and the 'BASIC' one. (Note that the verb classes 13–15 include auxiliaries attached with prepositions and/or conjunctions. It is unclear whether Adafre kept those together with the V class or with AUX; here they have been assumed to belong to the V class.)

Table 2 shows baselines for the tagsets. The "most frequent tag overall" baseline is the number of tokens tagged with the tag most frequent overall in the corpus (i.e., regular nouns; <N>) in relation to the total number of tokens (200,863). The "most frequent tag for word" baseline is the accuracy which would be obtained if labeling every word with the tag occurring most frequently with it in the corpus. The "most likely tag" baseline is the hardest to beat since it combines the other two, labeling known words with their most frequent tags and unknown words with <N>. It is fair to say that the higher the baseline, the easier it is to assign the tags of the tagset to a corpus: by simply guessing that a word is a standard noun, almost 6 out of 10 words would be correctly tagged using the 'SISAY' tagset or 'BASIC' tagsets. However, only 1 out of 3 words would be correctly tagged if applying the same strategy to the 'ELRC' tagset.

4.2. Splitting the Corpus into Folds

For evaluation of machine learning and statistical methods, it is common and useful to split a corpus into ten folds, each containing about 10% of the corpus. However, this can be done in several ways, for example, by taking the first 10th of the corpus as the first fold or by taking the folds to include every 10th word in the corpus (i.e., fold 1 consisting of words 1, 11, 21, etc.). The folds can also be of exactly equal size or be allowed to vary somewhat in size in order to preserve logical units (e.g., to keep complete sentences in the same fold). Furthermore they can be *stratified*, meaning that the distributions of different tags are equal over all folds.

The importance of how the folds are created points to one of the problems with n-fold cross validation: even though the results represent averages after n runs, the choice of the original folds to suit a particular machine learning or statistical strategy can make a major difference to the final result. For ease of straight-forward comparison between different studies, the same folds have to be used (cmf. the discussion in the next section, in particular the results of the MaxEnt tagger on different folds). We thus created a "standard" set of folds for the corpus.

The "standard" folds were created by chopping the corpus into 100 pieces, each of about 2000 words in sequence, while making sure that each piece contained full sentences (rather than cutting off the text in the middle of a sentence), and then merging sets of ten pieces into a fold. Thus the folds represent even splits over the corpus, to avoid tagging inconsistencies, but the sequences are still large enough to potentially make knowledge sources such as n-grams useful.

The resulting folds on average contain 20,086 tokens. Of those, 88.26% (17,727) are known, while 11.74% (2,359) are unknown, that is, tokens that are not in any of the other nine folds (if those were used for training, in a 10-fold evaluation fashion). Notably, the fraction of unknown words is about four times higher than in the English Wall Street Journal corpus (which, in contrast, is about six times larger).

5. On Tagging the Untagged Corpus

Having a tagged corpus is useful for many types of statistical and machine-learning based approaches to language processing. As an example application we will here look at automatic part-of-speech tagging, that is, the task of automatically assigning exactly those tags to the words that the human annotators assigned (or should have assigned). This could be a straight-forward way to extend the tagged corpus in itself, since the manually tagged portion of the corpus is only about 12% (1065 of 8715 news items). Clearly, tagging the remaining corpus would be useful, but as pointed out in Section 3.1., manually annotating the entire corpus would be an endeavour which would have to rely on human resources that are both scarce, expensive and inconsistent.

Part-of-speech tagging is a classification task, with the object of assigning lexical categories to words in a text. Within the computational linguistic community, part-of-speech tagging has been a fairly well-researched area. Most work so far has concentrated on English and on using supervised learning methods. The best results on the English Wall Street Journal corpus are now above 97%, using combinations of taggers: Spoustová et al. (2009) report achieving an accuracy of 97.43% by combining rule-based and statistically induced taggers. However, recently the focus has started to shift towards other languages and unsupervised methods.

The best reported figures for part-of-speech tagging for Arabic are comparable to those for English (Habash and Rambow, 2005; Mansour, 2008), while those for other Semitic languages are a bit lower: with 21 tags and a 36,000 word news text, Bar-Haim et al. (2008) report 86.9% accuracy on Hebrew. For Amharic, the best results reported before the availability of the corpus discussed here came from experiments by Adafre (2005) on using a Conditional Random Fields (CRF) tagger, an effort restricted by only having access to a news text corpus of 1,000 words. Using the 10-tag 'SISAY' tagset (see Table 1), the tagger achieved a 70.0% accuracy. Adding a machine-readable dictionary and bigram information improved performance to 74.8%, (i.e., just between the "most frequent tag overall" and "most likely tag" baselines obtained on the 200k word corpus).

The present corpus has been the topic of three independent sets of part-of-speech tagging experiments, each running on a differently cleaned version of the corpus, as described in Section 3. The differences in the cleaning strategies (Section 3.6.), as well as differences in the ways the corpus was split up into folds, might explain why the results of the three tagging experiments are not directly compatible.

Tachbelie (2010) achieved an overall accuracy on the 'ELRC' tagset of 84.4% using the Support Vector Machine-based tagger SVMTool (Giménez and Màrquez, 2004) and 82.9% with TnT, Trigrams and Tags (Brants, 2000), a tagger based on Hidden Markov Models, when training on 95% of the corpus and testing on 5%. For unknown words, she received an accuracy of 73.6% for SVMTool and 48.1% for TnT. The set of taggers was later on (Tachbelie et al., 2011) extended with MBT, a Memory-Based Learning tagger (Daelemans et al., 1996), and a CRF-based tagger developed in the toolkit CRF++ (Lafferty et al., 2001). Again SVMTool performed best on unknown words (75.1%) and overall (86.3%) while CRF++ was best on known words (87.6%). Interestingly (and surprisingly), no gain was achieved by combining taggers. However, adding segmentation and reducing the tagset to 16 word-classes did: the results improved by 7% over the board, including a 93.5% overall accuracy for SVMTool on the reduced (16 class) tagset.

Gambäck et al. (2009) report the average 10-fold cross-validated accuracy obtained from three taggers when trained on 90% of the corpus and evaluated on the remaining 10%. Then TnT performed best on known words (over 90% for all three tagsets, incl. 94% on 'SISAY'), but terribly on the unknown ones (only 52% 'ELRC' and 82% on the others), for 85.6% resp. 92.6% accuracy overall. SVMTool out-performed the other taggers both on the unknown words and overall, reaching 92.8% overall accuracy on the 'SISAY' and 'BASIC', and 88.3% on the more difficult 'ELRC' (with 78.9% on unknown words in 'ELRC' and 88.2–88.7% on the others). The third strategy tested was a Maximum Entropy (MaxEnt) tagger (Ratnaparkhi, 1996) as implemented in McCallum's java machine learning package MALLET (`http://mallet.cs.umass.edu`). The MaxEnt tagger performed more in the middle of the road: 87.9% overall accuracy on 'ELRC' and 92.6% on both the smaller sets.

However, the MaxEnt tagger clearly out-performed the other taggers on all tagsets when allowed to create its own, stratified folds: 94.5–94.6% on the'SISAY' and 'BASIC' tagsets, and 90.8% on 'ELRC'. The dramatic increase in the MaxEnt tagger's performance on the stratified folds is surprising, but a clear indication of why it is so important to create a "standard" set of folds, as discussed in Section 4.2.

Gebre (2010) tested both TnT and a CRF-based tagger on the corpus, as well as a version of Brill's transformation-based tagger (Brill, 1995), all on the 'ELRC' tagset. Then TnT and the Brill tagger performed on par, with a 10-fold cross-validated average accuracy of 87.1% and 87.4% respectively, while the tagger based on Conditional Random Fields did clearly better, reaching a 91.0% accuracy, the best result reported for Amharic part-of-speech tagging to date.

6. Conclusions

The paper has described how a 200,863 word part-of-speech tagged corpus of Amharic news texts was created by cleaning, normalising and verifying a publically available manually tagged corpus. The corpus has been marked up with three different tagsets (of 30, 11 and 10 tags each), and also been split into standardized folds for evaluation purposes.

The corpus has been used to train state-of-the-art part-of-speech taggers on Amharic. The best reported results so far show tagging accuracy of around 90% on the most difficult tagset, which is not very encouraging, and not useful for the task of tagging the remainder of the corpus. Rather, figures above 96% would be needed, a level which the best Amharic taggers currently fail to reach even on the easier tagsets.

The efforts on applying machine learning approaches to the task of tagging the corpus has still been useful, though, since it has meant that several errors and tagging inconsistencies in the corpus were spotted and subsequently corrected.

Acknowledgments

This work was carried out while the author was a researcher at SICS, Swedish Institute of Computer Science AB, and has been partially funded by Sida, the Swedish International Development Cooperation Agency, through the Swedish Programme for ICT in Developing Regions (SPIDER).

Thanks to Atelach Alemu Argaw, Lars Asker, Girma Awgichew Demeke, Binyam Gebrekidan Gebre, Mesfin Getachew, Fredrik Olsson, and the staff at ELRC.

7. References

ACL. 2005. *43rd Annual Meeting of the Assoc. for Computational Linguistics*, Ann Arbor, Michigan, June.

Sisay Fissaha Adafre. 2005. Part of speech tagging for Amharic using conditional random fields. In ACL (2005), pp. 47–54. Workshop on Computational Approaches to Semitic Languages.

Atelach Alemu Argaw and Lars Asker. 2005. Web mining for an Amharic-English bilingual corpus. In *1st Int. Conf. on Web Information Systems and Technologies*, pp. 239–246, Deauville Beach, Florida, May.

Roy Bar-Haim, Khalil Sima'an, and Yoad Winter. 2008. Part-of-speech tagging of modern Hebrew text. *Natural Language Engineering*, 14:223–251.

Thorsten Brants. 2000. TnT — a statistical part-of-speech tagger. In *6th Conf. on Applied Natural Language Processing*, pp. 224–231, Seattle, Washington, April. ACL.

Eric Brill. 1995. Transformation-based error-driven learning and Natural Language Processing: A case study in part of speech tagging. *Computational Linguistics*, 21:543–565.

CIA. 2012. *The World Factbook: Ethiopia*. The Central Intelligence Agency, Washington, DC, March. Webpage. https://www.cia.gov/library/publications/the-world-factbook/geos/et.html.

CSA. 2010. *Population and Housing Census of 2007*. Ethiopia Central Statistical Agency, Addis Ababa, Ethiopia, July. Online CD. http://www.csa.gov.et/index.php?Itemid=590.

Walter Daelemans, Jakub Zavrel, Peter Berck, and Steven Gillis. 1996. MBT: A memory-based part of speech tagger-generator. In *4th Workshop on Very Large Corpora*, pp. 14–27, Copenhagen, Denmark.

Girma Awgichew Demeke and Mesfin Getachew. 2006. Manual annotation of Amharic news items with part-of-speech tags and its challenges. *ELRC Working Papers*, 2:1–17, March.

EACL. 2009. *12th Conf. of the Europ. Chap. of the Assoc. for Computational Linguistics*, Athens, Greece, March.

Björn Gambäck, Fredrik Olsson, Atelach Alemu Argaw, and Lars Asker. 2009. Methods for Amharic part-of-speech tagging. In EACL (2009), pp. 104–111. 1st Workshop on Language Technologies for African Languages.

Michael Gasser, 2009. *HornMorpho 1.0 User's Guide*. Bloomington, Indiana, December.

Michael Gasser. 2011. HornMorpho: a system for morphological processing of Amharic, Oromo, and Tigrinya. In HLTD (2011), pp. 94–99.

Binyam Gebrekidan Gebre. 2010. Part of speech tagging for Amharic. MSc Thesis, Law, Social Sciences and Communications, Univ. of Wolverhampton, England, June.

Jesús Giménez and Lluís Màrquez. 2004. SVMTool: A general POS tagger generator based on support vector machines. In *4th Int. Conf. on Language Resources and Evaluation*, pp. 168–176, Lisbon, Portugal, May. ELRA.

Nizar Habash and Owen Rambow. 2005. Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In ACL (2005), pp. 573–580.

HLTD. 2011. *Conference on Human Language Technology for Development*, Alexandria, Egypt, May.

Grover Hudson. 1999. Linguistic analysis of the 1994 Ethiopian census. *Northeast African Studies*, 6:89–107.

John D. Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *18th Int. Conf. on Machine Learning*, pp. 282–289, Williamstown, Maryland, USA, June.

Saib Mansour. 2008. Combining character and morpheme based models for part-of-speech tagging of Semitic languages. MSc Thesis, Computing Science Dept., Technion, Haifa, Israel.

Adwait Ratnaparkhi. 1996. A maximum entropy model for part-of-speech tagging. In Eric Brill and Ken Church, editors, *1st Conf. on Empirical Methods in Natural Language Processing*, pp. 133–142, Univ. of Pennsylvania, Philadelphia, Pennsylvania, May. ACL.

Drahomíra Spoustová, Jan Hajič, Jan Raab, and Miroslav Spousta. 2009. Semi-supervised training for the averaged perceptron POS tagger. In EACL (2009), pp. 763–771.

Martha Yifiru Tachbelie, Solomon Teferra Abate, and Laurent Besacier. 2011. Part-of-speech tagging for under-resourced and morphologically rich languages — the case of Amharic. In HLTD (2011), pp. 50–55.

Martha Yifiru Tachbelie. 2010. Morphology-based language modeling for Amharic. PhD Thesis, Dept. of Informatics, Univ. of Hamburg, Germany, August.

Daniel Yacob. 1997. System for Ethiopic Representation in ASCII (SERA). Webpage. http://www.abyssiniacybergateway.net/fidel/sera-97.html.

Resource-Light Bantu Part-of-Speech Tagging

Guy De Pauw[†], Gilles-Maurice de Schryver[‡], Janneke van de Loo[†]

[†]CLiPS - Computational Linguistics Group
University of Antwerp
Antwerp, Belgium
guy.depauw@ua.ac.be
janneke.vandeloo@ua.ac.be

[‡]Dept of Languages and Cultures
Ghent University
Ghent, Belgium
gillesmaurice.deschryver@UGent.be

[‡]Xhosa Dept
University of the Western Cape
Cape Town, South Africa

Abstract

Recent scientific publications on data-driven part-of-speech tagging of Sub-Saharan African languages have reported encouraging accuracy scores, using off-the-shelf tools and often fairly limited amounts of training data. Unfortunately, no research efforts exist that explore which type of linguistic features contribute to accurate part-of-speech tagging for the languages under investigation. This paper describes feature selection experiments with a memory-based tagger, as well as a resource-light alternative approach. Experimental results show that contextual information is often not strictly necessary to achieve a good accuracy for tagging Bantu languages and that decent results can be achieved using a very straightforward unigram approach, based on orthographic features.

1. Introduction

Part-of-speech tagging is often considered as a prototypical classification task in the field of Natural Language Processing (NLP). It can more generally be described as *sequence tagging* and methods suitable for part-of-speech tagging can be directly applied to a wide variety of other NLP tasks, such as word sense disambiguation (Veenstra et al., 1999), phrase chunking (Ramshaw and Marcus, 1995) and concept tagging (Hahn et al., 2008). It is commonly accepted that contextual features constitute the most important information source to trigger the correct sequence tag of ambiguous words, although for the task of part-of-speech tagging (pseudo-)morphological features are often used as well (Ratnaparkhi, 1996; Daelemans et al., 2010).

In the English sentence in Example 1, part-of-speech disambiguation of the token *can*, which can be a modal, verb or noun, can be performed on the basis of the preceding determiner.

(1) *The can is empty .*

This disambiguation task can be automatically induced on the basis of annotated data: a statistical technique or machine learning algorithm observes the manually annotated data and automatically identifies the most important linguistic features towards disambiguation. This approach has the advantage of being language independent: all that is needed, is annotated data in the target language.

While these *data-driven* approaches have yielded state-of-the-art part-of-speech tagging accuracies for a number of sub-Saharan African languages, such as Swahili (De Pauw et al., 2006), Amharic (Gambäck et al., 2009), Wolof (Dione et al., 2010) and Northern Sotho (de Schryver and De Pauw, 2007), no research efforts exist that explore which type of linguistic features contribute to accurate part-of-speech tagging for the languages under investigation. This paper describes feature selection experiments with a memory-based tagger, as well as a resource-light alternative approach. Experimental results show that contextual information is often not strictly necessary to achieve a good accuracy for tagging Bantu languages and that decent results can be achieved using a very straightforward unigram approach, based on orthographic features.

This paper is organized as follows: in Section 2, we describe some of our previous research efforts on part-of-speech tagging of Bantu languages and we introduce the data sets that were used in the experiments described in this paper. Section 3 outlines experimental results using an off-the-shelf data-driven tagger and discusses the automatically determined optimal combination of features for disambiguation. We introduce a new data-driven approach to part-of-speech tagging of Bantu languages in Section 4 and further contrast it with the traditional, context-driven approach by means of learning curve experiments in Section 5. We conclude with a discussion of the main insights gained from these experiments and pointers for future research in Section 6.

2. Part-of-Speech Tagging for Bantu Languages

In this paper, we will investigate two different approaches to data-driven part-of-speech tagging for four different languages. The languages under investigation are: (i) Swahili, spoken by over fifty million people in eastern Africa, especially in Tanzania and Kenya, (ii) Northern Sotho and (iii) Zulu, two of the eleven official languages in South Africa, spoken by respectively four and ten million people, and (iv) Cilubà, spoken by six million people in the Democratic Republic of the Congo. There is no ideological reason behind the selection of these languages, other than the fact that they are among the few Sub-Saharan African languages that have part-of-speech tagged data available to them.

All four happen to be Bantu languages, which is a sub-group of one of Africa's four language phyla, Niger-Congo. The term sub-group is an understatement, as genetically the classification path goes through Niger-Congo > Mande-Atlantic-Congo > Ijo-Congo > Dogon-Congo > Volta-Congo > East Volta-Congo > Benue-Congo > East Benue-Congo > Bantoid-Cross > Bantoid > South Bantoid > Narrow Bantu. The Narrow Bantu languages, more of-

	Swahili	Northern Sotho	Zulu	Cilubà
Number of sentences	152,877	9,214	3,026	422
Number of tokens	3,293,955	72,206	21,416	5,805
POS-Tag set size	71	64	16	40
% of ambiguous words	22.41	45.27	1.50	6.70
Average % of unknown words	3.20	7.50	28.63	26.93

Table 1: Quantitative information for Swahili, Northern Sotho, Zulu and Cilubà data sets.

ten simply referred to as 'the Bantu languages', are thus truly a late offshoot historically speaking, and still there are about 500 Bantu languages, the largest unitary group on the African continent.

Data Sets

For Swahili, we will use the Helsinki Corpus of Swahili (Hurskainen, 2004a) as our data set of choice. The Helsinki Corpus of Swahili has been automatically annotated using SALAMA, a collection of finite-state NLP tools for Swahili (Hurskainen, 2004b). It is therefore important to point out that this resource constitutes silver-standard data, rather than gold-standard (i.e. manually annotated) data. Previous research (De Pauw et al., 2006) investigated the applicability of four different off-the-shelf, data-driven part-of-speech taggers and various system combination techniques, yielding an overall tagging accuracy of up to 98.6%. In this paper, we will use the same, cleaned-up version of the Helsinki Corpus of Swahili, containing over three million tokens, as described in De Pauw et al. (2006).

For Northern Sotho, we use a gold-standard, manually annotated part-of-speech tagged corpus, first described in de Schryver and De Pauw (2007). This corpus was used as training data for a maximum-entropy based tagger, which is able to deal with pseudo-morphological information that is more suitable to tagging a Bantu language, compared to existing off-the-shelf taggers. de Schryver and De Pauw (2007) describe experiments using this small data set of a little over 10,000 tokens and report a more than encouraging tagging accuracy of 93.5%. Since then, additional data has been automatically annotated by this tagger and manually corrected, yielding a modestly-sized, gold-standard corpus of about 70,000 tokens that will be used during the experiments described in this paper.

A publicly available, but at 21,000 tokens modestly sized, part-of-speech tagged corpus of Zulu is described in Spiegler et al. (2010). This data set will be included in the experiments as well.

Finally, a very small annotated corpus for Cilubà of just 6,000 tokens was manually annotated. While this data set is clearly too diminutive as training material for off-the-shelf data-driven taggers, experiments will show that the Bag-of-Substrings approach (cf. Section 4) is able to perform fairly accurate tagging of unknown words on the basis of a limited amount of training data.

Table 1 shows some quantitative information for the four data sets under investigation, such as the number of sentences and tokens in the first two lines. The third line displays the number of distinct POS tags, used in the corpus. As a rule of thumb, the larger the tag set, the more fine-grained the morpho-syntactic description and consequently the more difficult the task of part-of-speech tagging.

The fourth line in Table 1 expresses the percentage of words in the corpus that are lexically ambiguous, i.e. have been observed with more than one tag[1]. A low lexical ambiguity rate typically indicates a fairly straightforward part-of-speech tagging task. For the Zulu data set, for instance, 98% of the words do not need disambiguation. This is normal for languages which have both a rich morphology and are written conjunctively: here a token typically has considerable affixation, encoding its morpho-syntactic (and often also semantic) properties and is therefore less likely to be lexically ambiguous. This does not hold, of course, for languages with a disjunctive writing system, such as Northern Sotho, as is apparent from its lexical ambiguity rate in Table 1. The degree of conjunctiveness / disjunctiveness (Prinsloo and de Schryver, 2002) for Cilubà lies in-between that of Zulu and Northern Sotho, as does its lexical ambiguity rate.

The last line of Table 1 indicates the average expected number of unknown words in unseen data. For the larger data sets (Swahili and Northern Sotho), this percentage is reasonably low. For the smaller data sets (Zulu and Cilubà), however, this equals to more than one out of four unknown tokens in a typical data set. In Section 4, we will describe a novel approach that is able to classify unknown words with a higher degree of accuracy than a traditional, context-driven tagging method.

3. Context-Driven Tagging

Data-driven taggers have become staple tools in the field of Natural Language Processing and a wide range of different implementations, using a variety of machine learning and statistical techniques, are publicly available. For the world's most commercially interesting languages, these methods are well researched. For Sub-Saharan African languages however, they have only recently been applied. In this section, we will describe experiments with MBT, a data-driven tagger that uses a memory-based learning classifier as its backbone (Daelemans et al., 2010).

Memory-Based Tagging

MBT uses annotated data to automatically induce a tagger that is able to classify previously unseen sentences. To this end, it actually builds two separate memory-based learning classifiers: one for known words (i.e. words that have

[1] As a point of reference: the English Wall Street Journal Corpus (Marcus et al., 1993) contains about 35% lexically ambiguous words.

	Swahili		
	Known	**Unknown**	**Total**
Baseline	97.7	73.37	96.92
	±0.03	±0.37	±0.02
MBT	**98.43**	89.81	98.16
	±0.03	±0.59	±0.04
BoS	97.4	**93.52**	97.27
	±0.03	±0.28	±0.03
Best Combo	98.43	93.52	**98.27**

	Northern Sotho		
	Known	**Unknown**	**Total**
Baseline	90.38	61.81	88.24
	±0.44	±2.53	±0.45
MBT	**95.73**	81.72	94.68
	±0.27	±1.86	±0.24
BoS	85.71	**84.1**	85.59
	±0.38	±1.69	±0.34
Best Combo	95.73	84.1	**94.86**

	Zulu		
	Known	**Unknown**	**Total**
Baseline	99.65	52.02	86.00
	±0.16	±2.38	±1.13
MBT	99.61	75.51	92.71
	±0.19	±1.74	±0.66
BoS	**99.65**	**83.32**	**94.97**
	±0.16	±1.47	±0.50
Best Combo	99.65	83.32	94.97

	Cilubà		
	Known	**Unknown**	**Total**
Baseline	95.72	48.85	82.77
	±1.85	±5.72	±2.24
MBT	**95.91**	62.13	86.58
	±1.66	±5.66	±2.7
BoS	95.84	**72.24**	89.32
	±1.70	±5.56	±2.69
Best Combo	95.91	72.24	**89.37**

Table 2: Results (tagging accuracy & standard deviation (%)) of ten-fold cross validation experiment for Swahili, Northern Sotho, Zulu and Cilubà data sets.

previously been encountered in the training corpus) and another classifier that is able to tag unknown words. In a first phase, MBT constructs for each token in the training corpus a so-called `ambitag`, a single token encoding the tags that have been associated with that word in the training corpus. A word such as "*can*" for example, may receive an ambitag `MD_NN_VB`, which means that it has been encountered as a modal, a noun and a verb.

The training data is then transformed into a series of vectors, describing each word in its context. The known-words classifier uses the `ambitag` of each word as its primary feature. The user can then add extra contextual information: the tags of the left context of the word can be added to the vector to allow for disambiguation of cases such as "*can*" in Example 1. Also the right context of the token can be added, but since tagging is performed from left to right and the right context is not yet disambiguated, we need to refer to the ambitags of the tokens on the right hand side. Not only (ambi)tags can be added, but the actual tokens as well. This is usually only useful for highly frequent tokens, such as functors and punctuation marks.

For the unknown-words classifier, contextual information can be used as well (although no ambitag can be constructed for an unknown word). Additionally, some orthographic features are supported by MBT as well, such as the first or last *n* graphemes, or more general features such as hyphenation or capitalization features. While these features work well for unknown words, MBT is mostly driven by contextual features. To contrast it with the approach described in Section 4, we will refer to MBT as a *context-driven tagger*.

Experiments

To obtain reliable accuracy scores for our data sets, we use the technique of ten-fold cross validation to evaluate the respective techniques: the entire corpus is split into ten slices of equal size. Each slice is used as an evaluation set once, while eight slices are used to train the tagger and the remaining slice is used as a validation set to establish the optimal set of features.

The optimal features for part-of-speech tagging cannot be predetermined, as each data set requires different combinations of features. To facilitate the process of feature selection, we automated the search for optimal features: through the stepwise addition of features to the MBT classifiers and the observation of changes in the tagging accuracy on the validations set, we can dynamically establish optimal features for each data set (cf. infra for a discussion on feature selection).

The final evaluation of the tagger is performed on the held-out evaluation set, which is not used at any point during training, thereby establishing accuracy figures on truly unseen data. Table 2 displays the results of the 4x10 experiments. We provide scores for known words, unknown words and the overall tagging accuracy.

The first line in each sub-table displays baseline accuracies, achieved by using a simple unigram tagger, which always selects the most frequent tag for each known word and the overall most frequent tag for unknown words. For Swahili, the baseline accuracy is 97%, which can mostly be attributed to the sheer size and the silver standard nature of the corpus. Baseline tagging results for Northern Sotho (88%), Zulu (86%) and Cilubà (83%) on the other hand, are not so good.

Using MBT improves overall tagging accuracy over the baseline for all data sets under investigation. Swahili can be tagged with a projected accuracy of 98.16% and even unknown words are handled fairly well by this tagger (89.81%). The previously reported result for Northern Sotho (93.5%, cf. de Schryver and De Pauw (2007)) is sig-

	Known Words	Unknown Words
Swahili	wddfaaww	chsssppppppwddddFaww
N. Sotho	dddfaaa	csppdFa
Zulu	df	sssppwdF
Cilubà	wddf	sspppddFa

Table 3: Optimal features for memory-based tagging of Swahili, Northern Sotho, Zulu and Cilubà data sets (majority vote over ten folds).

nificantly improved up to 94.68%, thanks to the additional training data.

The Zulu tagger is able to tag known words almost perfectly, although MBT scores slightly lower than the baseline method. This indicates that tagging known tokens basically amounts to table look-up for this data and that the extra features MBT uses, are not helpful. The underwhelming score for (the copious number of) Zulu unknown words drags the overall tagging accuracy down to 92.71%. A similar situation can be observed for memory-based tagging of Cilubà, although known words accuracy barely exceeds baseline accuracy. This is undoubtedly due to the diminutive size of the corpus.

Feature Selection

Table 3 provides an overview of the optimal features, automatically selected during the development phase, for each of the four data set. Some general tendencies can be observed: for the prediction of unknown words, contextual features (**d** for left context and **a** for ambiguous right context) play a fairly unimportant role, except for Swahili. Particularly the right context does not seem very informative, neither for the prediction of tags for unknown words, nor for known words. Prefix (**p**) and suffix features (**s**) on the other hand are abundantly used for the prediction of tags for unknown words. Capitalization (**c**) and hyphenation (**h**) features are used for the Swahili and Northern Sotho data as well. The use of word tokens (**w**) as an information source is fairly limited, except for the more expansive data sets.

For known words, it can be observed that the disjunctively written language of Northern Sotho benefits heavily from contextual information, as is to be expected. Also for the large Swahili data set, contextual features play an important role. For the more diminutive Zulu and Cilubà data sets however, contextual features are fairly sparsely used during disambiguation.

The experimental results show that context-driven tagging is a viable solution for the languages under investigation, with the exception of Cilubà, which simply does not have enough data available to it to trigger any kind of useful contextual information source. For all of the languages under investigation, the handling of unknown words poses the biggest limitation on achieving state-of-the-art tagging accuracy. In the next section, we will describe an alternative, *resource-light* approach that is able to overcome this bottleneck, while also limiting the observed decrease in known words tagging accuracy.

4. Bag-of-Substrings Tagging

While state-of-the-art tagging accuracy can be achieved for Swahili and Northern Sotho and a fairly reasonable accuracy for Zulu, we need to take into account that the majority of the languages on the African continent are more akin to Cilubà in terms of linguistic resources. Most Sub-Saharan African languages are in fact decidedly resource-scarce: digital linguistic resources are usually not available and while data can be mined off the Internet in a relatively straightforward manner (Hoogeveen and De Pauw, 2011), annotated corpora are few and far between.

In previous research efforts, we have investigated ways to circumvent this by means of projection of annotation (De Pauw et al., 2010; De Pauw et al., 2011) and by means of unsupervised learning techniques (De Pauw et al., 2007; De Pauw and Wagacha, 2007). The latter research efforts attempted to automatically induce morphological features on the basis of a raw, unannotated lexicon of words in the respective target languages. In this section, we will describe how we can use the same technique, dubbed *Bag-of-Substrings*, to perform part-of-speech tagging on the basis of scarce linguistic resources.

The general idea behind the Bag-of-Substrings approach is simple: each token is described as a collection of its substrings, which function as features towards some kind of classification task, in this case part-of-speech tagging. We will illustrate the conversion on the basis of the tagged Swahili Example 2[2]:

(2) Adam$_{PROPNAME}$ alionekana$_{V}$ chumbani$_{N}$ kwake$_{PRON}$ hana$_{NEG}$ fahamu$_{N}$.$_{FULL-STOP}$

This is converted into the representation in Figure 1. For each word we list all of the possible substrings and indicate whether it occurs at the beginning (P=), end (S=) or middle (I=) of the word or whether it constitutes the word itself (W=). These orthographic features encode a lot of potentially useful morphological information, although most features are not relevant towards the actual prediction of the class, i.e. part-of-speech tag.

The advantage of this approach is that we do not need to predefine which features are needed for classification, nor do we require any knowledge about the morphology of the language in question. All of the features are presented to a maximum entropy-based, machine learning classifier (Le, 2004), which will automatically determine during the training phase which of these features are salient in terms of their predictive power. For example, the feature P=A for the word *Adam* implicitly encodes the capitalization of the word and will therefore probably be strongly correlated with the tag propname. Likewise, the features P=a and I=li for the word *alionekana* encode its prefixes. Furthermore, the W= features still enable basic table look-up functionality for known words.

This method of part-of-speech tagging effectively takes all contextual information out of the equation. Instead, all of

[2] *Adam appeared to be distraught.*

Class	Features
PROPNAME	P=A P=Ad P=Ada **W=Adam** I=d I=da S=dam I=a S=am S=m
V	P=a P=al P=ali P=alio P=alion P=alione P=alionek P=alioneka P=alionekan **W=alionekana** I=l I=li I=lio I=lion I=lione I=lionek I=lioneka I=lionekan S=lionekana I=i I=io I=ion I=ione I=ionek I=ioneka I=ionekan S=ionekana I=o I=on I=one I=onek I=oneka I=onekan S=onekana I=n I=ne I=nek I=neka I=nekan S=nekana I=e I=ek I=eka I=ekan S=ekana I=k I=ka I=kan S=kana I=a I=an S=ana I=n S=na S=a
N	P=c P=ch P=chu P=chum P=chumb P=chumba P=chumban **W=chumbani** I=h I=hu I=hum I=humb I=humba I=humban S=humbani I=u I=um I=umb I=umba I=umban S=umbani I=m I=mb I=mba I=mban S=mbani I=b I=ba I=ban S=bani I=a I=an S=ani I=n S=ni S=i
PRON	P=k P=kw P=kwa P=kwak **W=kwake** I=w I=wa I=wak S=wake I=a I=ak S=ake I=k S=ke S=e
NEG	P=h P=ha P=han **W=hana** I=a I=an S=ana I=n S=na S=a
N	P=f P=fa P=fah P=faha P=faham **W=fahamu** I=a I=ah I=aha I=aham S=ahamu I=h I=ha I=ham S=hamu I=a I=am S=amu I=m S=mu S=u
FULL-STOP	**W=.**

Figure 1: Bag-of-Substrings representation of Example 2.

the words are handled by one and the same orthography-based classifier (in contrast to MBT). While this classifier is basically a unigram classifier (cf. Baseline in Table 2) it draws its predictive power from the Bag-of-Substrings (henceforth BoS), presented as training material.

The BoS lines in Table 2 display the experiment results. For Swahili known words, the BoS-approach underperforms, compared to MBT and even the baseline tagger. For unknown words, on the other hand, tagging accuracy is significantly higher using a BoS-approach[3]. The same trend is visible for the Northern Sotho data. For Zulu, baseline accuracy is restored and a huge improvement is achieved in the handling of unknown words, compared to MBT. For the Cilubà data, the BoS approach does not significantly underperform for known words, while it substantially improves for unknown words.

Given the modular design of MBT, we can now envision a mixed tagging approach, using different, individual classifiers for known and unknown words. The results of this virtual experiment are displayed on the last line of Table 2. For Zulu, there is no advantage of using a mixed approach, but for Swahili, Northern Sotho and Cilubà, the optimal combination involves tagging known words with MBT and unknown words with BoS, yielding higher scores than the individual taggers. In practice, tagging accuracy will be even higher for such a combination, since more accurate unknown words tagging will lead to more accurate left-context information for the prediction of known words.

It is important to point out that the BoS approach does not necessarily need a tagged corpus as training material: a simple lexicon in which each word is associated with a part-of-speech tag, is all the training data this approach needs. This is good news, since state-of-the-art tagging accuracies can now be achieved without the development of a manually part-of-speech tagged corpus, unless of course, we are dealing with a language with a disjunctive orthography. As the experimental results for Northern Sotho in Table 2 show, contextual information is essential in such a case and this can only be induced from a tagged corpus. An additional advantage of the BoS approach is its efficiency: Swahili is tagged by MBT at a rate of 4,400 words per second, whereas BoS can tag words at 17,000 words per second.

5. Learning Curves

Table 2 shows that context-driven taggers still have the edge when trained on expansive data sets, whereas the situation is somewhat reversed for the resource-scarce languages of Zulu and Cilubà. In this section, we will perform a direct comparison between the two approaches, using steadily increasing amounts of data. We re-use the slices of the ten-fold cross validation experiments for this. The last slice is kept constant as the evaluation set throughout the experiments. In the first experiment Slice0 (10% of the data) is used as training material. In the next experiment Slice0+Slice1 (20% of the data) is used, etc.

This results in learning curves (Figure 2), which show how the two approaches compare to one another for different data set sizes. For Swahili and Northern Sotho, the learning curves confirm the results of the ten-fold cross validation experiment: MBT consistently outperforms BoS for known words, even for smaller data set sizes, while BoS has the edge for unknown words. For Northern Sotho, a peculiar situation arises in the middle of the experiment: BoS unknown word accuracy is actually higher than BoS known words accuracy. This proves that BoS is not suitable for tagging a language with a disjunctive writing system, even though as an unknown words predictor, it works better than a context-driven tagger does.

For both Zulu and Cilubà, performance for known words is more or less equal for the two techniques. The difference is made in the handling of unknown words. Increasing corpus size for Zulu does not allow MBT to make significantly more accurate predictions and at some points, the addition of new material appears to actually confuse the classifier. BoS is more stable in this respect, as accuracy for tagging Zulu unknown words steadily increases with data set size. For Cilubà BoS tagging of unknown words, on the other

[3] The McNemar Significance test for paired classifiers (McNemar, 1947) was used to establish statistical significance throughout this paper.

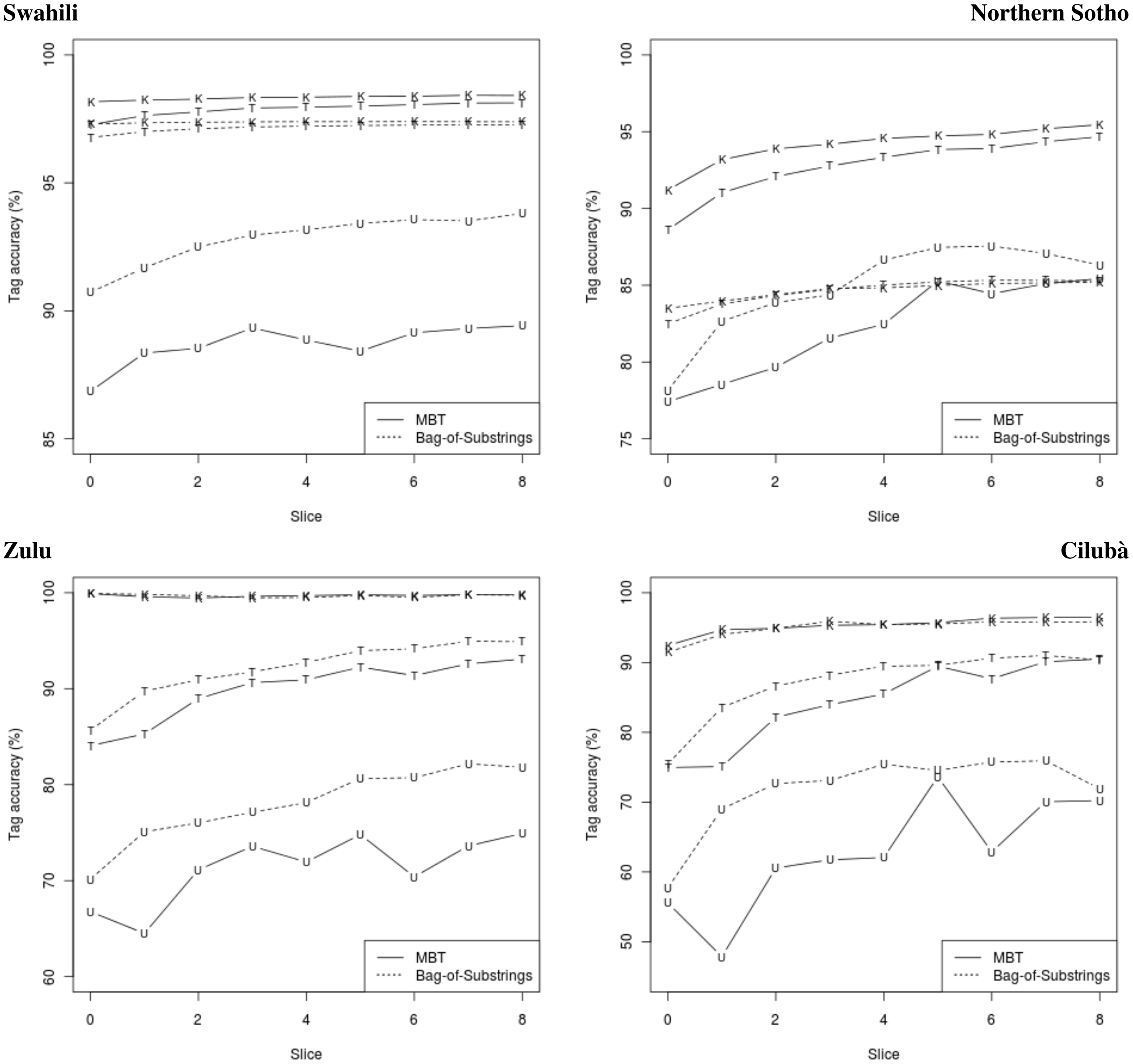

Figure 2: Graphs for learning curve experiments. Learning curves are displayed for (K)nown words, (U)known words and the (T)otal tagging accuracy.

hand, the learning curve seems to plateau and even drop at the end. The erratic curve for MBT however indicates that the small size of the corpus makes the results vulnerable to small distributional changes within the slices and it is not possible to draw any reliable conclusions from this experiment for Cilubà.

6. Conclusion and Future Work

This paper adds to the growing number of publications that describe data-driven approaches to natural language processing of African languages. We described experiments with an off-the-shelf, data-driven tagger and observed reasonable to excellent tagging accuracies for Swahili, Northern Sotho and Zulu. The underwhelming results for Cilubà can be attributed to the diminutive nature of the data set used for training.

The Bag-of-Substrings technique, introduced in this paper as a part-of-speech tagging approach, has been empirically shown to be able to hold its own against a context-driven tagger, provided there is a critical amount of data available. For a language with a disjunctive orthography, the BOS approach is only really useful for the prediction of tags for previously unseen words. The learning curves provided further evidence that the BOS approach establishes a fast, resource-light technique for part-of-speech tagging of agglutinative languages.

In future research efforts, we will also investigate other data-driven taggers. Preliminary experiments with TnT (Brants, 2000) and SVMTool (Giménez and Màrquez, 2004) exhibited the same trends as MBT using the same features, albeit with the former significantly and surprisingly underperforming compared to MBT. This unexpected result in itself warrants further research, which may provide some insight into best practices for Bantu part-of-speech tagging. Since the BOS approach is de facto a data-driven tagger, this experiment can be easily replicated for other languages and data sets. Future research will investigate whether the technique can prove valuable for other Bantu languages and other language groups as well. We will also attempt to introduce contextual features to the maxent classifier that underlies the BOS method, which may serve to get the best of both worlds: accurate unknown word POS-tagging, coupled with context-aware disambiguation of known tokens. Finally, we also aim to further explore the possibility of using lexicons, rather than the much less readily available annotated corpora, as training material for the BOS approach to bootstrap part-of-speech taggers for a wide range of resource-scarce languages.

Acknowledgments and Demos

The first author is funded as a Postdoctoral Fellow of the Research Foundation - Flanders (FWO). We would like to thank Sebastian Dom for his Cilubà annotation efforts and the reviewers for their insightful comments.

Demonstration systems can be found at AfLaT.org for part-of-speech tagging of
Swahili (`http://aflat.org/swatag`),
Northern Sotho (`http://aflat.org/sothotag`),
Zulu (`http://aflat.org/zulutag`) and
Cilubà (`http://aflat.org/lubatag`).

7. References

Brants, T. (2000). TnT a statistical part-of-speech tagger. In *Proceedings of the Sixth Conference on Applied Natural Language Processing (ANLP 2000).* Seattle, USA: pp. 224–231.

Daelemans, W., Zavrel, J., van den Bosch, A. & Van der Sloot, K. (2010). MBT: Memory-based tagger, version 3.2, reference guide. Technical Report 10-04, University of Tilburg.

De Pauw, G., de Schryver, G-M & Wagacha, P.W. (2006). Data-driven part-of-speech tagging of Kiswahili. In P. Sojka, I. Kopeček & K. Pala (Eds.), *Proceedings of Text, Speech and Dialogue, Ninth International Conference*. volume 4188/2006 of *Lecture Notes in Computer Science*, Berlin, Germany: Springer Verlag, pp. 197–204.

De Pauw, G., Maajabu, N.J.A. & Wagacha, P.W. (2010). A knowledge-light approach to Luo machine translation and part-of-speech tagging. In G. De Pauw, H. Groenewald & G-M de Schryver (Eds.), *Proceedings of the Second Workshop on African Language Technology (AfLaT 2010).* Valletta, Malta: European Language Resources Association (ELRA), pp. 15–20.

De Pauw, G. & Wagacha, P.W. (2007). Bootstrapping morphological analysis of Gĩkũyũ using unsupervised maximum entropy learning. In *Conference Program and Abstract Book of the Eighth Annual Conference of the International Speech Communication Association*. Antwerp, Belgium, 29 August: ISCA, p. 119.

De Pauw, G., Wagacha, P.W. & Abade, D.A. (2007). Unsupervised induction of Dholuo word classes using maximum entropy learning. In K. Getao & E. Omwenga (Eds.), *Proceedings of the First International Computer Science and ICT Conference*. Nairobi, Kenya: University of Nairobi, pp. 139–143.

De Pauw, G., Wagacha, P.W. & de Schryver, G-M. (2011). Exploring the SAWA corpus - collection and deployment of a parallel corpus English - Swahili. *Language Resources and Evaluation - Special Issue on African Language Technology*, 45(3), pp. 331–344.

de Schryver, G-M & De Pauw, G. (2007). Dictionary writing system (DWS) + corpus query package (CQP): The case of TshwaneLex. *Lexikos*, 17, pp. 226–246.

Dione, C.M.B., Kuhn, J. & Zarrie, S. (2010). Design and development of part-of-speech-tagging resources for Wolof (Niger-Congo, spoken in Senegal). In N. Calzolari, Kh. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner & D. Tapias (Eds.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10).* Valletta, Malta: European Language Resources Association (ELRA).

Gambäck, B., Olsson, F., Argaw, A.A. & Asker, L. (2009). Methods for Amharic part-of-speech tagging. In G. De Pauw, G-M de Schryver & L. Levin (Eds.), *Proceedings of the First Workshop on Language Technologies for African Languages (AfLaT 2009).* Athens, Greece: Association for Computational Linguistics, pp. 104–111.

Giménez, J. & Màrquez, L. (2004). A general POS tagger generator based on support vector machines. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004).* Lisbon, Portugal: pp. 43–46.

Hahn, S., Lehnen, P., Raymond, C. & Ney, H. (2008). A comparison of various methods for concept tagging for spoken language understanding. In N. Calzolari, Kh. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis & D. Tapias (Eds.), *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08).* Marrakech, Morocco: European Language Resources Association (ELRA).

Hoogeveen, D. & De Pauw, G. (2011). Corpuscollie - a web corpus mining tool for resource-scarce languages. In *Proceedings of Conference on Human Language Technology for Development*. Alexandria, Egypt: Bibliotheca Alexandrina, pp. 44–49.

Hurskainen, A. (2004a). HCS 2004 – Helsinki Corpus of Swahili. Technical report, Compilers: Institute for Asian and African Studies (University of Helsinki) and CSC.

Hurskainen, A. (2004b). Swahili language manager: A storehouse for developing multiple computational applications. *Nordic Journal of African Studies*, pp. 363–397.

Le, Z. (2004). Maximum entropy modeling toolkit for python and c++. Technical report, Available at: http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html. (Accessed: 2 March 2012).

Marcus, M., Santorini, B. & Marcinkiewicz, M. (1993). Building a large annotated corpus of English: The Penn treebank. *Computational Linguistics*, 19(2), pp. 313–330.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2), pp. 153–157.

Prinsloo, D.J. & de Schryver, G-M. (2002). Towards an 11 x 11 array for the degree of conjunctivism / disjunctivism of the South African languages. *Nordic Journal of African Studies*, 11(2), pp. 249–265.

Ramshaw, L.A. & Marcus, M.P. (1995). Text chunking using transformation-based learning. In D. Yarowsky & K. Church (Eds.), *Proceedings of the Third ACL Workshop on Very Large Corpora*. Cambridge, USA: Association for Computational Linguistics, pp. 82–94.

Ratnaparkhi, A. (1996). A maximum entropy model for part-of-speech tagging. In E. Brill & K. Church (Eds.), *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Philadelphia, USA: Association for Computational Linguistics, pp. 133–142.

Spiegler, S., van der Spuy, A. & Flach, P.A. (2010). Ukwabelana - an open-source morphological Zulu corpus. In Q. Lu & T. Zhao (Eds.), *Proceedings of the 23rd International Conference on Computational Linguistics*. Beijing, China: Tsinghua University Press, pp. 1020–1028.

Veenstra, J., van den Bosch, A., Buchholz, S., Daelemans, W. & Zavrel, J. (1999). Memory-based word sense disambiguation. In F. Van Eynde (Ed.), *Computational linguistics in the Netherlands 1998*, pp. 81–92: Rodopi, Amsterdam.

POS Annotated 50M Corpus of Tajik Language

Gulshan Dovudov, Vít Suchomel, Pavel Šmerk

Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
zarif_dovudov@mail.ru,xsuchom2@fi.muni.cz,smerk@mail.muni.cz

Abstract

Paper presents by far the largest available computer corpus of Tajik language of the size of more than 50 million words. To obtain the texts for the corpus two different approaches were used and the paper offers a description of both of them. Then the paper describes a newly developed morphological analyzer of Tajik and presents some statistics of its application on the corpus.

Keywords: Tajik language, Tajik corpus, morphological analysis of Tajik

1. Introduction

1.1. Tajik Language

The Tajik language is a variant of the Persian language spoken mainly in Tajikistan, where it plays a role of the national and official language. Tajik is spoken also in some few other parts of Central Asia, among which neighboring Uzbekistan is the most notable, because the biggest group of Tajik native speakers outside Tajikistan resides there.

Unlike closely related and mutually intelligible Iranian Persian (Farsi) and Afghan Persian (Dari), which are written in the Arabic script, Tajik is written mostly in the Cyrillic alphabet which is the official script.

According to its grammatical structure the Tajik language inflectionally belongs to analytical type of languages. Although the nominal morphology itself is rather poor—Tajik has neither gender nor case, only pluralization suffixes and specificity/undefiniteness marker—nouns, adjectives and participles can appear in many different forms thanks to the direct object marker, possesive enclitics (*dust-am: friend your*), copulas (*dust-am: friend [I] am*) and ezafe marker which are all written together with the word in the Cyrillic script and some of them even can combine together. Tajik verbal morphology is much richer: besides many compound verbal forms Tajik has affixes (mostly suffixes) for expressing person, number, tense, mood and voice. Together with participles, the number of distinct forms which can be generated from a single verbal root can easily exceed 1000 items.

1.2. Existing Tools and Resources

Since the Tajik language internet society (and consequently the potential market) is rather small and Tajikistan itself is ranked among developing countries, available tools and resources for computational processing of Tajik as well as publications in the field are rather scarce. Moreover, many of the publications were published only in Russian.

For Tajik there exist several online[1] or offline Tajik–Russian and Tajik–English dictionaries, from which the biggest (Usmanov et al., 2007) and (Usmanov et al., 2008) covers ca. 120,000 Russian words and phrases. The corresponding numbers for English are rather lower, around 35,000 words and phrases at most.

Megerdoomian and Parvaz (2008; 2009) aim to transliterate Tajik from Cyrillic alphabet to Arabic script, which would allow to employ tools developed for closely related Iranian Persian. More elaborated transliteration was offered by Usmanov et al. (2009). Usmanov et al. also created a morphological analyzer (2011) and a spellchecker (2012) for OpenOffice.org suite. Another spellchecker was made by Davrondjon and Janowski (2002). There is also available a Tajik text-to-speech system built by Khudoiberdiev (2009). If we add Tajik extension of the multilingual information extraction system ZENON (Hecking and Sarmina-Baneviciene, 2010), the list of all more interesting Tajik language processing tools and non-corpora resources we are aware of is complete.

The computer corpora of Tajik language are either small or even still only in the stage of planning or development. The University of Helsinki offers a very small corpus of 87 654 words.[2] Megerdoomian and Parvaz (2008; 2009) mention a test corpus of approximately 500 000 words, and the biggest and the only freely available corpus is offered within the Leipzig Corpora Collection project (Quasthoff et al., 2006) and consists of 100 000 Tajik sentences which equals to almost 1.8 million words.[3] Iranian linguist Hamid Hassani is said to be preparing a 1 million words Tajik corpus[4] and Tajik Academy of Sciences prepares a corpus of 10 million words[5]. Unfortunately, at least by now the latter is not a corpus of contemporary Tajik, but rather a collection of works—and moreover mainly a poetry—of a few

[1] E.g. http://lugat.tj, http://sahifa.tj, http://termcon.tj or http://www.slovar.kob.tj.

[2] http://www.ling.helsinki.fi/uhlcs/readme-all/README-indo-european-lgs.html

[3] Unfortunately, the encoding and/or transliteration vary greatly: more than 5 % of sentences are in Latin script, almost 10 % of sentences seem to use Russian characters instead of Tajik specific characters (e.g. х instead of Tajik ҳ, which sound/letter does not exist in Russian) and more than 1 % of sentences uses non-Russian substitutes for Tajik specific characters (e.g. Belarussian ў instead of proper Tajik ӯ) — and only the last case is easy to repair automatically.

[4] http://en.wikipedia.org/wiki/Hamid_Hassani, http://www.tajikistan.orexca.com/tajik_language.shtml

[5] http://www.termcom.tj/index.php?menu=bases&page=index3&lang=eng (in Russian)

notable Tajik writers (one of them is even from the 13th century).

1.3. Structure of the Paper

In this paper we present a newly built corpus of the contemporary Tajik language of more than 50 million words. All texts were taken from the internet. We used two different approaches to obtain the data for the corpus and we describe these methods and their results in the following two sections. In the Section 3 we present a new morphological analyzer of Tajik and the results of annotating the corpus with this analyzer. Finally, in the last section we discuss some planned future improvements.

2. Bulding the Corpus

As it was said in the Introduction, the new[6] corpus consists of two parts. The first one is supposed to contain data of a higher quality, the second one contains texts which we are able to download and process only in a less controlled way.

2.1. Semi-automatically Crawled Part

The first part of the corpus was collected by crawling several news portals in Tajik language.[7] If the articles of a particular portal were numbered, we tried to download all possible numbers, otherwise we got a list of articles either directly from the portal, or from the Google cache. Each single portal was processed separately to get the maximum of relevant (meta)information, i.e. correct headings, publication date, rubric etc.

In the Table 1 we present some statistics of the obtained data. Docs is a number of documents downloaded from the given source, pars is a number of paragraphs (including headings), words is a number of tokens which contain only characters from Tajik alphabet, w/doc is a words / document ratio (i.e. average length of possibly continuous texts), tokens is a number of all tokens (words, interpunction etc.) and MB is the size in megabytes of the data in vertical corpus format (i.e. plain text). The table is sorted by number of words. From the electronic library on `gazeta.tj` we choose only prose and omit all more structured texts as poetry, drama or e.g. computer manual. The articles in `gazeta.tj` archive are joined in one file on a weekly basis and that is why the words / document ratio is so high.

On almost all websites, alongside the articles in Tajik there were also many articles in Russian. Because both alphabets, Tajik and Russian, contain characters which do not occur in the other alphabet, it is easy to distinguish between the two languages and discard the Russian articles even without any further language analysis.

2.2. Automatically Crawled Part

SpiderLing, a web crawler for text corpora (Suchomel and Pomikálek, 2011), was used to automatically download documents in Tajik from the web. We started the crawl using 2570 seed URLs (from 475 distinct domains) collected with Corpus Factory (Kilgarriff et al., 2010). The crawler combines character encoding detection[8], general language detection based on a trigram language model trained on Wikipedia articles, and a heuristic based boilerplate removal tool jusText[9] which removes content such as navigation links, advertisements, headers and footers etc. and preserves only paragraphs (preferrably continuous groups of paragraphs) containing full sentences.

The crawler downloaded 9.5 GB of HTML data in ca. 300,000 documents over three days. That is not much compared to crawling documents in other languages by the same tool. For example the newly built web corpus of Czech, which has roughly only two times more native speakers compared to Tajik, has more than 5 billion words, of course, not obtained in three days. We conclude that the available online resources in Tajik are very scarce indeed. An overview of internet top level domains of URLs of the documents obtained can be found in Table 2.

TLD	docs downloaded	docs accepted
tj	55.0 %	51.7 %
com	23.0 %	28.1 %
uk	8.9 %	7.2 %
org	6.8 %	7.7 %
ru	2.6 %	1.4 %
ir	1.6 %	2.4 %
other	2.0 %	1.5 %

Table 2: Number of documents by internet top level domain

Since Russian is widely used in government and business in Tajikistan (and other language texts may appear), 33 % of the downloaded HTML pages were removed by the SpiderLing's inbuilt language filter. The obtained data was tokenized and deduplicated using Onion[10] with moderately strict settings[11]. Some statistics of the automatically crawled part of the corpus are in the Table 3 (only the ten most productive sources of data are detailed).

2.3. Corpus of Tajik Language

The two partial corpora were joined together and the result was deduplicated using the tool Onion. We obtained a corpus of more than 50 million words, i.e. corpus positions which consists solely of Tajik characters, and more than 60 million tokens, i.e. words, interpunction, numbers etc. Detailed numbers follow in the Table 4.

[6] It should be noted that this is not the very first information about the new corpus. We presented it at a local event few months ago (Dovudov et al., 2011). Since that time, as the following numbers and text show, we extended the first part of the corpus and repaired some errors in the tools for automatic crawling and then processed the previously crawled data again, which positively affected the second part of the corpus and among others it enabled its much better deduplication.

[7] Paradoxically, the two largest Tajik news portals are not located in Tajikistan, but in Czech Republic (ozodi.org, Tajik version of Radio Free Europe/Radio Liberty) and United Kingdom (bbc.co.uk, Tajik version of BBC).

[8] `http://code.google.com/p/chared/`

[9] `http://code.google.com/p/justext/`

[10] `http://code.google.com/p/onion/`

[11] removing paragraphs with more than 50 % of duplicate 7-tuples of words

source	docs	pars	words	w/doc	tokens	MB
ozodi.org	58228	348583	12597156	216	14738546	178
gazeta.tj archive	480	163565	5006469	10430	6032000	67
co.uk	9240	164252	4124668	446	4767707	57
jumhuriyat.tj	8104	104331	3703806	457	4397596	53
tojnews.org	9598	72407	2529370	264	3073951	37
khovar.tj	16860	67325	2502295	148	3055984	39
millat.tj	2540	47144	2131354	839	2511507	29
gazeta.tj	1940	37460	1352150	697	1622370	18
gazeta.tj library	130	98690	1053206	8102	1358510	15
ruzgor.tj	1753	26417	1046598	597	1241107	14
			...			
all	**115364**	**1200645**	**38269686**	**332**	**45473773**	**537**

Table 1: Statistics of the semi-automatically crawled part of the corpus.

source	docs	pars	words	w/doc	tokens	MB
*.wordpress.com	3298	85923	3683551	1117	4489937	50
bbc.co.uk	5194	94441	2480519	478	2857831	34
ozodi.org	5930	81147	2017199	340	2399314	28
khovar.tj	10345	41604	1599389	155	1953702	25
millat.tj	1349	21274	1080308	801	1268268	14
gazeta.tj	1704	27631	1040471	611	1244311	14
*.blogspot.com	793	21849	854055	1077	1060820	12
firdavsi.com	559	24315	804196	1439	966421	11
pressa.tj	2317	17130	637781	275	771169	9
ruzgor.tj	889	10408	489143	550	573800	7
			...			
all	**53424**	**675366**	**24723099**	**463**	**29709116**	**346**

Table 3: Statistics of the automatically crawled part of the corpus.

It was rather surprising for us that the fully automated crawling yielded even smaller data than the semi-automated approach. It has to be said that at least 25 % of semi-automatically crawled data were inaccessible to the general crawler, as it cannot extract texts from RAR-compressed archives (gazeta.tj archive and library) and because there does not seem to exist any link to the bigger part of older BBC articles although they remained on the server (we exploited Google cache to get the links). It is highly probable that also the other sites contain articles unreachable by any link chain and thus inaccessible for the general crawler. But even if we discount these data, the automated crawling did not outperform the semi-automated one in such an extent that we expected and which is common for many other languages. As we remarked in the previous subsection, we attribute it to the scarceness of the online texts in Tajik language. It means that we probably reach or almost reach the overall potential of internet resources, i.e. even if we somehow get all Tajik online texts, the corpus might be bigger by half, but surely not, for example, ten times or even three times.

3. Morphological Analysis of Tajik and Annotation of the Corpus

As it was mentioned in the Introduction, a morphological analyzer for the Tajik language already exists (Usmanov et al., 2011). Unfortunately, for the annotation of corpus data it showed up quite unusable, mainly for the following reasons:

- the program is too slow, it processes less than 2 words per second;
- its code is written in MS Visual Basic 6.0 and the executable can run only on MS Windows, but we process corpus data on Linux servers (and may be there are ways to compile such MS VB code under Linux, but we consider it too complicated);
- the program splits the input word form to morphs (among others), but it cannot offer the lemma (base form, citation form) of the word.

3.1. New Morphological Analyzer

For the creation of the new morphological analyzer we wanted to use information about Tajik morphs and their possible combinations from the current analyzer. It was not a simple task, because the information on morph combining is not described in some external file but “encoded” directly in the source code, so it is still a work in progress.

We use an approach of Jan Daciuk (Daciuk, 1998), who invented an algorithm for both space and time efficient building of minimal deterministic acyclic finite state automata (DAFSA). On his pages he offers source codes of tools

source	docs	pars	words	w/doc	tokens	MB
ozodi.org	59943	384932	13426445	224	15738683	189
gazeta.tj archive	480	163555	5006432	10430	6031951	67
co.uk	9288	164436	4129179	445	4772807	57
jumhuriyat.tj	8106	104776	3703685	457	4397650	53
*.wordpress.com	3080	74652	3235436	1050	3946319	44
tojnews.org	9653	72575	2532572	262	3077917	37
khovar.tj	17079	68022	2512232	147	3082293	39
millat.tj	2803	49846	2268000	809	2673004	30
gazeta.tj	2209	38060	1389318	629	1665672	19
kemyaesaadat.com	1863	33859	1182353	635	1404072	16
			...			
all	**138701**	**1541470**	**51722009**	**373**	**61837585**	**723**

Table 4: Statistics of the resulting corpus.

which allows to convert morphological data into such an automaton and to use it for morphological analysis.

The data for the analyzer are triplets *word:lemma:tag*, e.g. `kardem:kardan:tag` where the lemma is, in fact, not present in a full form, but encoded in the following way `kardem:Can:tag`, which means "to obtain a lemma, delete 2 (`A` = 0, `B` = 1, `C` = 2, ...) letters from the end of the wordform and add the string `an`".[12] The list of such triplets can be viewed as a finite formal language and for such languages always exists the minimal DAFSA. If we generate such triplets for all word forms the analyzer should know, the data will be quite big, but also very redundant: the conversion to an automaton serves as a very good compression.

The data for the analyzer are generated from our dictionary of Tajik base forms (lemmata), where each lemma has an information about its POS and optionally some additional information (e.g. that pluralization suffix can be not only common *-ho*, but also *-on*). The possible word forms are generated according to a rather simple description (80 lines) of possible suffixes and their allowed combinations. The process also uses an information about possible phonological or ortographical changes at morpheme boundaries.

The whole analysis is then just effortless travelling through the automaton: the analyzer simply (the automaton is deterministic) follows the path which corresponds to the analyzed word form and the delimiter, e.g. `kardem:`. Then the labels of each possible path to the final state represents one of the possible analyses. The new analyzer is therefore very simple: the whole source code has only around 450 lines in C++ (we have simplified Daciuk's code considerably). The new analyzer is also much faster then the current one: at the moment it processes around million words per second.[13] This number will get lower in the future as the size of the data will increase, but surely it always will remain very fast.

count of word+lemma+tag triplets	8,476,108
size of input data in bytes	175,845,264
size of automaton in bytes	1,138,480
bytes per line of input data	0.13
count of lemmata in dictionary	14934
average number of triplets per lemma	568

Table 5: Some statistics of the new analyzer data.

Because many words of our corpus were unknown to the current analyzer, we needed to enrich the lexicon. We took all unknown word forms from the corpus and for each such word form we generated possible lemmata. Then we manually evaluated such lemma candidates from the most frequent ones. In principle, our approach was very similar to (Sagot, 2005), but we avoided all the maths which Sagot uses to rank lemma candidates: the manual decision process was so fast that there was no need for some tiny optimization. But unlike Sagot, we did not use whole morphology at once, but we started with searching for possible proper nouns, then common nouns, then adjectives, which can express degree etc.—it simplifies the work of the annotator, because deciding only e.g. proper nouns is faster and less erroneous than deciding possible lemmata of all kinds in one pass. Before searching for lemma candidates we also tried to detect and lowercase words which started with capital letter yet not being a proper noun: we selected words whose lowercase form was more frequent than the form with the first letter capitalized. We obtained more than 7000 lemmata and almost 5000 of them were proper nouns. Of course, it is an ongoing process, we have evaluated only the most frequent candidates so far.

3.2. Annotation of the Corpus

We decided to use only a lemma and POS for the annotation of the corpus data. The information which is represented by the rest of the morphological tag is currently not in a fully consistent state[14] and also the current format is a subject to change, because the current analyzer uses tags which are too long and thus hard to read. The Table 6 shows the meaning of tags and counts of word forms for each POS.

[12] This example would work only for suffixes. To handle also the prefixes, it is possible to employ the same principle again, for example: `namekardem:ECan:tag`, where the first `E` denotes that to get the correct lemma kardan the first four letters are to be deleted (and nothing is to be added).

[13] The speed of analyzers was not compared on a same computer, but the difference is obvious.

[14] At least in our new analyzer—and it is not possible to directly use the "knowledge" of the current analyzer, as we have mentioned it in the previous subsection.

Meaning	Tag	# of forms
nouns	01	6267182
adjectives	02	941209
numerals	03	25572
pronouns	04	52
verbs	05	372778
infinitives	06	646500
adjectival participles	07	217273
adverbial participles	08	5253
adverbs	09	86
prepositions	10	44
postpositions	11	3
conjunctions	12	52
particles	13	35
interjections	14	64
onomatopoeia	15	0
numeratives	16	5

Table 6: Meaning of the tags and numbers of forms with the given tag in the data.

Our new analyzer is able to annotate 87.2 % of the 51,722,009 words from the corpus[15] (and 20.5 % of the wordlist). 25.6 % of known words are ambiguous (13.9 % of the wordlist) and for known words the analyzer offers 1.33 lemma+POS combinations in average. As this is the very first work in this field, we cannot offer any comparison with existing results. We also do not have any manually tagged data and thus we cannot calculate the standard measures like coverage, accuracy or F-measure.

dar	dar:01;dar:05;dar:10	1626855
ba	ba:10	1572867
va	va:12	1417227
ki	ki:04;ki:12	1226404
az	az:10	1173474
in	in:04;in:14	773985
bo	bo:10	513154
ast	ast:05	347578
on	on:04	301493
Tojikiston	Tojikiston:01	281627

Table 7: The top ten most frequent words, their analyses and frequency in corpus.

The annotated corpus is not freely available for a download at the moment, but eventual interested researchers can access it through the Sketch Engine on `http://ske.fi.muni.cz/open/`. This web interface displays concordances from the corpus for a given query. The program is very powerful with a wide variety of query types and many different ways of displaying and organising the results.

4. Future Work

The Table 8 shows statistics of the texts which were new in the automatically crawled part compared to the semi-automatically crawled data. It is worth mentioning the difference between the two parts of the corpus: the analyzer knows 89.1 % of words from the first part, but only 81.8 % words from the second part. Our interpretation is that data in the first part display a higher quality, higher regularity and we expect them to be less noisy than the data from the second part of the corpus. The numbers in the table indicate that there is still some room for an extension of the semi-automated part. We will prepare specialized scripts for the most productive portals to download their data in a some more controlled way. We would like to extend the first part of the corpus to at least 50 million words.

It is worth clarifying the cases of ozodi.org and kemyaesaadat.com, as data from these sites were downloaded also semi-automatically. The general crawler tries to get all reasonable texts on the page, which, on the news portals, may include the readers' comments. On the other hand, because the comments may contain a substandard language features, they were omitted during the semi-automated crawling. Thus the 1715 documents from ozodi.org and 346 from kemyaesaadat.com were not some newly added ones, but they were results of the deduplication which discarded the article itself and leaved only the comments as it processed the corpus by single paragraphs. This is also one of the reasons why we prefer the semi-automated crawling when it is possible: in the future we want to mark these comments to allow a creation of a subcorpus of the (presumably standard) language of articles as well as a subcorpus of the (potentially substandard) language of comments.

Another problem with the comments—but not only with them—is a common absence of Tajik-specific characters. The language model for the general crawler was trained using Tajik Wikipedia[16] so the crawler searches for texts in language, which looks like the language of Tajik Wikipedia. Unfortunately, in many Wikipedia articles the Tajik-specific characters are replaced by some other characters. The unambiguous replacements were trivially repaired in the whole corpus, but e.g. Cyrillic х can sometimes stand either for the Tajik-specific ҳ or also for х itself. On the one hand we plan to tag such texts to allow a creation of subcorpora with or without them, on the other hand we want to either develop a program which would be able to repair them or use (Usmanov and Evazov, 2011). We will also train the language model with another sets of texts to see how it will affect the crawled data.

Unlike the corpus—which is still "work in progress", but the progress is already rather moderated—the work on the analyzer and the morphological data is still at the beginning. It is neccessary to properly describe the "morphological" system of Tajik, design the tagset, enlarge the lexicon (preferably exploiting word derivation) and start working on at least partial disambiguation.

[15] Word is here a token which contain only characters from Tajik alphabet, because analyzer cannot analyze anything else. More than 10,000,000 of corpus tokens are numbers, interpunction etc., but also e.g. words in Latin alphabet.

[16] The use of Wikipedia to train the language model is a part of default settings or a default scenario of the process of building corpora for new languages without any other utilizable resources. Of course we have better Tajik texts at hand, but the automatically crawled part of our corpus had also to act as a test of a general suitability of our technologies for the case of building corpora for low-density languages.

source	docs	pars	words	w/doc	tokens	MB
*.wordpress.com	3080	74652	3235436	1050	3946319	44
ozodi.org	1715	36386	829816	484	1000827	11
*.blogspot.com	723	17901	690797	955	865401	10
firdavsi.com	428	19861	614062	1435	737959	8
abdulov.tj	99	9392	403686	4078	488243	6
nahzat.tj	2406	7096	394437	164	463538	6
sahifa.tj	56	478	390618	6975	480398	5
ozodagon.com	1425	6352	371081	260	446249	5
kemyaesaadat.com	346	8526	342394	990	411893	4
bayynattj.com	298	4810	293010	983	340423	4
			...			
all	**23337**	**340950**	**13454032**	**577**	**16365975**	**186**

Table 8: The contribution of automatically crawled part.

Acknowledgements

This work has been partly supported by Erasmus Mundus Action II lot 9: Partnerships with Third Country higher education institutions and scholarships for mobility, and by the Ministry of Education of CR within the Lindat Clarin Center.

5. References

Jan Daciuk. 1998. *Incremental Construction of Finite-State Automata and Transducers, and their Use in the Natural Language Processing*. Ph.D. thesis, Technical University of Gdańsk, Gdańsk.

Gafurov Davrondjon and Tomasz Janowski. 2002. Developing a Spell-Checker for Tajik using RAISE. In *Proceedings of the 4th International Conference on Formal Engineering Methods: Formal Methods and Software Engineering*. Springer Verlag.

Gulshan Dovudov, Jan Pomikálek, Vít Suchomel, and Pavel Šmerk. 2011. Building a 50M Corpus of Tajik Language. In *Proceedings of the Fifth Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 2011*, Brno. Masaryk University.

Leonid A. Grashenko, Zafar D. Usmanov, and Aleksey Y. Fomin. 2009. Tajik-Persian converter of the graphic writing systems. National patent 091TJ, National Patent Information Centre, Republic of Tajikistan. (In Russian).

Matthias Hecking and Tatiana Sarmina-Baneviciene. 2010. A Tajik Extension of the Multilingual Information Extraction System ZENON. In *Proceedings of the 15th International Command and Control Research and Technolgy Symposium (ICCRTS)*, Santa Monica, CA.

Khurshed A. Khudoiberdiev. 2009. *Complex Program of Tajik Text-to-Speech Synthesis*. Ph.D. thesis, Khujand Polytechnical Institute of Tajik Technical University, Dushanbe. (In Russian).

Adam Kilgarriff, Siva Reddy, Jan Pomikálek, and Avinesh PVS. 2010. A Corpus Factory for Many Languages. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC 2010*, Valleta, Malta.

Karine Megerdoomian and Dan Parvaz. 2008. Low-density Language Bootstrapping: The Case of Tajiki Persian. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation, LREC 2008*, Marrakech, Morocco.

Karine Megerdoomian, 2009. *Language Engineering for Lesser-Studied Languages*, chapter Low-density Language Strategies for Persian and Armenian, pages 291–312. IOS Press, Amsterdam.

Uwe Quasthoff, Matthias Richter, and Christian Biemann. 2006. Corpus Portal for Search in Monolingual Corpora. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC 2006*, Genoa.

Benoît Sagot. 2005. Automatic Acquisition of a Slovak Lexicon from a Raw Corpus. In *Text, Speech and Dialogue*, volume 3658 of *Lecture Notes in Computer Science*, pages 156–163. Springer.

Vít Suchomel and Jan Pomikálek. 2011. Practical Web Crawling for Text Corpora. In *Proceedings of the Fifth Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 2011*, Brno. Masaryk University.

Zafar D. Usmanov and Khisrav A. Evazov. 2011. Computer correction of Tajik text typed without using special characters. *Reports of Academy of Sciences, Republic of Tajikistan*, 54(1):23–26. (In Russian).

Zafar D. Usmanov, Sanovbar D. Kholmatova, and Odilchodja M. Soliev. 2007. Tajik–Russian computer dictionary. National patent 025TJ, National Patent Information Centre, Republic of Tajikistan. (In Russian).

Zafar D. Usmanov, Odilchodja M. Soliev, and Khurshed A. Khudoiberdiev. 2008. Russian–Tajik computer dictionary. National patent 054TJ, National Patent Information Centre, Republic of Tajikistan. (In Russian).

Zafar D. Usmanov, Gulshan M. Dovudov, and Odilkhodja M. Soliev. 2011. Tajik computer morfoanalyzer. National patent ZI-03.2.220, National Patent Information Centre, Republic of Tajikistan. (In Russian).

Zafar D. Usmanov, Odilchodja M. Soliev, and Gulshan M. Dovudov. 2012. Tajik language package for system OpenOffice.Org. National patent ZI-03.2.222, National Patent Information Centre, Republic of Tajikistan. (In Russian).