AfLaT 2010

Proceedings of the Second Workshop on African Language Technology

Held in collocation with the Seventh International Conference on Language Resources and Evaluation (LREC 2010)

May 18, 2010
Valletta, Malta

Edited by

Guy De Pauw
Handré Groenewald
Gilles-Maurice de Schryver

LREC 2010

Proceedings of the Second Workshop on African Language Technology AfLaT 2010

Edited by
Guy De Pauw
Handré Groenewald
Gilles-Maurice de Schryver

May 18, 2010
Mediterranean Conference Centre
Valletta, Malta

ISBN 2-9517408-6-7

Preface

African Language Technology is rapidly becoming one of the hottest new topics in computational linguistics. The increasing availability of digital resources, an exponentially growing number of publications and a myriad of exciting new projects are just some of the indications that African Language Technology has been firmly established as a mature field of research. The AfLaT workshops attempt to bring together researchers in the field of African Language Technology and provide a forum to present on-going efforts and discuss common obstacles and goals.

We are pleased to present to you the proceedings of the Second Workshop on African Language Technology (AfLaT 2010), which is held in collocation with the Seventh International Conference on Language Resources and Evaluation (LREC 2010). We were overwhelmed by the quantity and quality of the submissions we received this year, but were lucky enough to have a wonderful program committee, who sacrificed their valuable time to help us pick the cream of the crop. We pay tribute to their efforts by highlighting reviewers' quotes in the next paragraphs.

Grover et al. kick off the proceedings with a *comprehensive overview of the HLT situation in South Africa*, followed by **Bański and Wójtowicz**'s description of an initiative that is *beneficial to the creation of resources [...] for African languages*. **De Pauw et al.** describe techniques that *could be used to develop a plethora of [...] HLT resources with minimal human effort*, while **Shah et al.** present *impressive* results on tackling *the problem of NER in MT systems between languages, one of which at least is poorly resourced*. **Groenewald and du Plooy**'s paper tackles the *all too-often overlooked problem* of *text anonymization in corpus collection*, followed by **Chege et al.**'s effort that is *significant [...] to the open source community, not just for Gĩkũyũ but for the African languages in general*. **Faaß** presents *a useful resource for further computational processing of the language* of Northern Sotho.

Tachbelie and Menzel provide *a clear and concise overview of the general issues affecting language models for morphologically rich languages*, while **Van der Merwe et al.** go into an *informative discussion of the properties of the Zulu verb, its extensions, and deverbatives*. The paper by **Oosthuizen et al.** aptly discusses *the issue of quantifying and correcting transcription differences between inexperienced transcribers*, while **Davydov**'s paper is an interesting case study for collecting *corpora for "languages recently put into writing"*. Ng'ang'a presents *the key resource for the identification of a machine-readable dialectal dictionary for Igbo* and **Purvis** concludes by discussing a corpus that *contributes to the development of HLT tools for Dagbani*.

We are proud to have Justus Roux as the invited speaker for this year's edition of AfLaT to discuss one of the most often asked and rarely answered questions in our field of research: *Do we need linguistic knowledge for speech technology applications in African languages?* We hope you enjoy the AfLat 2010 workshop and look forward to meeting you again at AfLaT 2011.

The AfLaT 2010 Organizers
Guy De Pauw
Handré Groenewald
Gilles-Maurice de Schryver
Peter Waiganjo Wagacha
HTTP://AFLAT.ORG/AFLAT2010

Workshop Organizers

Guy De Pauw
CLiPS - Computational Linguistics Group, University of Antwerp, Belgium
School of Computing & Informatics, University of Nairobi, Kenya

Handré Groenewald
Centre for Text Technology (CTexT), North-West University, South Africa

Gilles-Maurice de Schryver
African Languages and Cultures, Ghent University, Belgium
Xhosa Department, University of the Western Cape, South Africa
TshwaneDJe HLT, South Africa

Peter Waiganjo Wagacha
School of Computing & Informatics, University of Nairobi, Kenya

Workshop Program Committee

Tunde Adegbola - African Languages Technology Initiative (Alt-i), Nigeria
Tadesse Anberbir - Addis Ababa University, Ethiopia
Winston Anderson - University of South Africa, South Africa
Lars Asker - Stockholm University, Sweden
Etienne Barnard - Meraka Institute, South Africa
Piotr Bański - University of Warsaw, Poland
Ansu Berg - North-West University, South Africa
Sonja Bosch - University of South Africa, South Africa
Chantal Enguehard - LINA - UMR CNRS, France
Gertrud Faaß - Universität Stuttgart, Germany
Björn Gambäck - Swedish Institute of Computer Science, Sweden
Katherine Getao - NEPAD e-Africa Commission, South Africa
Dafydd Gibbon - Universität Bielefeld, Germany
Arvi Hurskainen - University of Helsinki, Finland
Fred Kitoogo - Makerere University, Uganda
Roser Morante - University of Antwerp, Belgium
Lawrence Muchemi - University of Nairobi, Kenya
Wanjiku Ng'ang'a - University of Nairobi, Kenya
Ọdẹ́ṭúnjí Ọdẹ́jọbí - University College Cork, Ireland
Chinyere Ohiri-Anichie - University of Lagos, Nigeria
Sulene Pilon - North-West University, South Africa
Laurette Pretorius - University of South Africa, South Africa
Rigardt Pretorius - North-West University, South Africa
Danie Prinsloo - University of Pretoria, South Africa
Justus Roux - North-West University, South Africa
Kevin Scannell - Saint Louis University, United States
Gerhard Van Huyssteen - Meraka Institute, South Africa

Table of Contents

The Workshop Program

Tuesday, May 18, 2010

09:00–10:00	**Invited Talk** - *Do we need linguistic knowledge for speech technology applications in African languages?* Justus Roux
10:00–10:30	*An HLT profile of the official South African languages* Aditi Sharma Grover, Gerhard B. van Huyssteen and Marthinus W. Pretorius
10:30–11:00	**Coffee Break**
11:00–11:30	*Open-Content Text Corpus for African languages* Piotr Bański and Beata Wójtowicz
11:30–12:00	*A Knowledge-Light Approach to Luo Machine Translation and Part-of-Speech Tagging* Guy De Pauw, Naomi Maajabu and Peter Waiganjo Wagacha
12:00–12:30	*SYNERGY: A Named Entity Recognition System for Resource-scarce Languages such as Swahili using Online Machine Translation* Rushin Shah, Bo Lin, Anatole Gershman and Robert Frederking
12:30–13:00	*Processing Parallel Text Corpora for Three South African Language Pairs in the Autshumato Project* Hendrik J. Groenewald and Liza du Plooy
13:00–14:30	**Lunch Break**

Tuesday, May 18, 2010 (continued)

14:30–15:00	*Developing an Open source spell checker for Gĩkũyũ* Kamau Chege, Peter Waiganjo Wagacha, Guy De Pauw, Lawrence Muchemi, Wanjiku Ng'ang'a, Kenneth Ngure and Jayne Mutiga
15:00–15:30	*The verbal phrase of Northern Sotho: A morpho-syntactic perspective* Gertrud Faaß
15:30–16:00	*Capturing Word-level Dependencies in Morpheme-based Language Modeling* Martha Yifiru Tachbelie and Wolfgang Menzel
16:00–16:30	**Coffee Break**
16:30–17:00	*Towards the Implementation of a Refined Data Model for a Zulu Machine-Readable Lexicon* Ronell Van der Merwe, Laurette Pretorius and Sonja Bosch
17:00–17:30	*Improving Orthographic Transcriptions of Speech Corpora* Nicolas Laurens Oosthuizen, Martin Johannes Puttkammer and Martin Schlemmer
17:30–18:00	*Towards The Manding Corpus: Texts Selection Principles and Metatext Markup* Artem Davydov
18:00–18:30	*Towards a Comprehensive, Machine-readable Dialectal Dictionary of Igbo* Wanjiku Ng'ang'a
18:30–19:00	*Corpus Building in a Predominantly Oral Culture: Notes on the Development of a Multi-Genre Tagged Corpus of Dagbani* Tristan Purvis

- Invited Talk - Do we need linguistic knowledge for speech technology applications in African languages?

Justus C Roux

School of Languages, North West University
Potchefstroom, South Africa
Justus.roux@nwu.ac.za

Abstract

Given the fact that this workshop focuses on African languages, the main thrust of the presentation will be on these languages, however, certain points of discussion could just as well apply to languages in general. The topic of this presentation makes reference to three concepts that will be discussed individually against the background of an "uneasy relationship" that seemingly exists between linguists, engineers and computer scientists working in the domain of Human Language Technologies.

The concepts under discussion in the topic are *linguistic knowledge*, *speech technology applications* and *African languages*.

In the discussion of *linguistic knowledge* the following questions will be raised and addressed:

- What types of knowledge are we speaking about?
- Are these types of knowledge available in African languages for the development of systems?

The discussion on *Speech technology applications* will entertain some arguments why it is necessary to prioritise speech-based applications in research and development (R&D) agendas as far as African languages are concerned. This will include a discussion of real world needs and available infrastructures in developing environments. Turning to the question of the application of linguistic knowledge in the development of speech-based systems, the difference between rule based, and data driven speech synthesis systems will be addressed and demonstrated.

The discussion on *African languages* will focus on the sustainability of generating new and applicable knowledge as a prerequisite for development of applications in African languages. Given the efforts of the current generation of scholars in African languages, and the seemingly declining interest in African linguistics, at least within the Southern African context, the question arises how to ensure and maintain growth in this sector. The discussion will be concluded with a number of suggestions and interventions across disciplines and countries that could possibly contribute to sustainable activities in the field of HLT in Africa.

An HLT profile of the official South African languages

Aditi Sharma Grover[1,2], Gerhard B. van Huyssteen[1,3], Marthinus W. Pretorius[2]
Human Language Technology Research Group, CSIR[1],
Graduate School of Technology Management, University of Pretoria[2],
Centre for Text Technology (CTexT), North-West University[3]
HLT RG, Meraka Institute, CSIR, P.O. Box 395, Pretoria 0001, South Africa
asharma1@csir.co.za, gvhuyssteen@csir.co.za, tinus.pretorius@up.ac.za

Abstract

Human language technologies (HLT) have been identified as a priority area by the South African government to enable its eleven official languages technologically. We present the results of a technology audit for the South African HLT landscape, and reveal that a number of HLT components are available in SA but are of a very basic and exploratory nature and much investment is needed in the development of HLT language resources (LRs) in SA. The South African HLT landscape is analysed using a number of complementary approaches and based on the interpretations of the results, recommendations are made on how to accelerate HLT development in SA.

1. Introduction

Over the past few years, the South African government has realised the role that human language technology (HLT) could play in bridging the digital divide in South Africa. Various research and development (R&D) projects and initiatives have been funded by government, notably through its Department of Arts and Culture (DAC), Department of Science and Technology (DST), and National Research Foundation (NRF). For a historical perspective on HLT policy and non-R&D initiatives in South Africa, see Roux & Du Plessis (2005) and Sharma Grover *et al*. (submitted) for recent initiatives.

In 2009 the National HLT Network (NHN), funded by the DST, conducted the South African HLT audit (SAHLTA). The need for a technology audit is evident in the HLT community where discourse with respect to R&D is vibrant, but with a lack of a unified picture that presents the technological profile of the South African HLT landscape. We present in this paper the results of SAHLTA, focussing on a technological profile of the official South African languages.

2. SAHLTA Process: A Brief Overview

The BLaRK concept (Binnenpoorte *et al.*, 2002 and Maegaard *et al.,* 2009) was chosen to guide the audit, since it provides a well-defined structure to capture the different HLT components as data, modules, and applications.

A questionnaire was used as the primary means to gather data, capturing relevant information according to set criteria. This questionnaire was sent to all major HLT role-players in the country, with the request to supply detailed information regarding LRs and applications developed at their institutions. This audit questionnaire consisted of four major sections: one for each HLT component category (i.e. 'Data', 'Module', 'Application'), as well as a section, 'Tools/Platforms', which was added to accommodate technologies that are typically language-independent, or that aid the development of HLTs (e.g. annotation tools, or corpus searching tools); each section includes the most relevant audit criteria (e.g. maturity, accessibility, quality) for that particular category.

The audit questionnaire was sent to all major HLT role-players in the country. Organisations approached were classified as primary (universities, science councils, and companies-15) or secondary (national lexicography units, government departments-12) participants, based on their historical core HLT competence in R&D. All primary participants were paid a minimal honorarium to compensate for the considerable effort that was required from them.

For further details on the SAHLTA process and instruments used, see Sharma Grover *et al*. (2010).
In order to compare data (e.g. languages with each other), we experimented with various (subjective) ways to quantify the data. We developed a number or indexes in order to represent the technological profiles of the South African languages comparatively; these indexes are discussed and presented below.

3. Maturity Index

The Maturity Index measures the maturity of components by taking into account the maturity stage (i.e. development stage) of an item against the relative importance of each maturity stage. The 'maturity sum' per item grouping (e.g. 'pronunciation resources') for each language is calculated as:

$$MatureSum = 1\times UD + 2\times AV + 4\times BV + 8\times RV \qquad (1)$$

where UD is the number of components in the 'under development' phase, AV is the number of 'alpha version' components, BV the number of 'beta version' components, and RV the number of 'released version' components; the weights for the different versions are relative weights, in order to give greater importance to the final, released versions of components. Table 1 illustrates the maturity sum calculation for 'pronunciation resources'; the maturity sum for English will be 17, since English has one 'under development' item, no 'alpha' or 'beta version' items and two 'released' items; thus its maturity sum is

calculated to be (1×1+0×2+0×4+2×8)=17.

Data item grouping	Maturity Stage	Maturity Weight	Eng	Afr	Zul	Xho	Ndb	Ssw	Ses	Sep	Sts	Xit	Tsv	Lang. independent
Pronunciation resources	Under development	1	1	1	1	0	0	0	1	1	1	0	0	-
(Phoneme sets, Pronunciation dictionaries, Multilingual pronunciation lexicons, Pronunciation models, Intonation models)	Alpha version	2	0	0	0	0	0	0	0	0	0	0	1	-
	Beta version	4	0	0	0	0	0	0	0	1	0	0	0	-
	Released	8	2	5	4	4	3	3	4	3	3	3	3	-
	No. of items		**3**	**6**	**5**	**4**	**3**	**3**	**5**	**5**	**4**	**3**	**4**	-
	Maturity sum		**17**	**41**	**33**	**32**	**24**	**24**	**33**	**29**	**25**	**24**	**26**	-

Table 1: Example of a 'maturity sum' calculation for pronunciation resources.

Maturity sums were calculated across component groupings for all data, modules and applications per language. To obtain a comparative approximation of the maturity across languages, the Maturity Index (per language) was calculated by normalising the total of all the maturity sums (i.e. all item groupings across data, modules and applications for a language) by the sum of weights for the maturity stages (1+2+4+8=15). Table 2 presents this Maturity Index per language; note that this index is a relative index, based on the number of components that exist in a language.

Lang	MatureInd	AccessInd	LangInd
SAE	26.0	28.2	54.2
Afr	37.9	36.7	74.6
Zul	21.7	25.0	46.7
Xho	20.9	22.3	43.2
Ndb	11.5	11.0	22.5
Ssw	11.6	11.4	23.0
Ses	17.7	20.4	38.1
Sep	18.1	22.3	40.4
Sts	18.5	21.9	40.4
Xit	10.9	11.0	21.9
Tsv	11.9	12.1	24.0
L.I	10.2	8.6	18.8

Table 2: Maturity index, Accessibility index and Language index per language[1].

4. Accessibility Index

The Accessibility Index provides a measure of the accessibility of HLT components in a language by considering the accessibility stage of an item as well as the relative importance of each accessibility stage. The 'accessibility sum' is calculated per HLT component grouping for each language as follows:

$$AccessSum=1\times UN+2\times NA+4\times RE+8\times CO+12\times CRE \qquad (2)$$

where UN is the number of components that are classified as 'Unspecified' in terms of the accessibility stage, NA the number of components that are listed as 'Not available (proprietary or contract R&D)', RE the number of components 'available for research and education (R&E)', CO the number of components 'available for commercial purposes', and CRE the number of components 'available for commercial purposes and R&E'. Relative weights were assigned to the different accessibility stages, with higher weights for stages that make a component more accessible (e.g. available for commercial purposes). Also, since the 'available for commercial purposes and R&E' stage is a combination of the previous 'commercial only' and 'R&E only' categories, it was assigned only 1.5 times the weight of the preceding score (i.e. 1.5×8=12).

Accessibility sums were calculated across component groupings for all data, modules and applications per language. The Accessibility Index (per language) provides a comparative approximation of the accessibility of HLT components across all the languages. It was calculated by normalising the grand total of the accessibility sums from all the data, modules and applications component groupings per language, by dividing it with the sum of the weights of the accessibility stages (1+2+4+8+12=27). Results for the Accessibility Index are also presented in Table 2.

5. Language Index

The Language Index provides an impressionistic comparison on the overall status of HLT development for the eleven South African languages and was calculated by summation of the Maturity Index and the Accessibility Index for each language (across all HLT components):

$$LangInd = Maturity\ Index + Accessibility\ Index^{2} \qquad (3)$$

From the Language Index presented in Table 2, it emerges that Afrikaans has the most prominent technological profile of all the languages, followed by the local vernacular of South African English. The fact that Afrikaans scores higher than English on this index, can be attributed to the fact that very relatively little work on South African English is required within the text domain; South African English will therefore almost always only be measured in terms of activity related to speech technologies.

The two languages with the most native speakers, isiZulu and isiXhosa (both Nguni languages) follow behind English, and have both slightly more prominent profiles compared to the Sotho languages (Sepedi, Setswana and Sesotho). This can be attributed to the fact that isiZulu and isiXhosa are often of larger commercial and/or academic interest, because they are used more widely throughout South Africa. At the tail-end are the lesser-used languages, viz. Tshivenda, Siswati, isiNdebele and Xitsonga. These four languages significantly lag behind in terms of HLT activity; the majority of items available for these languages were developed quite recently, and are mainly due to the South African government's investment in these languages.

[1] SAE – South African English, Afr – Afrikaans, Zul – isiZulu, Xho – isiXhosa, Ndb – isiNdebele, Ssw – SiSwati, Ses – Southern Sotho (Sesotho), Sep – Northern Sotho (Sesotho sa Leboa/Sepedi), Sts – Setswana, Xit – Xitsonga, Tsv – Tshivenda, L.I – language independent.

[2] The Maturity Index and the Accessibility Index here is on a per language basis, taken across all data, modules, applications as discussed in section 3 and 4 respectively.

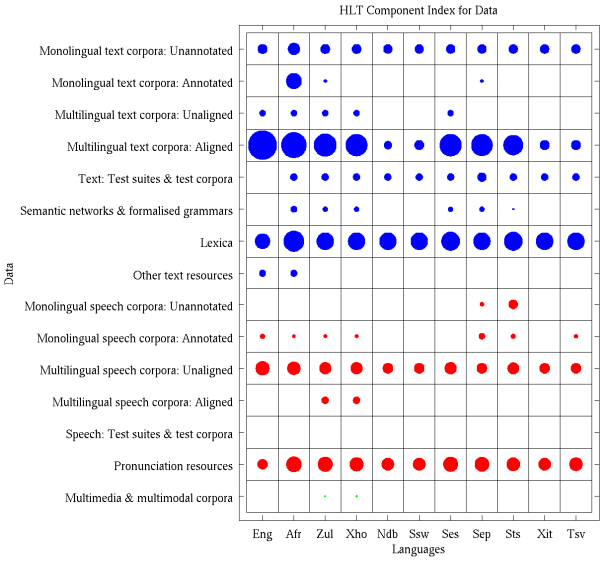


Figure 1: HLT Component Index for Data

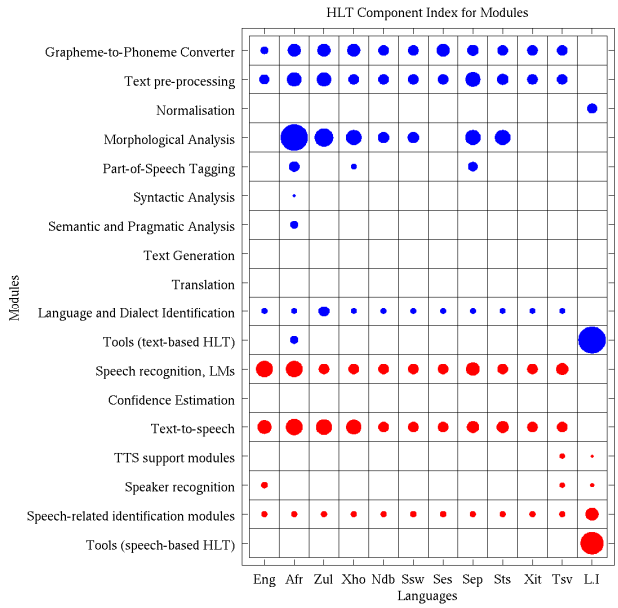


Figure 2: HLT Component Index for Modules

6. Component Indexes

The Component Indexes provide an alternative perspective on the quantity of activity taking place within the data, modules and application categories on a component grouping level (e.g. pronunciation resources), and is calculated as follows:

Component Index= Maturity Index (per item grouping) + Accessibility Index (per item grouping)[3] *(4)*

The Component Indexes for all languages are plotted in a grid using a bubble plot (see Figures 1 and 2; for the index on applications, see Sharma Grover *et al*., submitted). The value of the Component Index for a particular component grouping determines the size of the bubble (i.e. the higher the index the larger the bubble). It is important to note that the size of the bubbles plotted within a plot is proportional to the highest value of the Component Index within that specific plot. Thus, this index provides a relative comparison of the HLT activity within the various groupings of data, modules or applications within a single plot, as opposed to an absolute comparison of languages.

Figure 1 depicts the plot for the Component Index for data, where aligned multilingual text corpora have the highest score (which implies the greatest quantity of mature and accessible activities), followed by 'lexica' and so forth. In general it can be seen that speech data resources (in red) have less activity compared to text resources (in blue).

The figure also reveals that although there may be activity in many of the data sub-categories (e.g. 'semantic networks and formalised grammars'), it is very small (implying less maturity and accessibility), or does not exist across all the languages.

Figure 2 reveals that the largest amount of activity is in the field of morphological analysis in Afrikaans, as well as in text-based and speech-based tools that are language-independent (indicated by 'LI'). In general, there is some medium-scale activity in the basic HLT modules for both text and speech domains, such as grapheme-to-phoneme conversion, morphological analysis, speech recognition, and text-to-speech, while the more advanced modules are mostly non-existent or barely available for a few languages.

7. Discussion

The audit's findings reveal that while there is a significant level of HLT activity in South Africa, there are considerable differences in the amount of activity across the languages, and in general the LRs and applications currently available are of a very basic nature. In order to understand this, we need to take a holistic view of the current HLT landscape in South Africa; below we reflect on several factors that might have an influence on this.

HLT expert knowledge: Linguistic knowledge plays a crucial role in HLT enabling a language. In general, the availability of linguistic experts in South Africa is, compared to Europe or the USA, limited. Due to historical imbalances of the past, more linguistic expertise and foundational work are available for Afrikaans and South African English. This is followed by languages such as isiZulu, isiXhosa, Sepedi, Setswana and Sesotho, which have a larger pool of native speakers in South Africa, and thus a greater likelihood of linguistic experts' availability (as opposed to the smaller languages like Tshivenda, isiNdebele, SiSwati and Xitsonga).

Availability of data resources: The frequency and availability of text (e.g. newspapers, books, periodicals, documents) and speech sources (e.g. audio recordings) is far greater for languages such as Afrikaans and South African English, as opposed to the African languages (and even more so for the smaller languages). This is a challenge since the R&D community is constantly faced with limited data collections when working with the

[3] The Maturity Index and Accessibility Index used here are calculated for each grouping of HLT components within data, modules and applications (for example, lexica, corpora, morphological analysis, translation, etc.).

African languages, which in turn severely inhibits the HLT development of the latter group.

Market needs of a language: The market needs of HLT in a particular language can be viewed as a combination of supply-and-demand factors, and the functional status of the language in the public domain. By supply-and-demand, one mostly refers to the size and nature of the target population for the language, while the functional status refers to the usage of a language in various public domains. In South Africa, English (and to a somewhat lesser extent Afrikaans) is by and large the lingua franca in the business domain, while the African languages are less widely used in such commercial environments. This significantly lowers the economic feasibility of HLT endeavours for these languages.

Relatedness to other world languages: Cross-language information reuse and bootstrapping approaches (Davel & Barnard, 2003; De Pauw *et a.l*, 2006) based on other linguistically similar languages can be used to initiate HLT development of new languages. Linguistically, Afrikaans is very similar to Dutch, and thus has benefitted and leveraged on the HLT approaches and developments made in Dutch. Conversely, South African English has not received significant attention in HLT R&D, since researchers leverage on adapting and reusing other international English LRs rather than investing in South African English LR generation from scratch. This leads to a lesser amount of home-grown South African English LRs, and explains the lesser position South African English takes to Afrikaans.
In contrast to the above, African languages are linguistically very dissimilar to any of the European or other world languages, and thus cannot leverage on an existing pool of closely related language resources. This fact, coupled with the complexity of African languages (e.g. tone, clicks, linguistic structure, etc.), leads to these languages having to commence their HLT efforts from the bottom of the development lifecycle, and start by investing in basic LR and linguistic knowledge generation.

Interplay of the above-mentioned factors with the socio-economic and political background of South Africa has shaped the HLT efforts across the South African languages, which has resulted in significant differences in the level of HLT activity across the eleven official languages of South Africa.

8. Conclusion

The South African HLT community is faced with the challenging task of balancing political needs (i.e. attention to all official languages equally) with economic viability (i.e. create a thriving HLT industry, with return on investment on HLT for a certain language). South Africa is far from being unique in this sense, and a number of recommendations can be made for accelerating HLT development in SA and other (African) countries.

Resource development and distribution: From the results it is discernible that basic core LRs need to be built for all languages. However, it is also important to note that while building basic LRs should be prioritised, one needs to start building experience in developing more advanced LRs for future fast-tracking of HLT applications. In addition, market needs and trends should be a prime consideration in the development of such LRs. It was also observed in the audit that licensing agreements were often not defined for numerous LRs (often for government funded research projects).
Thus, although some of these LRs may be declared as accessible (available for commercial and R&E usage) the ambiguity around the licensing leads to delays and obstacles in using them. Therefore, in order to encourage innovation, LRs should preferably be made freely available in the open source domain; alternatively, where these are subject to intellectual property rights for commercial use, LRs should be available at a price that does not prohibit their usage.

Funding: The principal sponsors of HLT development for resource-scarce languages are often the governments of those countries. In contrast, the HLT industries in such countries often only comprise a handful of companies that focus on a few languages, since the initial investment required does not cover the potential income from the projected market needs for most languages. Thus, in the formative years governments needs to continue to invest in HLT efforts (especially in the development of LRs) to build a strong foundation of HLT outputs, which could create thriving HLT industries in such countries.

Industry stimulation programmes: Besides funding, governments need to ensure that there are more initiatives to encourage the existing industry's participation in national HLT activities, and to enable the establishment of new HLT-based start-up companies. For example, it has been noted that there is little awareness in the South African commercial sector about the opportunities and positive impact of HLT (e.g. in the financial or ICT sectors). HLT-focussed initiatives could be launched to stimulate R&D partnerships between academia and industry. In addition, industry participation in lesser resourced languages may need to be motivated proactively by such governments.

*Collaborations: C*losely related to the above-mentioned stimulation programmes is the need for greater collaborations within local HLT communities and the larger international community. One of the challenges is to harness the knowledge and skills developed in local pockets of excellence into a collaborative endeavour. Thus, a more coordinated effort across an HLT community is required to ensure that there is a well-mapped trajectory for LR creation and HLT market development. Also, collaboration across disciplines (e.g. linguistics, engineering, mathematics) should be encouraged, since HLT involves crossing silos of academic disciplines and national borders.

Human capital development (HCD): The shortage of linguistic and HLT expertise (and general scientific capacity) is a prohibitive factor in the progress of HLT; thus, HCD efforts within the field of HLT should be accelerated. For example, there is currently only one South African undergraduate HLT degree programme of its kind (Pilon, *et al,,* 2005), while most other training

courses are at the postgraduate level. A greater investment needs to be made in generating HLT practitioners who can feed into the emergent HLT industry's pipeline.

Cultivation of niche expertise: It was observed in the audit that a number of language-independent methods have been adopted in creating HLT components for SA. This approach (depending on the LR in question) has the potential to fast-track the development of HLTs across other resource-scarce languages.

9. References

Binnenpoorte, D., De Friend, F., Sturm, J., Daelemans, W., Strik, H. & Cucchinari, C. (2002). A Field Survey for Establishing Priorities in the Development of HLT Resources for Dutch. In *Proc. LREC 2002*, Spain.

Davel, M. & Barnard, E., (2003). Bootstrapping in Language Resource Generation. In *Proc. Symposium of Pattern Recognition Society of South Africa*, November 2003.

De Pauw G., de Schryver, G-M, & Wagacha, P.W., (2006) Data-driven part-of-speech tagging of Kiswahili. In P. Sojka, I. Kopeˇcek, and K. Pala *(ed)*, *Proc. Text, Speech and Dialogue, 9th International Conference*, LNCS, Berlin: Springer Verlag, vol. 4188, pp. 197–204.

Maegaard, B., Krauwer, S. & Choukri, K. (2009). *BLaRK for Arabic. MEDAR – Mediterranean Arabic Language and Speech Technology*. [Online]. Available: http://www.medar.info/MEDAR_BLARK_I.pdf (accessed June 2009)

Roux J. & Du Plessis, T., (2005). *The Development of Human Language Technology Policy in South-Africa*. In Daelemans, W., Du Plessis, T., Snyman, C. &Teck, L., Multilingualism and Electronic Language Management, Pretoria: Van Schaik, pp. 24–36.

Pilon, S., van Huyssteen, G.B. & Van Rooy, B. (2005). Teaching Language Technology at the North-West University. In *Proc. Second ACL-TNLP Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*. June. Ann Arbor, Michigan: University of Michigan. pp. 57–61.

Sharma Grover, A., van Huyssteen G..B. & Pretorius, M.W. (submitted). *The South African Human Language Technology Audit*. [submitted for publication in Journal for Language Resources and Evaluation]

Sharma Grover, A., van Huyssteen G.B & Pretorius, M.W. (2010). The South African Human Language Technologies Audit. In *Proc. LREC 2010*, Malta [accepted].

Acknowledgements

We would like to thank the Department of Science and Technology (DST) for funding this audit. We would also like to acknowledge Professor S. Bosch and Professor L. Pretorius from UNISA whose 2008 BLaRK questionnaire results (for the 2008 NHN-NTU workshop) and preliminary language-specific inventories were used to build the first draft of the cursory inventory of HLT items available in South Africa.

Open-Content Text Corpus for African languages

Piotr Bański[1], Beata Wójtowicz[2]

[1] Institute of English Studies, [2] Faculty of Oriental Studies
University of Warsaw
[1] Nowy Świat 4, [2] Krakowskie Przedmieście 26/28
[1] 00-497, [2] 00-927
Warszawa, Poland
pkbanski@uw.edu.pl, b.wojtowicz@uw.edu.pl

Abstract

The paper addresses the issue of scarce availability of linguistic resources for African languages and proposes a systemic solution that has a chance to improve the current state of affairs in more than one domain of language technology. We introduce the Open-Content Text Corpus, a freely available and transparently built resource that may combine independently maintained individual language corpora into parallel subcorpora that can be used, among others, in bilingual lexicography and machine translation. The individual language corpora feature multi-layer stand-off architecture and can be added gradually and allow for multiple views of the data as well as for comparisons of tools and corpus-linguistic approaches.

1. Introduction

It is the currently accepted state of affairs in the field of Language Resources, especially in under-funded areas, such as the area of African linguistics, that data are produced locally and guarded against dissemination, on the understandable premise that one had to go to enormous pains sometimes in order to obtain funding, interview speakers or gather permissions to use the relevant data (and the permissions most often include a non-disclosure clause). On the other hand, it is common for researchers to shun data produced by others – this is popularly dubbed the "not invented here" syndrome, where the motivation, if it is rationalized at all, is that the origin of the given resource is not obvious and the way it is constructed may be in conflict with best practices, as understood by the party refusing to accept the data.

The combination of the two creates a very serious obstacle to an individual researcher, without an institutional backing in terms of money and data access – compare the research and data-gathering that needed to be done e.g. by Minah Nabirye, who conducted eight year long data collecting research in order to produce her Lusoga dictionary (Nabirye, 2009) with the amount of data available at hand for e.g. a Google or Microsoft researcher. This grudgingly but widely accepted state of affairs promotes the creation of separate islands of data bound to the individual researcher or the individual institution. This also concerns well as what de Schryver (2002) calls Web-for-Corpus (which partly concerns Ide and Suderman, 2006, and is most prominently featured in the Crúbadán project, cf. Scannell, 2007), because due to the data that the corpora derived from the Web automatically incorporate, they are not free to disseminate, and therefore, although huge and with an enormous potential, they remain the closed property of their creators (or data distribution centres such as the LDC or ELRA) that they can at most make accessible on an individual basis and often with non-disclosure constraints.[1]

We would like to challenge this state of affairs. We have been observing the field of Language Resources for some time, and we have shared the suffering from the lack of free, unlimited access to language data (and, occasionally, from the lack of financial resources to obtain paid access). One way to address our need for access to a parallel corpus of Swahili and Polish would be to build our own, and in doing so, to replicate the efforts of others and, in the end, create another data island that we could use and write about, but that, crucially, we would not be able to share.

We do not want to do that, because that would mean perpetuating the access model that we strongly disagree with. Instead, we have taken a different approach: to create an open resource that will be free for all to access and use, and that will have (i) a transparent format that properly exposes the data and guarantees their sustainability, (ii) explicit version control that is critical for objective measurements of research results, (iii) a license that will encourage everyone to use it, while at the same time preserving information about the authorship, and (iv) a potential to evolve beyond the sum of its parts. Enter OCTC: the Open-Content Text Corpus.

The OCTC is hosted by Sourceforge.net – one of the world's largest repositories of open source/open content material. Sourceforge provides all feedback-gathering and

[1] Note that while Web-as-Corpus projects (cf. e.g. Kilgarriff and Grefenstette, 2003) are seemingly free of these considerations, they have other shortcomings, as De Schryver (2009) points out: the inherent instability of the Web makes it suitable for only a fraction of the purposes that corpora are used for. In the context of African languages, Web-as-Corpus has even more limited applicability.

community-forming facilities (forums, bug/idea/patch trackers, mailing lists), a version-control system together with a web front-end,[2] and a distribution mechanism.

In section 2, we look at the general architecture of the OCTC and point out its advantages over a plain-text single-language approach. We use fragments of the Swahili subcorpus to illustrate the overview. In section 3, we place the OCTC in the context of African languages. Section 4 discusses the potential uses of the corpus as a nexus for research and community-building, and section 5 concludes the paper.

2. OCTC: general architecture

The corpus has a transparent, multi-layer stand-off architecture, whereby annotations are held separately from the data stream (see 2.1 below) – that makes it easy to manipulate, extend, and use for various purposes. It is encoded in XML, according to the Guidelines of the Text Encoding Initiative – a *de facto* standard for corpus encoding.[3] The general architecture follows the principles adopted by the National Corpus of Polish (NCP, cf. Przepiórkowski and Bański, 2009), because they have already been tested in practice and because such a move will make it easy to adopt the open-source tools developed by the NCP: Anotatornia – the hand-annotation tool (Przepiórkowski and Murzynowski, 2009), and Poliqarp – the query engine that can efficiently handle multi-layer corpora (Janus and Przepiórkowski, 2007).

The OCTC has a monolingual part and a parallel part. They are discussed in the two subsections below. The third subsection looks at the finer divisions within the two parts and the structuring of meta-information.

2.1. Monolingual part

Each text in the corpus is placed in a separate directory, together with, minimally, the header that describes it and a catalog.xml file that provides the locations of the crucial parts of corpus infrastructure (RelaxNG schemas, and the higher-level headers). This is shown below.

(1) Contents of leaf directories of the monolingual part of the OCTC

catalog.xml – cross-corpus navigation
header.xml – metadata, changelog, etc.
text.xml – source text with structural markup
(morph.xml, etc.) – stand-off annotation file(s)

The file text.xml contains the source text with the structural encoding of the most general kind: the identification of chapters, sections and paragraphs. Only the "narrative" text content gets included, most lists and all graphics, tables and the like are removed. The content of title pages is placed in the header.xml that accompanies the text. The header provides the metadata: it states the authorship information, specifies any special techniques used in the process of "up-translation", i.e. converting the text to the OCTC format, lists the changes that the files in the given directory have undergone, and classifies the text according to various criteria (e.g. the genre, licensing of the original, subcorpus membership). The header is included by text.xml and other, optional files that contain stand-off annotation information (e.g., the morph.xml file can contain morphosyntactic information output by a tagger).

The stand-off architecture of the corpus is modelled on that of the National Corpus of Polish and involves a hierarchy of linguistic levels, beginning with the level of text segments that the other corpus files may refer to (e.g., the morph.xml file can specify morphosyntactic information for each text token in the source, and the syntax.xml file can gather the morphosyntactic pieces into syntactic chunks). The crucial notion here is that of reference to the source text, which is realised by @corresp attributes, as shown below, where the <seg/> element identifies the first token in the element identified as "swh_par_1"; the token begins at index 0 and contains five characters.[4]

(2)
```
<seg xml:id="seg_1" type="token"
     corresp="text.xml#string-range(swh_par_1,0,5)"/>
```

The annotation hierarchy in a single leaf directory is planned to look schematically as follows:

(3)
```
source text (with structural markup)
        ↑
segmentation information
        ↑
morphosyntactic information
        ↑
etc. (syntactic/semantic annotation)
```

Note that this kind of setup makes it possible to have more than one file with e.g. morphosyntactic information (using different tagsets or different taggers), pointing at the same file with segmentation information. This is the

[2] Despite appearances, version control, together with a milestone/tagging mechanism is crucial for dynamic corpora if we want to maintain that corpus linguistics is an empirical science where experiments are *reproducible*. Without versioning, an experiment performed today may easily *not* yield the same results a year from now, due e.g. to the corpus size changing in the meantime. We are grateful for this observation to Henry S. Thompson (p.c.). Not that this issue is also relevant to de Schryver's (2009) criticism of Web-as-Corpus approaches.

[3] See http://www.tei-c.org/Activities/Projects/ for an incomplete list of encoding projects using the TEI and e.g. http://www.tractor.bham.ac.uk/tractor/catalogue.html for a list of Language Resources containing numerous applications of the TEI.

[4] Other mechanisms of expressing correspondence across levels exist in the TEI. The syntax of the string-range() XPointer scheme is described in ch. 16 of the TEI Guidelines (TEI Consortium, 2009).

well-known advantage of stand-off annotation (see e.g. Ide and Suderman, 2006; Ide and Romary, 2007), where possibly conflicting or overlapping annotations are kept in separate files and thus provide different views of the annotated data, and can even be compared among themselves.

2.2. Parallel part

Parallel subcorpora are kept in a separate corpus directory and conform to a separate TEI schema. They contain <linkGrp> elements providing alignment information at various levels of granularity. Currently, in the demonstration Polish-Swahili aligned subcorpus, the lowest alignment level is that of paragraph-sized chunks, but as De Pauw *et al.* (2009) demonstrate on the basis of the SAWA corpus, going below this level is a feasible task.
Below, we provide a modified fragment of a Polish-Swahili alignment file.

(4) Fragments of an alignment document

a. Gross alignment with <link> elements

```
<linkGrp type="texts">
 <link targets="pol/UDHR/text.xml#text_body
            swh/UDHR/text.xml#text_body"/>
</linkGrp>
<linkGrp type="divs">
 <link
 targets="pol/UDHR/text.xml#pol_txt_3-div
        swh/UDHR/text.xml#swh_txt_3-div"/>
 <link
  targets="pol/UDHR/text.xml#pol_txt_4-div
        swh/UDHR/text.xml#swh_txt_4-div"/>
 </linkGrp>
```

b. Lower-level alignment with typed <ptr> elements

```
<linkGrp type="para">
 <ptr
  target="pol/UDHR/text.xml#swh_txt_4.2-p"
  type="pol" xml:lang="pl"/>
 <ptr
  target="swh/UDHR/text.xml#swh_txt_4.2-p"
  type="swh" xml:lang="sw"/>
</linkGrp>
```

As can be seen in (4b), the pointers are typed with the appropriate ISO code. This is because there may be more than one <ptr> element for a particular language, in case of one-to-many alignment relationships. Unaligned chunks of text may also be indicated in a similar manner, in order to maintain the text flow for visualization purposes.
Note that it is possible for the linkGrp elements to hold pointers to more than two monolingual subcorpora, and thus it is possible to create alignment files that put together translations from more than two languages at once. The aligned part of the OCTC is what we mean by saying that the corpus may become more than a sum of its parts – this is how monolingual texts can participate in the creation of bi- or multi-lingual derived resources.

2.3. Subcorpora and headers

The OCTC is subdivided along many dimensions. As mentioned above, the primary division is into the monolingual and the parallel part. Within the monolingual part, each language subcorpus is in a directory of its own, identified by an ISO 639-3 code (hence we talk about "OCTC-swh" to refer to the Swahili subcorpus). Within the bilingual part, the particular subcorpora are also divided on a per-directory basis, according to ISO codes arranged alphabetically, hence "pol-swh" is the directory name for the parallel Polish-Swahili corpus. Each of such subcorpora (we call them *primary subcorpora*) is expected to have a separate subcorpus header, gathering all the relevant meta-information.
However, a text's membership in a subcorpus is not always a biunique relationship: any Swahili text is part of OCTC-swh, but on top of that, a Swahili text of the Universal Declaration of Human Rights is part of the (*secondary*) subcorpus of UDHR texts, and may also become part of other secondary subcorpora, for example a subcorpus of legal texts. In such cases, rather than using the tree-like file-directory structure or XML architecture, we delegate the task of identifying membership in a secondary subcorpus to the text itself, or more precisely, to its header.
Recall that each corpus file (whether containing the source text or its annotation) includes the local header, which records all the modifications and properties of the given text. This is the place where information about the text's membership in subcorpora is indicated. The main corpus header, on the other hand, contains the taxonomy of subcorpora. The local headers are only permitted to use the values defined in the main header. This is illustrated below.

(5) Subcorpus taxonomy

a. Fragment of the main corpus header

```
<classDecl>
 <taxonomy xml:id="OCTC-subcorpora">
  <category xml:id="subc_UDHR"
   corresp="#subc_UDHR-description">
    <catDesc>Subcorpus of Universal
    Declaration of Human Rights translations
    (…)</catDesc>
  </category>
  (other taxonomy definitions follow)
```

b. Fragment of a local header

```
<profileDesc>
 <textClass>
  <catRef scheme="#OCTC-subcorpora"
    target="#subc_UDHR"/>
  <catRef scheme="#OCTC-text_types"
    target="#ttyp_legal-official
            #ttyp_net_non-interactive"/>
  </textClass>
 </profileDesc>
```

We have mentioned three header types: (i) the local

header, (ii) the (primary) subcorpus header, and (iii) the main corpus header. In fact, each text is expected to incorporate all three, and thus three levels of information: (i) local information concerning the text and its annotations, (ii) information concerning the maintenance of the given primary subcorpus, and (iii) information concerning the entire corpus (licensing, major changes, administrative contacts, definitions of the taxonomy of texts and subcorpora). Each header is included by means of the XInclude mechanism, and the non-local headers are located thanks to the catalog.xml files, implementing the OASIS XML Catalogs standard (Walsh, 2005).

3. The corpus seeds

The first material committed to the OCTC repository were translations of the Universal Declaration of Human Rights (henceforth UDHR), sometimes accompanied by introductory text (uniformly marked with the attribute @n="intro"). These texts form the seeds for numerous subcorpora of individual languages. The reason was partly symbolic and partly practical: we used the text addressing the right to, and the need for, individual freedom, in an enterprise that is meant to address the painful lack of freedom as far as language data and linguistic resources are concerned. Additionally, such little pieces of texts were ideal seeds for the individual language subcorpora, but also for parallel subcorpora.

Our selection of the texts to be included was partly subjective: we concentrated on Sub-Saharan African languages and transduced most of those texts which had HTML versions with Unicode fonts. We have then added texts for major languages of the world (again, those with HTML versions of the UDHR), because these are likely to form one pole of bilingual studies and MT tasks involving African languages. Finally, we added Polish, because we need it for Polish-Swahili alignment, and some other Slavic languages.[5]

We have planted corpus seeds for, among others, Bambara, Ewe, Hausa, Igbo, Northern Sotho, Shona, Somali, Swati, Southern Sotho, Tonga, Tswana, Wolof, Xhosa, Yoruba, and Zulu (and we slowly proceed with others, to present as many of them as possible to the AfLaT community). This means that researchers in the above-mentioned languages can submit further directories with open-content texts, according to the examples that already exist. For others, the procedure is minimally more complicated, as the subcorpus for the given language has to be created first. We are going to delegate the maintenance of primary subcorpora to researchers interested in extending the resources for their languages. All the project facilities that exist for the OCTC in general can be delegated to the particular subprojects as well – it is possible to have mailing lists or sections of the bulletin board devoted to a particular subproject and administered by that project's maintainers.

Because our interests focus on Swahili, we have already prepared additional free texts beyond the UDHR, and are actively looking for others to include in OCTC-swh. It is worth mentioning that non-free texts may also be considered for inclusion in the OCTC, if their chapters or sections are sampled for individual paragraphs or sentences, in order to preserve the linguistic content but at the same time ensure e.g. commercial publishers that their financial interests are not at risk.

4. OCTC's potential for research and community building

Potential data islands among parallel corpus projects targeting African languages are relatively easy to find and we refer to just two: De Schryver (2002) mentions a parallel Web-derived corpus for 11 official languages of South Africa, Chiarcos *et al.* (2009) mention a corpus of 25 African languages. Such and similar efforts might, at least partially (given copyright restrictions on texts) be conducted in the context of the OCTC.

On the other hand, what may be called predecessors of the OCTC have also appeared: Scannell (2007) reports on the Crúbadán project, partly community-driven and containing resources for ca. 300 languages (unfortunately, the legal status of some of the texts downloaded from the Web does not allow them to be distributed under a fully open license, though we acknowledge Kevin Scannell's efforts towards making linguistic data maximally available for research. The Crúbadán community also demonstrates our point about willingness to share data and work towards common good for free, given an established framework. Tyers and Salperen (forthcoming) report on SETIMES, a free, open-content parallel corpus of 8 Balkan languages. While the data in the American National Corpus are not open, it is worth mentioning the ANC's sub-initiative called Open Linguistic Infrastructure, grouping free annotations of the ANC's data, supplied by a wide community of researchers (cf. Ide and Suderman, 2006).

Apart from data, we hope that the OCTC will create environment conducive towards producing or customizing tools that can be used for tagging, visualization or querying multiple languages. Again, some such tools have already been reported on, e.g. the data-driven, language-independent methods for diacritic-restoration, lemmatisation or POS-tagging reported on by De Pauw and De Schryver (2009). The National Corpus of Polish, while itself a closed resource, makes sure to release all its tools under open-source licenses. These tools – the robust query engine Poliquarp (Janus and Przepiórkowski, 2007) or manual annotation environment Anotatornia (Przepiórkowski and Murzynowski, forthcoming) only need fine-tuning to work for the OCTC, whose architecture resembles that of the NCP. Note that the single TEI XML schema requires a single converted to e.g. the PAULA format with its associated tools (see e.g. Chiarcos *et al.*, 2009).

[5] We encourage further submission of the text of the Declaration, both to the OCTC and the OHCHR. This also concerns texts which have no HTML representation in OHCHR and would require retyping or transduction into Unicode.

Above, we have mentioned the possibility to use multiple annotations of the same kind, e.g. POS-tagging output by different taggers, addressing a single body of primary data, i.e. the data in text.xml documents. This and other examples show the need not so much for unifying grammatical descriptions as for establishing references to common data categories, such as those created in the GOLD ontology (http://linguistics-ontology.org/, cf. (Farrar and Langendoen, 2003)) or ISOCat (http://www.isocat.org/, cf. (Kemps-Snijders *et al.*, 2008)). The fact that OCTC tools should best be applicable across language subcorpora might increase the incentive to work on such standardization issues.

What we cannot stress enough is that such a project creates a potential for minimizing the number of data islands – it can host many diverse subprojects and many sub-communities. It is also worth mentioning that such an environment may play a role in community-building, creating a platform for information exchange and co-operation among researchers.

5. Conclusion

It is hard to begin a project from scratch – it is much easier to add to something that already has a foundation. We have built such a foundation, with seeds for many languages already planted, in the form of the texts of the Universal Declaration of Human Rights. We are convinced that it will be easier for individual researchers to donate even single texts, encoded in simple XML and structured on the example of the existing ones. All texts will be uniformly indexed, and some of them, e.g. translations of the same basic text, will be able to become part of the parallel corpus simply by virtue of being part of the monolingual subcorpus of the OCTC. Recall that besides being seeds for monolingual languages, the texts of the UDHR are at the same time seeds for the possible parallel corpora.

Being licensed under the Lesser GNU Public License means that the corpus can be used as part of other corpora, including closed-content resources. Naturally, these resources will have to acknowledge the use of the OCTC, but will not be exposed to the so-called viral nature of some open-source licenses, such as the GNU Public License or the Creative Commons ShareAlike licenses. This is intentional, because we do not want to alienate researchers or institutions that have been creating their own corpora, which are not able to be distributed for various reasons. Quite on the contrary: we are hopeful that in exchange for texts gathered by, or donated to, the OCTC, the maintainers of closed-content corpora will consider giving back texts that they can afford to give: either public-domain texts or texts available on non-viral licenses (e.g. Creative Commons Attribution), or possibly sampled and reshuffled versions of their non-free texts. The only reflex that such an institution will have to suppress is the "it's mine and you can't have it" reflex, hopefully not very strong.[6]

At the beginning, we have referred to two situations hampering research and causing the emergence of data islands: the abundance of situations responsible for aversion to share data or not being allowed to do it, as well as the "Not-Invented-Here" syndrome. The OCTC has a chance to be a solution to both: those who need data can get it for free (and hopefully will give something back), and those who are not sure of the data origins will be able to see its history and adjust their measurements accordingly.

This is an experiment and we are hopeful that the African Linguistic community, realising that in many cases it is data islands that hamper the progress of research, especially research on "non-commercial" languages, will join us in expanding the OCTC. It is worth bearing in mind that data exchange does not hurt because it is precisely *exchange*: as a reward for donating one's data, one gets more data to operate on. While a researcher may contribute her data to one subcorpus, she may decide to use tools used by another subcorpus project, already tuned to the OCTC data format.

6. Acknowledgements

We are grateful to four anonymous AfLaT-2010 reviewers for their encouragement and constructive remarks. We have attempted to address all of them in the final version of the paper.

We would like to acknowledge the work performed by the Office of the United Nations High Commissioner for Human Rights (OHCHR), which solicited and published the text of the UDHR, most of it in relatively regular HTML that was easy to transduce into TEI P5.

We wish to thank Kevin Donnelly and Kevin Scannell for valuable discussions of the OCTC initiative, and to Jimmy O'Regan for being a good spirit right when we needed one.

Some of the ideas expressed here come from our earlier involvement with the National Corpus of Polish. We are especially indebted to the NCP for the text taxonomy that we have adopted in the OCTC with minimal changes.

Our research was partly financed by the Polish Ministry of Science and Higher Education project # N104 050437 "Small Swahili-Polish and Polish-Swahili dictionary".

References

Chiarcos, C., Fiedler, I., Grubic, M., Haida, A., Hartmann, K., Ritz, J., Schwarz, A., Zeldes, A. & Zimmermann, M. (2009). Information Structure in African Languages: Corpora and Tools. In G. De Pauw, G-M. de Schryver & L. Levin (Eds.) *Proceedings of the First Workshop on Language Technologies for African Languages.* Athens: ACL, pp. 17–24.

[6] We are grateful to Kevin Donnelly for his discussion of the licensing issues, which may result in our moving away from the LGPL towards GPL, in the interest of the creators of the data rather than closed-content data manufacturers.

De Pauw, G. & de Schryver, G.-M. (2009). African Language Technology: The Data-Driven Perspective. In V. Lyding (eds.). 2009. *LULCL II 2008 – Proceedings of the Second Colloquium on Lesser Used Languages and Computer Linguistics*, Bozen-Bolzano, 13th-14th November 2008 (EURAC book 54), Bozen-Bolzano: European Academy, pp. 79–96.

De Pauw, G., Wagacha, P.W. & de Schryver, G. M. (2009). The SAWA Corpus: a Parallel Corpus English-Swahili. In G. De Pauw, G-M. de Schryver & L. Levin (Eds.) *Proceedings of the First Workshop on Language Technologies for African Languages*. Athens: ACL, pp. 9–16.

De Schryver, G-M. (2002). Web for/as Corpus: A Perspective for the African Languages. *Nordic Journal of African Studies* 11/2, pp. 266–282.

De Schryver, G-M. (2009). An Analysis of *Practical Lexicography: A Reader* (Ed. Fontenelle 2008). *Lexikos* 19, pp. 458–489.

Farrar, S.; Langendoen, D.T. (2003). A linguistic ontology for the Semantic Web. *GLOT International*. 7 (3), pp. 97–100.

Ide, N. & Suderman, K. (2006). Integrating Linguistic Resources: The American National Corpus Model. *Proceedings of the Fifth Language Resources and Evaluation Conference (LREC)*, Genoa, Italy.

Ide, N. & Romary, L. (2007). Towards International Standards for Language Resources. In Dybkjaer, L., Hemsen, H., Minker, W. (eds.), Evaluation of Text and Speech Systems, Springer, pages 263–84.

Janus, D. & Przepiórkowski, A. (2007). Poliqarp: An open source corpus indexer and search engine with syntactic extensions. In *Proceedings of the ACL 2007 Demo Session.* Prague. pp. 85–88.

Kilgarriff, A. & Grefenstette, G.. (2003). Introduction to the Special Issue on the Web as Corpus. *Computational Linguistics* 29(3), pp. 333–347.

Nabirye, M. (2009). *Eiwanika ly'Olusoga* [Lusoga Dictionary]. Kampala: Menha Publishers.

Przepiórkowski, A. & Bański, P. (forthcoming). XML Text Interchange Format in the National Corpus of Polish. In S. Goźdź-Roszkowski (Ed.) *The proceedings of Practical Applications in Language and Computers PALC 2009*, Frankfurt am Main: Peter Lang.

Przepiórkowski, A. & Murzynowski, G. (forthcoming). Manual annotation of the National Corpus of Polish with Anotatornia. In S. Goźdź-Roszkowski (Ed.) *The proceedings of Practical Applications in Language and Computers (PALC-2009)*. Frankfurt: Peter Lang.

Scannell, K. P. (2007). The Crúbadán Project: Corpus building for under-resourced languages. In Fairon et al. (Eds.) *Building and Exploring Web Corpora.* Louvain-la-Neuve: PUL, pp. 5–15.

TEI Consortium (Eds.) (2010). TEI P5: Guidelines for Electronic Text Encoding and Interchange. [Online]. Available: http://www.tei-c.org/Guidelines/P5/ (accessed February 2010).

Tyers, F. & Salperen, M. (forthcoming) South-East European Times: A parallel corpus of Balkan languages. Forthcoming in the proceedings of the LREC workshop on "Exploitation of multilingual resources and tools for Central and (South) Eastern European Languages".

Walsh, N. (Ed.) (2005). XML Catalogs. OASIS Standard V1.1. [Online]. Available: http://www.oasis-open.org/committees/download.php/14809/xml-catalogs.html (accessed March 2010).

Kemps-Snijders, M., Windhouwer, M.A., Wittenburg, P. & Wright, S.E. (2008). ISOcat: Corralling Data Categories in the Wild. In European Language Resources Association (ELRA) (Ed.), *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, May 28-30, 2008.

A Knowledge-Light Approach to Luo Machine Translation and Part-of-Speech Tagging

Guy De Pauw[1,2], Naomi Maajabu[1], Peter Waiganjo Wagacha[2]

[1]CLiPS - Computational Linguistics Group
University of Antwerp, Belgium
{firstname.lastname}@ua.ac.be

[2]School of Computing & Informatics
University of Nairobi, Kenya
waiganjo@uonbi.ac.ke

Abstract

This paper describes the collection and exploitation of a small trilingual corpus English - Swahili - Luo (Dholuo). Taking advantage of existing morphosyntactic annotation tools for English and Swahili and the unsupervised induction of Luo morphological segmentation patterns, we are able to perform fairly accurate word alignment on factored data. This not only enables the development of a workable bidirectional statistical machine translation system English - Luo, but also allows us to part-of-speech tag the Luo data using the projection of annotation technique. The experiments described in this paper demonstrate how this knowledge-light and language-independent approach to machine translation and part-of-speech tagging can result in the fast development of language technology components for a resource-scarce language.

1. Introduction

In recent years quite a few research efforts have investigated the applicability of statistical and machine learning approaches to African language technology. Digitally available corpora, many of which are compiled through web mining (de Schryver, 2002; Scannell, 2007), have proved to be the key component in the development of accurate and robust *performance* models of language.

Unfortunately, linguistically annotated gold-standard corpora, publicly available digital lexicons and basic language technology components, such as morphological analyzers and part-of-speech taggers for African languages are still few and far between. While great advances have been made for many of Africa's larger languages, such as Swahili, Zulu and Hausa, very few research efforts are ongoing for the majority of the continent's 2000+ languages.

This is particularly problematic for many vernacular languages that do not have official status, as they run the risk of being politically and technologically marginalized. An example of such a language is Luo (Dholuo), spoken by about three million people in Kenya, Tanzania and Uganda. This language was recently met with renewed interest, as it is the language of the Kenyan Luo tribe, which constitutes half of United States President Barack Obama's heritage.

This Nilo-Saharan language can undoubtedly be considered as *resource scarce*, as digital text material, as well as the commercial interest to develop it, is largely absent. Apart from the morphological clustering tool described in De Pauw et al. (2007), we are not aware of any published research on Luo in the field of computational linguistics.

This paper explores the applicability of a *knowledge light* approach to the development of Luo language technology. By compiling and exploiting a small corpus of translated texts in Luo, English and Swahili, we will show how a basic machine translation system for the language can be built and how linguistic annotation can be transferred from a *resource-rich* language to the resource-scarce language in question.

In Section 2 we describe the trilingual parallel corpus that forms the building blocks of our approach. We then describe how we can use standard techniques to develop a basic statistical machine translation system for the language pairs in question (Section 3). By further exploiting the automatically induced word alignment patterns, we show in Section 4 how we can project part-of-speech tag annotation from English and Swahili onto Luo. We conclude the paper with a discussion of the advantages and limitations of the approach and outline some options for future research.

2. A Trilingual Corpus English - Swahili - Luo

Unlike many other smaller vernacular languages, such as Gĩkũyũ, Luo has relatively little data available on the Internet. A small web-mined corpus is nevertheless available in the archive of the Crúbadán project (Scannell, 2007). We used this corpus as a seed to perform further web mining, resulting in an updated monolingual Luo corpus of about 200,000 words.

2.1. Parallel Data

A modernized translation of the New Testament in Luo was recently made digitally available on-line by the International Bible Society (2005). Not only does this document effectively double the size of our monolingual Luo corpus, it also enables the compilation of a parallel corpus, by aligning the text with the same documents in other languages. This parallel corpus can consequently be used to power a statistical machine translation system[1].

We used the New Testament data of the SAWA corpus (De Pauw et al., 2009) to construct a small trilingual parallel corpus English - Luo - Swahili. The chapter and verse indications that are inherently present in the data hereby function as paragraph alignment markers, further facilitating automatic sentence alignment, using the approach described in Moore (2002).

[1]Interestingly, an audio-book version of the New Testament is also available from *Faith Comes by Hearing*, opening up the possibility of investigating data-driven speech recognition for Luo as well.

Token counts for the trilingual parallel corpus can be found in Table 1. The corpus was randomly divided into an 80% training set, a 10% validation set and a 10% evaluation set. The validation set allows us to tweak algorithm settings for language modeling, while the evaluation set can be used to measure the accuracy of the machine translation system on previously unseen data.

	English	**Swahili**	**Luo**
New Testament	192k	156k	170k

Table 1: Token counts for trilingual parallel NT corpus

English	You/PP/you have/VBP/have let/VBN/let go/VB/go of/IN/of the/DT/the commands/NNS/command ...
Swahili	Ninyi/PRON/ninyi mnaiacha/V/mnai amri/N/amri ya/GEN-CON/ya Mungu/PROPNAME/Mungu ...
Luo	Useweyo/useweyo Chike/chike Nyasaye/nyasaye mi/mi koro /koro umako/mako ...

Table 2: Factored Data (Mark 7:8)

2.2. Annotation

Standard statistical machine translation tools can be used to *align* the words of the source and target language, by looking for translation pairs in the sentences of the parallel corpus. This typically requires a large amount of data, as it involves scanning for statistically significant collocation patterns and cognates in the orthography. For the language pairs under investigation, the latter information source is limited, as named entities are often transliterated to match the Swahili and Luo pronunciation patterns.

Introducing a layer of linguistic abstraction, however, can aid the alignment process: knowing that a word in the source language is a noun, can help connecting it with a corresponding noun in the target language. Similarly, lemmatization can aid word alignment, as it allows scanning for word types, rather than tokens.

We therefore part-of-speech tagged and lemmatized the English part of the corpus using the TreeTagger (Schmid, 1994). We used the systems described in De Pauw et al. (2006) and De Pauw and de Schryver (2008) to tag and stem the Swahili data. The result is illustrated in the first two rows of Table 2. Each English token consists of three parts: the word form itself, its part-of-speech tag and its lemma. Each Swahili token likewise consists of three parts: the word itself, its part-of-speech tag and its stem.

For Luo, no such tools are available. We therefore made use of the MORFESSOR algorithm (Creutz and Lagus, 2005), which attempts to automatically induce the morphotactics of a language on the basis of a word list. While this fully automated approach obviously generates a flawed morphological description of the language in question, both Koehn et al. (2007) and Virpioja et al. (2007) indicate that even a very rough type of morphological normalization can still significantly aid word alignment of highly inflecting languages.

A lexicon compiled from the monolingual Luo corpus and the training and validation sets of the Luo portion of the parallel corpus was used to train the MORFESSOR algorithm. The output contains a segmented word list, indicating prefixes, stems and suffixes, as well as a segmentation model that allows for the segmentation of previously unseen words. This model was consequently used to morphologically segment the words of the evaluation set. The automatically induced stemming information is then added to the Luo part of the parallel corpus, giving us the type of factored data, illustrated in the last row of Table 2.

A small portion of the evaluation set (about 2,000 words) was also manually annotated for part-of-speech tags. We restricted ourselves to a very small set of 13 part-of-speech tags[2]. This gold-standard data allows us to evaluate the accuracy with which the part-of-speech tags are projected from English and Swahili (Section 4).

3. Machine Translation

In a first set of experiments, we explore the feasibility of statistical machine translation between the two language pairs under investigation: English ↔ Luo and Swahili ↔ Luo. We use MOSES, the standard toolkit for statistical machine translation (Koehn et al., 2007), which incorporates word-alignment using GIZA++ (Och and Ney, 2003) and a phrase-based decoder. The big advantage of MOSES is its ability to efficiently process *factored* data (Table 2) during word alignment, the building of phrase tables and final decoding.

The four translation models were trained using the factored data, described in Section 2.2. For each of the three target languages we built an n-gram language model using the SRILM toolkit (Stolcke, 2002). The value for n was tuned by optimizing perplexity values on the respective validation sets. The Gigaword corpus (Graff, 2003) served as source material for the English language model. For the Swahili language model, the twenty million word *TshwaneDJe Kiswahili Internet Corpus* (de Schryver and Joffe, 2009) was used. To construct the Luo language model, we used the 200,000 word monolingual Luo corpus (Section 2), complemented by the training and validation sets of the parallel corpus.

After training, MOSES used the resulting translation model to translate the evaluation set. By comparing the generated translations to the reference translations, we can estimate the quality of the translation systems using the standard BLEU and NIST metrics (Papineni et al., 2002; Doddington, 2002). The experimental results can be found in Table 3. For the sake of comparison, we have performed each experiment twice: once on just word forms and once on the factored data (indicated by **[F]** in Table3). This allows us to quantitatively illustrate the advantage of using factored data.

[2]The Luo tag set is listed in Table 6 (Addendum).

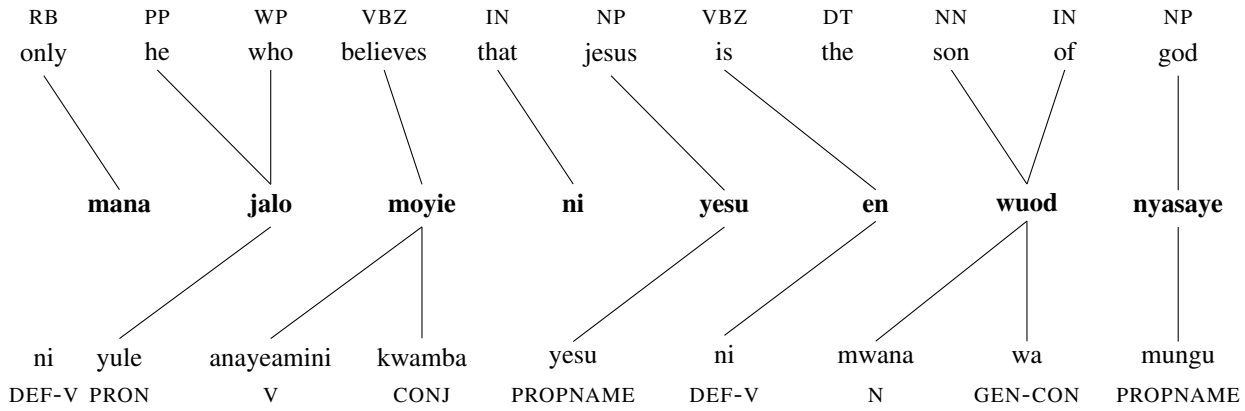


Figure 1: Projection of Part-of-Speech Tag Annotation Using Automatically Induced Word Alignment

The OOV rate expresses the percentage of out-of-vocabulary words, i.e. how many words of the evaluation set of the target language are unknown to the language model. Unsurprisingly, Luo has the highest OOV rate, which can be attributed to the limited amount of corpus data and the morphological complexity of the language.

	OOV rate	NIST	BLUE
Luo → English	4.4%	5.39	0.23
Luo → English [F]		6.52	0.29
English → Luo	11.4%	4.12	0.18
English → Luo [F]		5.31	0.22
Luo → Swahili	6.1%	2.91	0.11
Luo → Swahili [F]		3.17	0.15
Swahili → Luo	11.4%	2.96	0.10
Swahili → Luo [F]		3.36	0.15

Table 3: Results of Machine Translation Experiments

For the language pair English ↔ Luo, NIST and BLEU scores are quite encouraging. While not in the same league as those reported for Indo-European language pairs, the scores are fairly reasonable, considering the very limited amount of training data. We also digitized a translation dictionary English - Luo (Bole, 1997), but adding this information source during word alignment, unfortunately did not lead to significantly higher BLUE or NIST scores during decoding.

The use of factored data seems to have significantly aided word alignment. While it can do nothing to remedy the problem of out-of-vocabulary words, paired T-tests show that the difference in accuracy between the unfactored and factored translation models is statistically significant.

It is important to point out, however, that the encouraging results can be largely attributed to the fact that the system was trained and evaluated on texts from the same domain, i.e. biblical data. Trials on secular data revealed a much higher OOV rate (up to 40% for Luo as a target language) and rather sketchy translations, as illustrated in the translation examples displayed in Table 4.

The language pair Swahili ↔ Luo fairs much worse. While this may seem surprising given the geographical proximity of the two languages, they are genealogically very different, with Swahili being a Bantu language, as opposed to the Nilotic language of Luo.

(1)	**Source**	*en ng'a moloyo piny ? mana jalo moyie ni yesu en wuod nyasaye*
	Reference	*who is it that overcomes the world ? Only he who believes that jesus is the son of god*
	Translation	*who is more than the earth ? only he who believes that he is the son of god*
(2)	**Source**	*atimo erokamano kuom thuoloni*
	Reference	*I am thankful for your leadership*
	Translation	*do thanks about this time*

Table 4: Translation examples for religious data (1) and secular data (2)

The low BLEU and NIST scores can be attributed to the highly inflectional nature of both the source and target language in this case. Despite the fact that word alignment was performed using factored data, there is no morphological generation component for either language in the translation models, so that often an erroneously inflected word form will be generated during decoding, thereby adversely affecting translation quality.

4. Projection of Part-of-Speech Tags

The idea behind projection of annotation is to use the automatically induced word alignment patterns to project the part-of-speech tags of the words in the source language to the corresponding words in the target language. This is illustrated in Figure 1. The Luo sentence in the middle is word aligned with the English sentence at the top and the Swahili sentence at the bottom.

The direct correspondence assumption (Hwa et al., 2002) suggests that the part-of-speech tag of a source language word can be safely projected onto the corresponding word in the target language. The Luo word *moyie* for example can receive the English verbal part-of-speech tag VBZ. Likewise, the word *nyasaye* can be tagged using the Swahili

part-of-speech tag PROPNAME.

In many cases, there is a one-to-many pattern, for example in the alignment of the Luo word *moyie* and the Swahili phrase *anayeamini kwamba*. In such cases, we refer to a predefined tag priority list (see Addendum), which would discard the CONJ in favor of the verbal V tag.

After transferring the source language part-of-speech tags to the target language, a look-up table (see Addendum) maps the projected part-of-speech tags to those of the target tag set (see Footnote 1). We can then compare the projected and converted tag to that of the small gold-standard evaluation set to estimate the feasibility of the approach.

Table 5 displays the results of the experiment. The *Rate* score expresses how many words in the target language received a tag. *Precision* expresses how many of these projected tags are correct. Finally, the *accuracy* score expresses the overall tagging accuracy on the evaluation set.

	Rate	**Precision**	**Accuracy**
English → Luo	73.6%	69.7%	51.3%
Swahili → Luo	71.5%	68.4%	48.9%
Exclusive → Luo	66.5%	78.5%	52.2%
Inclusive → Luo	75.4 %	69.5%	52.4%

Table 5: Results of Projection of Part-of-Speech Tag Annotation Experiments

Projecting part-of-speech tags from English onto Luo works reasonably well. Almost 75% of the words receive a tag. Quite a few words do not receive a part-of-speech tag, either through erroneous word alignment patterns or because there simply was no corresponding word in the target language. Tags are projected with a precision of close to 70%. Overall, more than half of the words in the evaluation set received the correct tag.

The BLEU and NIST scores for the language pair Swahili - Luo (Table 3) were significantly lower than those for the language pair English - Luo. Interestingly, this performance drop is much less significant in this experiment. This further indicates that the problem in translating between these two languages is mostly due to the absence of a morphological generation component during decoding, rather than to fundamental issues in the word alignment phase.

The last two rows of Table 5 show the scores of combined models. The *Exclusive* model only retains a part-of-speech tag for the target word, if both source languages project it. This obviously results in a lower tagging rate, but significantly improves the precision score. The *Inclusive* model likewise combines the two projections, but does not require the two projected tags to be the same. In the case of a projection conflict, the English part-of-speech tag is preferred. This further improves on the Rate and Accuracy scores, but looses out on Precision.

Error analysis shows that many of the tagging errors are being made on the highly frequent function words. This is encouraging, since these constitute a closed-class and mostly unambiguous set of words which can be tagged using a simple look-up table. The translation dictionary (Bole, 1997) can also be used to further complement the part-of-speech tagging database, hopefully further increasing rate and possibly accuracy of the tagger.

Furthermore, the automatically part-of-speech tagged corpus can now be used as training material for a data-driven part-of-speech tagger. Particularly morphological clues can be automatically extracted from this data that can help in tagging previously unseen words.

5. Discussion

To the best of our knowledge, this paper presents the first published results on statistical machine translation for a Nilotic language. A very small trilingual parallel corpus English - Swahili - Luo, consisting of biblical data, was compiled. Morphosyntactic information was added, either by using existing annotation techniques, or by using the unsupervised induction of morphological segmentation patterns.

The results show that using factored data enables the development of a basic machine translation system English - Luo, as well as the projection of part-of-speech tag information for the resource-scarce language of Luo. We hope to replicate this experiment for other vernacular languages as well, for example Gĩkũyũ, which may benefit from its genealogical proximity to Swahili (Wagacha et al., 2006).

One of the bottlenecks in the current approach is the limited morphological analysis and generation capabilities for Swahili and particularly Luo. The unsupervised approach used to stem the Luo data can only serve as a stop-gap measure and a more intricate morphological component will be needed to improve on the current BLEU and NIST scores.

For part-of-speech tagging we will investigate how the Luo corpus, annotated through projection, can be used as training material for a morphologically aware data-driven tagger. Such a part-of-speech tagger may significantly outperform the current approach, as it will be able to process out-of-vocabulary words and smooth over errors introduced by erroneous word alignment patterns or errors made in the annotation of the source language.

The experiments described in this paper serve as a proof-of-concept that language technology components and applications for African languages can be developed quickly without using manually compiled linguistic resources for the target language in question. While we certainly do not claim that the resulting part-of-speech tagger or machine translation system can serve as an end product, we are confident that they can aid in the further development of linguistically annotated Luo corpora.

Acknowledgments and Demonstration

The first author is funded as a Postdoctoral Fellow of the Research Foundation - Flanders (FWO).

The authors wish to thank the reviewers for their constructive and insightful comments. We are also indebted to Kevin Scannell for making his Dholuo data available for research purposes.

A demonstration system and the trilingual parallel corpus is available at `http://aflat.org/luomt`.

6. References

Bole, O.A. (1997). *English-Dholuo dictionary*. Kisumu, Kenya: Lake Publishers & Enterprises.

Creutz, M. & Lagus, K. (2005). Inducing the morphological lexicon of a natural language from unannotated text. In T. Honkela, V. Könönen, M. Pöllä & O. Simula (Eds.), *Proceedings of AKRR'05, International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning*. Espoo, Finland: Helsinki University of Technology, Laboratory of Computer and Information Science, pp. 106–113.

De Pauw, G. & de Schryver, G-M. (2008). Improving the computational morphological analysis of a Swahili corpus for lexicographic purposes. *Lexikos*, 18, pp. 303–318.

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, 9th International Conference*. Berlin, Germany: Springer Verlag, pp. 197–204.

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. (2009). The SAWA corpus: a parallel corpus English - Swahili. 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. 9–16.

de Schryver, G-M. & Joffe, D. (2009). *TshwaneDJe Kiswahili Internet Corpus*. Pretoria, South Africa: TshwaneDJe HLT.

de Schryver, G-M. (2002). Web for/as corpus: A perspective for the African languages. *Nordic Journal of African Studies*, 11(2), pp. 266–282.

Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In M. Marcus (Ed.), *Proceedings of the second international conference on Human Language Technology Research*. San Francisco, USA: Morgan Kaufmann Publishers Inc., pp. 138–145.

Faith Comes by Hearing. (2010). *Luo Audio Bible*. [Online]. Available: http://www.faithcomesbyhearing.com (accessed March 2010).

Graff, D. (2003). *English Gigaword*. [Online]. Available: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2003T05 (accessed March 2010).

Hurskainen, A. (2004). HCS 2004 – Helsinki Corpus of Swahili. Technical report, Compilers: Institute for Asian and African Studies (University of Helsinki) and CSC.

Hwa, R., Resnik, Ph., Weinberg, A. & Kolak, O. (2002). Evaluating translational correspondence using annotation projection. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, USA: Association for Computational Linguistics, pp. 392–399.

International Bible Society. (2005). *Luo New Testament*. [Online]. Available: http://www.biblica.com/bibles/luo (accessed March 2010).

Koehn, Ph., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A. & Herbst, E. (2007). MOSES: Open source toolkit for statistical machine translation. In *Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session*. Prague, Czech Republic: Association for Computational Linguistics, pp. 177–180.

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

Moore, R.C. (2002). Fast and accurate sentence alignment of bilingual corpora. In S.D. Richardson (Ed.), *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users*. Berlin, Germany: Springer Verlag, pp. 135–144.

Och, F.J. & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1), pp. 19–51.

Papineni, K., Roukos, S., Ward, T. & Zhu, W-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Morristown, USA: Association for Computational Linguistics, pp. 311–318.

Scannell, K. (2007). The Crúbadán project: Corpus building for under-resourced languages. In C. Fairon, H. Naets, A. Kilgarriff & G-M Schryver (Eds.), *Building and Exploring Web Corpora - Proceedings of the 3rd Web as Corpus Workshop*. volume 4 of *Cahiers du Cental*, Louvain-la-Neuve, Belgium: Presses Universitaires de Louvain, pp. 5–15.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In D. Jones (Ed.), *Proceedings of the International Conference on New Methods in Language Processing*. Manchester, UK: UMIST, pp. 44–49.

Stolcke, A. (2002). SRILM - an extensible language modeling toolkit. In J.H.L. Hansen & B. Pellom (Eds.), *Proceedings of the International Conference on Spoken Language Processing*. Denver, USA: International Speech Communication Association, pp. 901–904.

Virpioja, S., Väyrynen, J.J., Creutz, M. & Sadeniemi, M. (2007). Morphology-aware statistical machine translation based on morphs induced in an unsupervised manner. In B. Maegaard (Ed.), *Proceedings of the Machine Translation Summit XI*. Copenhagen, Denmark: European Association for Machine Translation, pp. 491–498.

Wagacha, P.W., De Pauw, G. & Getao, K. (2006). Development of a corpus for Gĩkũyũ using machine learning techniques. In J.C. Roux (Ed.), *Networking the development of language resources for African languages*. Genoa, Italy: ELRA, pp. 27–30.

Addendum: Tag Projection Table

Table 6 describes how English part-of-speech tags (Marcus et al., 1993) and Swahili part-of-speech tags (Hurskainen, 2004; De Pauw et al., 2006) are projected onto the Luo tag set. The Luo tag lists expresses tag priority (from top to bottom) in the event of projection conflicts. The Luo tag `particle` was used as a retainer for particles and language specific function words in English and Swahili.

English	Luo	Swahili
NN NNS SYM	noun	ABBR CAP IDIOM N (all concords)
MD VB VBD VBG VBN VBP VBZ	verb	DEF-V V (all concords) IMP
JJ JJR JJS	adjective	A-INFL A-UINFL AD-ADJ ADJ
RB RBR RBS WRB	adverb	ADV NEG
PP PP$ WP$	pronoun	DEM PRON
CD	number	NUM
FW	loan word	AR ENG
IN TO	preposition	PREP
NP NPS	proper name	PROPNAME
UH	exclamation	EMPH EXCLAM RHET
CC	conjunction	CC CONJ
DT EX PDT POS RP WDT WP	particle	AG-PART GEN-CON INTERROG NA-POSS REL SELFSTANDING
LS	punctuation	PUNC

Table 6: Tag Projection Table

SYNERGY: A Named Entity Recognition System for Resource-scarce Languages such as Swahili using Online Machine Translation

Rushin Shah, Bo Lin, Anatole Gershman, Robert Frederking

Language Technologies Institute, School of Computer Science, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh 15213, PA, USA
{rnshah, bolin, anatoleg, ref+}@cs.cmu.edu

Abstract

Developing Named Entity Recognition (NER) for a new language using standard techniques requires collecting and annotating large training resources, which is costly and time-consuming. Consequently, for many widely spoken languages such as Swahili, there are no freely available NER systems. We present here a new technique to perform NER for new languages using online machine translation systems. Swahili text is translated to English, the best off-the-shelf NER systems are applied to the resulting English text and the English named entities are mapped back to words in the Swahili text. Our system, called SYNERGY, addresses the problem of NER for a new language by breaking it into three relatively easier problems: Machine Translation to English, English NER and word alignment between English and the new language. SYNERGY achieves good precision as well as recall for Swahili. We also apply SYNERGY to Arabic, for which freely available NERs do exist, in order to compare its performance to other NERs. We find that SYNERGY's performance is close to the state-of-the-art in Arabic NER, with the advantage of requiring vastly less time and effort to build.

1. Introduction

Online machine translation tools such as (Google Translate™, 2010) and (Microsoft Bing Translator™, 2010) support many languages, including some for which few resources exist and few NLP tools have been developed. In this paper, we focus on one such resource-scarce language, Swahili. As of August 2009, there is no freely available Swahili NER system (Borovikov et al., 2009). However, Google Translate supports two-way translation between Swahili and English. Admittedly, the quality of translation is far from perfect. Many words are simply not translated from Swahili and are carried over as they are into the English version. However, the output can be leveraged to achieve our purposes.

We will show that by performing NER on the translated English text and then matching back the English named entities to words in the Swahili text, our SYNERGY system can perform Swahili NER with a high degree of accuracy. Moreover, SYNERGY can be easily applied to perform NER for other languages too. This is significant because we no longer need to acquire and annotate large amounts of training data ourselves, which is costly and time-consuming for each new language.

Since there is no other Swahili NER system available, we need to apply SYNERGY to another language, for which freely available NER systems do exist, in order to compare its performance to these systems and evaluate the effectiveness of our Machine Translation (MT) based approach to NER. We use Arabic for this purpose, because it is typically considered a 'hard' language for which to do NER, and hence presents a good test for SYNERGY.

The remainder of this paper is organized as follows. In section 2, we discuss relevant prior work done in this area. Section 3 details the various datasets and tools that we have used. In section 4 we describe the development of our SYNERGY NER architecture, presenting in section 5 the results of this system. In section 6, we present examples that illustrate SYNERGY's performance. Finally, in section 7, we discuss our conclusions and plans for future work in this area.

2. Related Work

The problem of NER can be thought of as a subtask of the general task of information extraction. It involves identifying named entities in a text, and in some cases, classifying them according to various types such as persons, locations, organizations, etc. In this paper, we restrict ourselves to the task of identifying named entities only and do not try to predict their types. The NER task was formally defined at the (MUC6, 1995) conference, and this definition was expanded upon at the (CoNLL, 2002) and (CoNLL, 2003) conferences. Various NER systems were evaluated at these conferences, and NER systems produced in subsequent years have also been evaluated on the (CoNLL, 2003) test set. (Zhao et al., 2007) and (Huang, 2006) have shown that NER can be used to help machine translation; it is important to note that our work focuses on showing the reverse case, which we believe to be novel. (De Pauw et al., 2009) illustrate the difficulties inherent in English-Swahili word alignment during their development of the SAWA English-Swahili parallel corpus. In addition, (De Pauw et al., 2006) have also developed a memory-based part-of-speech (POS) tagger for Swahili, using the Helsinki Corpus of Swahili. (Benajiba and Rosso, 2008) use Conditional Random Fields (CRFs) to perform Arabic NER and report a best-case F1 score of 0.792. (Benajiba et al., 2008) use Support Vector Machines (SVMs) and CRFs with optimized feature sets to perform Arabic NER and report a best-case F1 score of 0.835.

3. Resources

We use (Google Translate™, 2010) as our online Machine Translation system for Swahili and Arabic text. We attempted to use (Microsoft Bing Translator™, 2010) (which

doesn't support Swahili yet) for Arabic but were unable to do so because currently it leaves many words (including many Named Entities) untranslated and this leads to poor NER performance. We would have liked to use additional MT systems, but for these languages, we could find only two freely available machine translation tools. Although various other MT systems are available online, they mainly support translation only between English and European languages. By contrast, Google Translate and Bing Translator not only support a wide variety of languages but are also steadily adding support for new languages over time. Currently, no other African languages are supported by these or any other freely available MT systems that we know of, but a key advantage of SYNERGY is that as soon as new languages become available on any MT system, SYNERGY can be used with minimal modification to perform NER for these new languages. SYNERGY does not necessarily have to use online MT systems, it can also use MT systems automatically generated from parallel data using freely available toolkits such as (Moses, 2010).

Named Entity (NE) labeled data is scarce for both languages. For Swahili, we use a test set of slightly more than 27000 tokens from the Helsinki Corpus of Swahili (HCS, 2004). For Arabic, we use a test set of 25000 tokens from the ANERcorp corpus (Benajiba et al., 2007). For Named Entity recognition, we use two well-known systems: Stanford's Conditional Random Field (CRF) based NER system (Finkel et al., 2005) and UIUC's Learning-Based Java (LBJ) Named Entity Tagger (Ratinov and Roth, 2009). We use only off-the-shelf named entity recognizers, to demonstrate that a good NER system can be developed for new languages in short order, with minimal training data. We also use the (CoNLL, 2003) dataset.

Post-processing improves the quality of the system, but is not required for its operation. For post-processing, we make use of the Swahili-English dictionary by (Kamusi Project, 2010), and the Linguistic Data Consortium's Arabic TreeBank (Maamouri et al., 2005).

4. SYNERGY Architecture

As described in our introduction, SYNERGY performs NER on a new language, which we shall henceforth refer to as the source language, by translating source language text to English, running off-the-shelf state-of-the-art NER systems on the translated English text, and then matching back the English named entities to words in the source language. It also performs some post-processing to improve NER performance. This section describes each of these steps in detail.

4.1. Translation to English

SYNERGY uses a Perl interface to Google Translate. It translates Swahili and Arabic documents to English by translating each sentence individually. Unlike its web interface, the Google Translate API does not accept sentences that are longer than a certain length (which in our testing we found to be 700 characters). Most corpora, including the ones we use, are composed of newswire articles, where sentence lengths frequently exceed that number. Therefore, when the system encounter sentences whose length exceeds the maximum amount, it performs preprocessing and splits them into smaller sentences of acceptable size. Although splitting sentences in this manner may consequently lead to poor translation results, we find that named entities are not affected and hence this does not present a problem for our system. We also observe that the Arabic translation is of a better quality than the Swahili translation; in particular, many Swahili words are simply not recognized and hence not translated. However, named entities are not strongly affected by this issue either. Let d_{src} denote a source language document; d_{en} denotes its translated English version.

4.2. NER on Translated English Text

For English NER, the freely available systems developed by Stanford and UIUC achieve a high degree of accuracy and represent the current state-of-the-art in the field. We use both these systems in SYNERGY. We initially also tested the (LingPipe™, 2010) and (AlchemyAPI™, 2010) NER systems, but found no improvement in precision or recall from adding either of these systems to SYNERGY.

The Stanford and UIUC systems do not return identical results: Approximately 3% of the named entities are correctly identified by only one of the two systems. We exploit this discrepancy and create a combined NER system called Union that achieves a higher F1 score than either of the individual systems. Formally, the Union system performs NER as follows: A token t is categorized as being part of a named entity if it is classified as being part of a named entity by either the Stanford or UIUC system. We define the system in this manner to try to maximize recall, with the expectation that the increase in recall will compensate for a drop in precision. The Stanford and UIUC systems have their own distinct tokenizers, and there are many subtle differences in the way they treat non-word characters, dates, URLs, etc. So, synchronizing the outputs of the two systems proves to be a non-trivial task.

Table 1 gives a summary of the performance of these systems. The systems were not retrained; we used their off-the-shelf versions. The systems were evaluated on the test subset of the (CoNLL, 2003) dataset. We performed the evaluation on a per-token basis, as opposed to a per-entity basis. This eliminates the need to address potential ambiguities that may be caused if a token is classified as being a part of different named entities by different NER systems. As expected, we find that the Union system gets a slightly higher F1 score than both its component systems. Therefore, Union is the default NER module used in SYNERGY.

System	Precision	Recall	F1
Stanford	0.962	0.963	0.963
UIUC	0.968	0.964	0.966
LingPipe	0.624	0.554	0.587
AlchemyAPI	0.680	0.506	0.580
Union	0.952	0.985	0.968

Table 1: Performance of various NER systems

We run the Union NER system on the translated English document d_{en} and extract a list L_{ne} of named entities and the locations at which they occur in d_{en}.

4.3. Alignment of Named Entities with Source Language Words

This is the final and most challenging operation performed by the SYNERGY system. After creating the list L_{ne} of named entities in the translated English document d_{en}, we need to find the words in the source language document d_{src} that match the entities in L_{ne}. We try two different algorithms: brute force alignment, and word alignment using GIZA++.

4.3.1. Brute Force Alignment

In our initial approach, we translated each word w_{en} in L_{ne} back to a source language word, which we denote as w_{re-src}, and proceeded to search in the source document d_{src} for a word w_{src} that matches w_{re-src}, by looking at a window of words, centered around the location that w_{en} occurs in the translated English document d_{en}. If a word w_{src} matching w_{re-src} was found, it was tagged as being part of a named entity. If no such match was found for this particular w_{re-src} in this window, w_{re-src} was discarded and we repeated this search with the next word in L_{ne}. The parameter in this search algorithm was the size s of the window of words that we examine for each w_{re-src}. We found that this initial approach gave very poor results for both our Swahili and Arabic test sets. It found matches for only 40% of the named entity words in L_{ne}. There was an increase in the number of matches found with an increase in s, until $s = 30$, after which there was no increase. Subsequent error analysis shows that a major cause of this poor performance is that when a source language word w_{src} is translated to an English word w_{en} which is then retranslated to a word w_{re-src} in the original language, it is often the case that $w_{src} \neq w_{re-src}$. For Arabic, additional mismatches are introduced while converting text from UTF-8 to Latin encoding and back.

Hence, we try a different version of this algorithm: After creating L_{ne}, we create a new English document d_{en-tl} by translating each word in the source language document d_{src} one at a time to English. We now search in d_{en-tl} for matches for each word in L_{ne}. To deal with the possibility that a source word w_{src} may produce multiple English words, we keep pointers from each English word to the source word w_{src} that produces it. The search proceeds in the same manner as outlined in the previous algorithm: by considering a window of words of size s centered at the location of the named entity word in d_{en}. If a word w_{en-tl} is found in d_{en-tl} that matches a word in L_{ne}, we label the source language word w_{src} that produces w_{en-tl} as being part of a named entity.

This second version produces much better results and finds matches in the source document d_{src} for 75% of the named entity words. As in the previous algorithm, the number of matches found increases with an increase in s, until $s = 30$. Furthermore, since this algorithm compares English words, we can use stemming to perform a more sophisticated comparison than simple equality testing. We use the well-known (Porter, 1980) stemmer. This augmented version of our brute force alignment algorithm finds matches for 77% of the named entity words. It is important to emphasize that this figure only refers to the fraction of named entity words for which words in the source language are found; it does not address the issue of whether the source language words are in fact the correct ones that produced the named entity words. If many of the matched source language words are incorrect, this results in poor overall NER performance in spite of a large fraction of named entities being matched. Hence, the fraction of matches found is ultimately a metric of limited utility and cannot be reliably construed as a predictor of overall NER performance.

4.3.2. Word Alignment using GIZA++

The GIZA++ package (Och and Hey, 2003) is the state-of-the-art for the task of word alignment in the field of statistical Machine Translation. It contains a number of statistical alignment models specifically designed for word alignment between different languages. We modify SYNERGY to use GIZA++ to perform word alignment. GIZA++ takes an input language corpus and an output language corpus, and automatically generates word classes for both languages using these corpora. It then finds for each sentence in the input language corpus the most probable alignment of the corresponding output language sentence (also known as the Viterbi alignment). A detailed description of the various statistical models used in word alignment and available in GIZA++ can be found in (Och and Hey, 2003). For our task, the translated English documents serve as the input corpus and the source language documents as the output corpus. In order to use GIZA++, SYNERGY first preprocesses these corpora to make sure that they are sentence-aligned.

4.4. Post-Processing

We compile exhaustive part-of-speech tagged English, Swahili and Arabic dictionaries, and divide all of these into Named Entity (NE) and Non-Named Entity (NNE) sections, depending on whether a word may be used as a proper noun or not. Of course, many words may occur in both sections, since in every language many proper nouns are often derived from common words. The post-processing (PP) module scans through the NE-annotated source language documents returned by the previous modules and applies the following two rules:

- If a word annotated as a Named Entity occurs in the NNE section of a dictionary but not in its NE section, the word's annotation is removed.
- If a word not annotated as a Named Entity occurs in the NE section of a dictionary but not in its NNE section, the word is accordingly annotated as a Named Entity.

Even though the above rules are relatively simple, they nevertheless lead to significant improvements in NER performance.

5. SYNERGY Results

Table 2 shows the results of SYNERGY on Swahili data. We use the gold-standard labels in the HCS corpus to check the accuracy of our labeling. Evaluation is performed on a per-token basis, as described earlier. We find

that without post-processing, brute force alignment performs slightly better than GIZA++ based alignment, but with post-processing the situation is reversed and GIZA++ based alignment performs much better. Moreover, post-processing leads to major improvements in the F1 score for both algorithms. In particular, the "GIZA++ with Post-Processing" version of SYNERGY yields an F1 score of 0.815, which we believe is very good for a first effort in the field of Swahili NER.

Version	Prec	Recl	F1
Brute Force w/o PP	0.676	0.694	0.685
Brute Force with PP	0.818	0.704	0.757
GIZA++ w/o PP	0.534	0.900	0.670
GIZA++ with PP	0.754	0.886	0.815

Table 2: SYNERGY Results for Swahili NER

Since there are no other freely available Swahili NER systems, in order to measure the effectiveness of SYNERGY's MT-based approach to NER, we compare SYNERGY's results on Arabic data with the results obtained by state-of-the-art systems in Arabic NER. The results are shown in Table 3. We use the gold-standard labels in ANERcorp to check the accuracy of our labeling and perform evaluation on a per-token basis. We must point out here although (Benajiba and Rosso, 2008) uses the same ANERcorp corpus that we use, (Benajiba and Rosso, 2008) does not, and we do not have access to the testing data used by them. As a result, this is not an exact comparison. However we wish to obtain a larger picture of the effectiveness of the approach used by SYNERGY, and this comparison serves that purpose reasonably.

System	Prec	Recl	F1
SYNERGY BF w/o PP	0.680	0.502	0.578
SYNERGY BF with PP	0.848	0.600	0.703
SYNERGY GIZA++ w/o PP	0.530	0.702	0.603
SYNERGY GIZA++ with PP	0.761	0.817	0.788
(Benajiba and Rosso, 2008)	0.869	0.727	0.792
(Benajiba et al., 2008)	N/A	N/A	0.835

Table 3: SYNERGY Results for Arabic NER and comparison with other systems

We find that the F1 score achieved by the "GIZA++ with Post-Processing" version of SYNERGY comes quite close to the scores achieved by the state-of-the-art systems. In addition, SYNERGY has the advantage of requiring vastly less time and effort to build and adapt to new languages than other systems. It does not require collecting and annotating an NER training corpus for each new language.

There are three possible types of errors that SYNERGY may produce:

- Named Entities that are lost during translation from the source language to English.
- Errors made by the English NER module of SYNERGY
- A correctly recognized English NE word that gets mapped to the wrong word in the source language document during alignment

It would be interesting to analyze the distribution of SYNERGY's errors across each of these categories. However, that would require the presence of NE gold standard data for the English translations of our Swahili and Arabic test sets in addition to native gold standard data, and since there are no other known systems that employ an MT based approach to NER, such data is not currently available. Therefore, we are unable to perform this analysis.

6. Examples

We illustrate the performance of each stage of SYNERGY on sample Swahili and Arabic sentences. The following NE tagged Swahili sentence is taken from a newswire article in the HCS corpus (Here, we use italics to indicate true Named Entities):

"*Lundenga* aliwataja mawakala ambao wameshatuma maombi kuwa ni kutoka mikoa ya *Iringa Dodoma Mbeya Mwanza* na *Ruvuma*."

. Using Google Translate, SYNERGY translates this sentence to the following English one:

"Lundenga mentioned shatuma agents who have a prayer from the regions of Dodoma Iringa Mwanza Mbeya and Ruvuma."

SYNERGY then adds the following NE labels to this sentence (Here, we use italics to indicate NE labels added by our system):

"*Lundenga* mentioned shatuma agents who have a prayer from the regions of *Dodoma Iringa Mwanza Mbeya* and *Ruvuma*."

Finally, after applying GIZA++ alignment and post-processing, the sentence returned as SYNERGY's final output is:

"*Lundenga* aliwataja mawakala ambao wameshatuma maombi kuwa ni kutoka mikoa ya *Iringa Dodoma Mbeya Mwanza* na *Ruvuma*."

Now, consider the following NE tagged Arabic sentence taken from the ANERcorp corpus, shown here after being transliterated according to the Buckwalter encoding (Buckwalter, 2002):

"n$yr AlY h*h AlZAhrp l¿nhA t$kl Aljw Al*y yEml fyh fryq Alr}ys *jwrj bw$ wrAys* wAlsfyr *jwn bwltwn* fy *AlAmm AlmtHdp* , wAl*y symyz Alkvyr mn AlEnASr Alty qd yqdmhA bED AlAwrwbyyn lrdE tmAdy *AlwlAyAt AlmtHdp* fy AlAstvmAr bqrAr AlEdwAn AlAsrA}yly ElY *lbnAn*."

The NE tagged English translation of this sentence produced by SYNERGY is:

"Refer to this phenomenon because it is the atmosphere with a team of President *George W. Bush Rice* and Ambassador *John Bolton* at the *United Nations* which will recognize a lot of elements that might make some *Europeans* to deter the persistence of the *United States* decision to invest in the *Israeli* aggression on *Lebanon*."

Finally, SYNERGY maps the Named Entities in this sentence back to the original Arabic sentence and gives the following output:

"n$yr AlY h*h AlZAhrp l¿nhA t$kl Aljw Al*y yEml fyh fryq Alr}ys *jwrj bw$ wrAys wAlsfyr jwn bwltwn* fy *AlAmm AlmtHdp* , wAl*y symyz Alkvyr mn AlEnASr Alty qd yqdmhA bED AlAwrwbyyn lrdE tmAdy *AlwlAyAt AlmtHdp* fy AlAstvmAr bqrAr AlEdwAn AlAsrA}yly ElY *lbnAn*."

From these examples, we see that although there are many inaccuracies in both the Swahili and Arabic translations, a vast majority of NE words are preserved across translation and successfully recognized by an English NER system. Moreover, performing NER in English helps us to avoid the difficulties inherent in native Swahili and Arabic NER, e.g. ambiguous function words, recognizing clitics, etc. In other words, SYNERGY addresses the problem of NER for the source language by breaking it into three relatively easier problems: Machine Translation to English, English NER and word alignment between English and the source language.

7. Conclusion and Future Work

We achieve best-case NER F1 scores of 0.788 for Arabic and 0.815 for Swahili. The F1 score for Arabic comes quite close to the state-of-the-art achieved by (Benajiba et al., 2008) and (Benajiba and Rosso, 2008), and the F1 score for Swahili is impressive. Moreover, building SYNERGY and adapting it to new languages is much less expensive than creating an NER training corpus from scratch. Of course, someone had to build the translation system first, but these exist for many more languages than do NER systems. We believe ours is the first freely available Named Entity Recognition system for Swahili[1], and we hope it will be a valuable tool for researchers wishing to work with Swahili text. We intend to use SYNERGY to perform NER for various other resource-scarce languages supported by online translators.

One important language technology that is even less widely available than Named Entity Recognition for many languages is Co-Reference Resolution (CRR), and this is a natural problem to approach using SYNERGY. But unlike NER, in the case of CRR, to test SYNERGY's output we would need a parallel English-Swahili corpus. The SAWA corpus (De Pauw et al., 2009) has not been released yet, and we are not aware of any other such parallel corpus. With its release, this need will be addressed. In the future, we plan to augment SYNERGY to perform CRR for Swahili and also other languages.

[1] `http://www.cs.cmu.edu/~encore/`

8. References

Alchemy API™ (2010). [Online]. Available: http://www.alchemyapi.com (accessed March 2010).

Alias-i. LingPipe™ (2010). [Online]. Available: http://alias-i.com/lingpipe (accessed March 2010).

Benajiba, Y. & Rosso, P. (2008). Arabic Named Entity Recognition using Conditional Random Fields. In *Proceedings of Workshop on HLT and NLP within the Arabic World. LREC 2008*.

Benajiba, Y., Diab, M. & Rosso, P. (2008). Arabic Named Entity Recognition using Optimized Feature Sets. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*.

Benajiba, Y., Rosso, P. & Ruiz, J. M. B. (2007). ANERsys: An Arabic Named Entity Recognition system based on Maximum Entropy. In *Proceedings of Computational Linguistics and Intelligent Text Processing*, pp. 143–153.

Buckwalter, T. (2002). Arabic Transliteration. [Online]. Available: http://www.qamus.org/transliteration.htm (accessed March 2010).

Borovikov, A., Borovikov, E., Colquitt, B. & Summers K. (2009). The EDEAL Project for Automated Processing of African Languages. In *Proceedings of MT Summit XII*. Ottawa, Canada.

CoNLL. (2002). CoNLL-2002 shared task: Language-independent named entity recognition. [Online]. Available: http://www.cnts.ua.ac.be/conll2002/ner (accessed March 2010).

CoNLL. (2003). CoNLL-2003 shared task: Language-independent named entity recognition. [Online]. Available: http://www.cnts.ua.ac.be/conll2003/ner (accessed March 2010).

De Pauw, G., de Schryver, G. & Wagacha, P. W. (2006). Data-driven part-of-speech tagging of Kiswahili. In *Proceedings of Text, Speech and Dialogue, 9th International Conference.* Berlin, Germany: Springer Verlag, pp. 197–204.

De Pauw, G., Wagacha, P. W. & de Schryver, G. (2009). The SAWA Corpus: a Parallel Corpus English - Swahili. In *Proceedings of the First Workshop on Language Technologies for African Languages (AfLaT 2009), Association for Computational Linguistics*. Athens, Greece, pp. 9–16.

Finkel, J. R., Grenager, T. & Manning, C. (2005). Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*.

Google Translate™ (2010). [Online]. Available: http://www.translate.google.com (accessed March 2010).

Helsinki Corpus of Swahili (HCS). (2004). Institute for Asian and African Studies (University of Helsinki) and CSC Scientific Computing Ltd.

Huang, F. (2006). *Multilingual Named Entity Extraction*

And Translation From Text And Speech. Phd Thesis, LTI, Carnegie Mellon University.

Kamusi Project. (2010). [Online]. Available: http://www.kamusiproject.org (accessed March 2010).

Maamouri, M., Bies, A., Buckwalter, T., Jin, H. & Mekki, W. (2005). Arabic Treebank: Part 3 (full corpus) v 2.0 (MPG + Syntactic Analysis). Linguistic Data Consortium, Philadelphia.

Microsoft Bing Translator™ (2010). [Online]. Available: http://www.microsofttranslator.com (accessed March 2010).

Moses: Open source toolkit for statistical machine translation. [Online]. Available: http://www.statmt.org/moses/ (accessed March 2010).

MUC6. (1995). *Proceedings of the Sixth Message Understanding Conference (MUC6)*. DARPA, Morgan-Kaufmann.

Och, F. J. & Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1), pp. 19–51.

Porter, M. F. (1980). An Algorithm for Suffix Stripping. *Program*, 14(3), pp. 130–137.

Ratinov, L. & Roth, D. (2009). Design Challenges and Misconceptions in Named Entity Recognition. *CoNLL 2009*.

Zhao, B., Bach, N., Lane, I. & Vogel, S. (2007). A Log-linear Block Transliteration Model based on Bi-Stream HMMs. In *Proceedings of NAACL-HLT 2007*. Rochester, NY, USA.

Processing Parallel Text Corpora for Three South African Language Pairs in the Autshumato Project

Hendrik J. Groenewald, Liza du Plooy
Centre for Text Technology (CTexT™),
North-West University (NWU), Potchefstroom Campus, South Africa
{handre.groenewald, liza.duplooy}@nwu.ac.za

Abstract

Multilingual information access is stipulated in the South African constitution. In practise, this is hampered by a lack of resources and capacity to perform the large volumes of translation work required to realise multilingual information access. One of the aims of the Autshumato project is to develop machine translation systems for three South African languages pairs. These machine translation systems are envisaged to serve as tools that will enable the human translator to do his/her work more efficiently and thereby indirectly contributing to improving access to multilingual information services. This paper describes the gathering and processing of parallel corpora for the purpose of developing machine translation systems. We highlight the difficulties in gathering parallel corpora and provide detailed information about the anonymisation of parallel corpora. We also describe various data processing methods to prepare the parallel corpora for use in the development of machine translation systems.

1. Introduction

South Africa is known for diversity in cultures, languages and religious beliefs. The South African constitution (Republic of South Africa, 1996) recognises no fewer than eleven official languages. The South African government aims to meet the constitutional obligations on multilingualism, by providing equitable access to government services, knowledge and information (Republic of South Africa, 2003). Promoting access to information in all of the official South African languages is a noble idea, but it is difficult to apply in practise, due to large volumes of translation work that are required to realise this.

It is therefore no surprise that government translation agencies such as the National Language Service (NLS) are struggling to keep up with large volumes of translation work. This problem is aggravated by the fact that machine-aided translation tools and resources are often not available for South African languages. Another contributing factor is the fact that proprietary computer-assisted translation software suites are not widely used by government translation agencies due to high licensing costs and incompatibilities with open source computer operating systems.

The consequence of this is that English often acts as the *lingua franca* (although it is the mother tongue of only 8.2% of the South African population (Van der Merwe and Van der Merwe, 2006)), with the unfortunate result that the remaining indigenous South African languages are yet even further marginalised. The dominance of English has the resulting effect that a large number of South African citizens are deprived of their constitutional right of access to information in their language of choice.

Innovative solutions are required to ensure multilingual access to information in South Africa. The South African Government is aware of the potential impact that Human Language Technology may have in this regard and is therefore involved in a number of initiatives and projects to promote multilingualism through the use of Human Language Technology. One such a project is the Autshumato Project, which is discussed in Section 2. The rest of this paper is organised as follows: In Section 3 we provide more information about the machine translation approach that is followed in the Autshumato project. Section 4 focuses on data providers, while Section 5 provides detailed information about the anonymisation system that we have developed. Further processing of parallel corpora is discussed in Section 6. The paper concludes in Section 7 with some directions for future work.

2. Autshumato Project

The Autshumato[1] project was commissioned in 2006 by the South African Department of Arts and Culture. The aim of the project is to develop open source machine-aided translation tools and resources for South African languages. One of the interesting outcomes of the Autshumato Project is the development of machine translation systems for three South African languages pairs, namely English – > isiZulu, English – > Afrikaans, and English – > Sesotho sa Leboa. The purpose of these machine translation systems is to provide government translators with a tool that may help them to perform their work more efficiently and increase consistency and productivity. Obtaining and processing parallel data to develop machine translation systems for three South African Language pairs is the central theme of this paper.

3. Machine Translation

Interest in machine translation as an area of research started to gain momentum after World War II. The complexity of machine translation was grossly

[1] Autshumato was a Khoi-khoi leader that worked as in interpreter between Europeans and the Khoi-khoi people during the establishment of the Dutch settlement at the Cape of Good Hope in the 17th century (Giliomee and Mbenga, 2007).

underestimated in those early years and together with the publication of the ALPAC report (ALPAC, 1966) resulted that interest in (and funding of) machine translation research gradually decreased. Despite the pessimism of the ALPAC report, machine translation research continued and a large number of machine translation systems have been developed over the years with varying degrees of success (De Pauw *et al.*, 2009).

The rapid expansion of the internet and computer processing power in the 1990's, together with the resulting increase in the availability of electronic data led to an increased interest in the use of the statistical machine translation (SMT) method. The performance of statistical machine translation systems is to a large extent dependent on the amount of and quality of parallel text corpora available.

All of the official South African languages, except English, are considered to be resource scarce languages. The availability of parallel text corpora for these languages is limited. The lack of parallel text corpora implies that alternative methods to the mere addition of parallel data must be sought to improve the quality of the machine translation systems that are being developed in the Autshumato project. The approach followed in the Autshumato project is to utilise statistical machine translation to create baseline machine translation systems that can be augmented with rules based on expert linguistic knowledge. It was decided that SMT would form the basis for the development of machine translation systems in the Autshumato Project for the following reasons:

1) SMT is currently the preferred approach of numerous industrial and academic research laboratories.
2) State-of-the art open source SMT toolkits are readily available.
3) Less expert linguistic knowledge is required to create a working baseline system in comparison to rule-based systems.

4. Data Providers

As mentioned in the previous section, it is difficult to obtain parallel text corpora for the language pairs for which machine translation systems are currently being developed in the Autshumato project. Apart from government sources, few other sources of parallel text corpora exist. It is often necessary to approach private translation companies and freelance translators to obtain parallel text corpora. The unavailability of parallel text data can be ascribed to the following reasons:

1) Computer-assisted translation software suites are not widely used, with the result that translation memories are not readily available for the indigenous South African languages.
2) Lack of publications like books, newspapers, magazines and websites in the indigenous South African Languages.
3) Lack of sound document management practices, which makes it difficult to obtain parallel documents from translators.
4) Unwillingness of translators and private companies to make their data available for purposes of machine translation research.

5. Text Anonymisation

5.1 Introduction

As mentioned in the previous section, several potential data providers are unwilling to provide their parallel corpora for the development of machine translation systems. One of the main reasons for this is that their corpora often contain confidential information. For this reason, we are developing text anonymisation software to identify confidential information in parallel corpora. We will then anonymise the corpora by replacing the entities conveying the confidential information, by randomly selected entities from the same category.

Text anonymisation can be seen as a subcategory of named entity recognition. It differs from named entity recognition in the respect that it focuses only on the entities in text containing information of a confidential nature.

Neamatullah *et al.* (2008) did research on the anonymisation of medical records and developed a *Perl*-based de-identification software package that can be used on most free text medical records like notes from nurses. It was developed using gazetteers (i.e. dictionaries containing specialised entity lists), regular expressions and simple rules. This software was exclusively developed for use in the medical domain.

Text anonymisation software is being developed as part of this research for English, Sesotho sa Leboa, isiZulu and Afrikaans (the four languages involved in the development of the Autshumato machine translation systems). Research on named-entity recognition for Afrikaans was done by Puttkammer (2006) as part of the development of an Afrikaans tokeniser. A named-entity recogniser for isiZulu was developed as a part of the project *REFLEX: Named Entity Recognition and Transliteration for 50 Languages* (Sproat, 2005). For Sesotho sa Leboa research has been done on the recognition of spoken personal names as a part of the *Interactive Voice Recognition (IVR)*-system (Modipa, Oosthuizen & Manamela, 2007).

5.2 Method and Implementation

A rule-based approach was followed in the development of the anonymisation system, using gazetteers, regular expressions and simple context rules. A proof-of-concept system was first developed for Setswana, Afrikaans and English. This system has been recently adapted to include isiZulu and Sesotho sa Leboa. We aimed to make the anonymisation system as language independent as possible, to ensure rapid expansion to the remaining official South African languages in the future. The operation of the anonymisation system is as follows: Firstly, entities with a predictable form like dates,

addresses, contact numbers, email addresses and URL's are marked with regular expressions. Next, all the words that appear in any of the gazetteers are marked. Lastly, context rules are applied to find entities that do not appear in any of the gazetteers. A logic flow of the various steps in the anonymisation process is shown in Figure 1. Each of these steps will now be discussed in more detail in the rest of Section 5.2.

5.2.1 Regular Expressions

As mentioned in the previous section, regular expressions are used to identify entities with a predictable form. Dates are recognised by combining regular expressions with lists containing words that are regularly found in dates, like the names of months and days, in the different languages (see step 3 in Figure 1). Dates in formats like *2009-09-27, 05/12/2010* and *16 Feberware 1978* are also identified. To recognise addresses, lists with location names are combined with regular expressions. Due to the numerous valid formats of addresses, several regular expressions are implemented to recognise the different formats and types of addresses. Contact numbers are written differently in different parts of the world, but for the goal of this research specific attention has been given to numbers in the formats commonly used in South Africa.

5.2.2 Gazetteers

The gazetteers were assembled from a variety of sources and are continuously updated. All of the words in the text that appear in any of the gazetteers are marked by the system (see step 6 in Figure 1).

First Names and Surnames

The list of first names of Puttkammer (2005) was expanded by the addition of names that are found amongst to various ethnic groups in South Africa. The list currently consists of 8,853 unique names. The list of surnames used by Puttkammer (2005) consists of 3,174 surnames. The list used by Neamatullah et al. (2008) is freely available on the web and was added to the existing list of Puttkammer. Surnames were also extracted from the address book of the North-West University, available on the university's website (North-West University, 2004). Several of the first names and surnames contained in the above-mentioned lists are also valid words in the indigenous languages when not used in the "first name or surname sense". For example, the Setswana sentence *Ke na le khumo* means "I have wealth", but *Khumo* can also be a first name. These words were removed from the list by comparing it to lexica consisting of valid lower case words of the different languages. The final list consists of 81,711 surnames.

Company, organisation and product names

Once again the list used by Puttkammer (2006) was expanded to be used in the anonymisation system. A small amount of company names were obtained from the website of the Johannes Stock Exchange (JSE, 2005) and appended to the list of Puttkammer (2006). This list needs to be expanded considerably in future.

5.2.3 Context Rules

Context rules are applied to find entities that do not appear in any of the gazetteers (step 9 in Figure 1). An example of a context rule is: "if a word following a word that has been tagged as a first name starts with a capital letter, and if that word does not appear in one of the lexica, it is considered to be a personal name".

The last step is to compare every capital letter word against lowercase lexica. If the particular word does not appear in any of the lexica, it is considered to be a named entity (see step 11 in Figure 1). This has a positive effect on the recall of the system, but it also has a negative effect on precision. For the purposes of our research, high recall is more important than high precision, since we do not want any named entities that can convey confidential information to go unnoticed by the system.

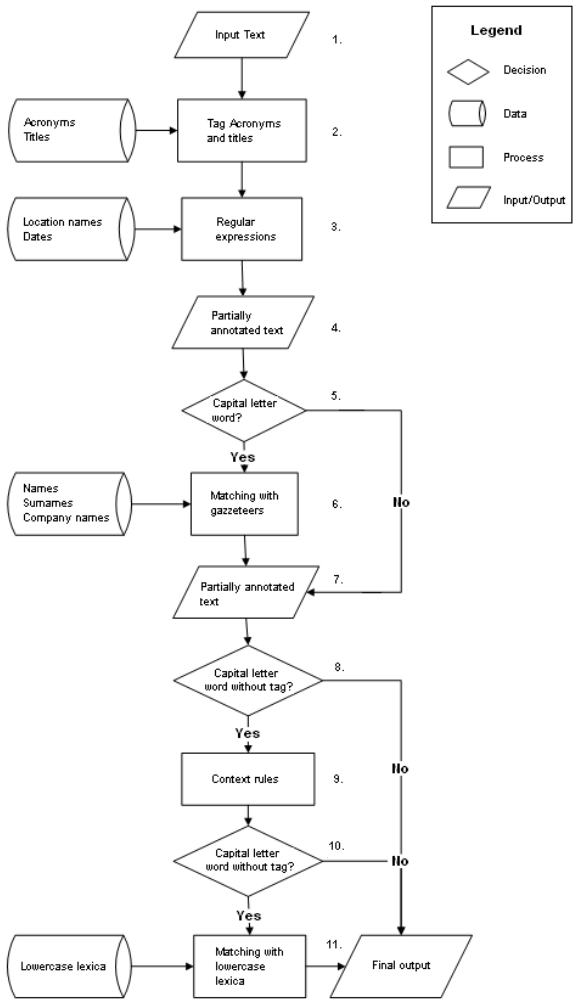


Figure 1: Logical flow of the anonymisation system

5.3 Results

The anonymisation system was tested on data from the government, university and agricultural domains. The test data were annotated manually. The results obtained for Setswana, Afrikaans and English are shown in Figure 2. Setswana obtains remarkable lower scores than Afrikaans

and English. One reason for this is the fact that many first names and surnames in Setswana (and many other indigenous South African and African languages) are also valid words when not used in the "first name or surname" sense of the word (see Section 5.2.2).

	English	Afrikaans	Setswana
Recall	0,796	0,799	0,777
Precision	0,832	0,805	0,756
F-Score	0,814	0,802	0,767

Table 1: Results obtained by the anonymisation system

6. Processing of Data

6.1 Sentencisation

After the parallel text corpora is obtained from the various data suppliers, the data must be sentencised (i.e. split into sentences) before it can be aligned at sentence level. Sentencisation algorithms for all the official South African languages have been developed as part of the Autshumato project. These sentencisation algorithms are based on a combination of language specific rules and abbreviation lists.

6.2 Alignment

Once the data have been sentencised, the alignment process can begin. Microsoft's bilingual sentence aligner (Moore, 2002) is used to automatically align sentences. The sentences that are not aligned during the automatic alignment process are manually aligned. After the parallel corpora have been aligned at sentence level, it is ready to be used to train statistical machine translation systems. The current number of aligned text units is shown in Table 2.

Language Pair	Aligned Text Units
English-Afrikaans	439,890
English-Sesotho sa Leboa	50,238
English-isiZulu	92,560

Table 2: Number of aligned text units

7. Conclusion and Future Work

In this paper we motivated the importance of developing machine translation systems for South African languages. We described the problematic aspects associated with the gathering of parallel corpora for resource scarce South African languages. We also focused extensively on the anonymisation algorithm that has been developed to motivate data suppliers to make their data available to the Autshumato project. We concluded with a brief description of the processing steps that are carried out on the data after the anonymisation process.

Future work includes the improvement of the performance of the current anonymisation system. Possible ways to achieve this is to expand the gazetteers (especially the company and product names), "cleaning" the gazetteers by removing ambiguous words, adding more context rules and refining the existing rules. We also want to augment the current system with machine learning techniques to determine if further improvements in recall and precision can be obtained.

8. References

ALPAC. (1966). *Languages and machines: Computers in translation and linguistics.* A report by the Automatic Language Processing Advisory Committee, Division of Behavioural Sciences, National Academy of Sciences, National Research Council. Washington, D.C.: National Academy of Sciences, National Research Council. Publication 1416.

Bekker, A., Groenewald, H.J. & McKellar, C.A. (2009). *Progress Report: DAC Project B: Machine Translation Systems*. Potchefstroom: CTexT.

De Pauw, G., Wagacha, P.W. & de Schryver, G. (2009). The SAWA Corpus: a Parallel Corpus English-Swahili. In *Proceedings of the EACL 2009 Workshop on Language Technologies for African Languages*. Athens: Association for Computational Linguistics, pp 9–16.

Giliomee, H. & Mbenga, B. (2007). *New History of South Africa*. Cape Town: Tafelberg.

Johannesburg Stock Exchange. (2009). Listed Companies. [Online]. Available: http://www.jse.co.za/listed_companies.jsp (accessed October 2009).

Modipa, T., Oosthuizen, H. & Manamela, M. (2007). Automatic Speech Recognition of Spoken Proper Names. Sovenga: University of Limpopo.

Moore, R.C. (2002). Fast and accurate sentence alignment of bilingual corpora. In *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas*, volume 2499 of Lecture Notes in Computer Science, Berlin: Springer Verlag, pp. 135–144.

Neamatullah I., Douglass M.M., Lehman L.W., Reisner A., Villarroel M., Long W.J., Szolovits P., Moody G.B., Mark R.G. & Clifford G.D. (2008). Automated de-identification of free-text medical records. In *BMC Medical Informatics and Decision Making* 2008, 8(32).

North-West University. (2009). [Online]. Available: http://galactrix.puk.ac.za/opendir.asp (accessed October 2009).

Puttkammer, M.J. (2006). *Outomatiese Afrikaanse Tekseenheididentifisering "Automatic Text Unit Identification"*. Potchefstroom: NWU.

Republic of South Africa. (1996). *Constitution of the Republic of South Africa*. Act 108 of 1996. Pretoria: Government Printer.

Republic of South Africa. (2003). *National Language Policy Framework*. Pretoria: Department of Arts and Culture.

Sproat, R. 2005. REFLEX: Named Entity Recognition and Transliteration for 50 Languages. [Online]. Available: http://serrano.ai.uiuc.edu/reflex (accessed December 2009).

Van der Merwer, I.J. & Van der Merwe, J.H. (2006). *Linguistic Atlas of South Africa.* Stellenbosch: Sun Press.

Developing an Open source Spell-checker for Gĩkũyũ

Kamau Chege[1], Peter Wagacha[1], Guy De Pauw[1,2], Lawrence Muchemi[1]
Wanjiku Ng'ang'a[1], Kenneth Ngure[3], Jayne Mutiga[3]

[1]School of Computing and Informatics
University of Nairobi
Kenya
kamauchege@gmail.com
{waiganjo,lmuchemi,wanjiku.nganga}@uonbi.ac.ke

[2]CLiPS - CLG
University of Antwerp
Belgium
guy.depauw@ua.ac.be

[3]Department of Linguistics
University of Nairobi
Kenya
kamurink@yahoo.com
jaynemutiga@yahoo.co.uk

Abstract

In this paper, we describe the development of an open source spell checker for Gĩkũyũ, using the Hunspell language tools. We explore the morphology of Gĩkũyũ and highlight the inflection of various parts of speech, including verbs, nouns, and adjectives. In Hunspell, surface words are realized as a set of continuation classes, with each class providing a morpheme with a specific function. In addition, circumfixation, which is prevalent in Gĩkũyũ, is implemented. Hunspell also provides for word suggestion, using character prevalence and replacement rules. Given that the developed Gĩkũyũ spellchecker and the Hunspell tools are open source, the spell checking function developed in this work can be adopted in major open-source products such as Mozilla and OpenOffice products. The spell checker has a fairly representative Gĩkũyũ lexicon and achieves an acceptable realization of a Gĩkũyũ spellchecker. When tested on a test corpus, the spell checker attains a precision of 82%, recall of 84% and an accuracy of 75%.

1. Introduction

A grammatically and orthographically correct text is necessary in ensuring high quality textual documents for effective communication. There is therefore a need to avail tools and utilities to support electronic document preparation. A spell checker is a design feature or a program that verifies the spelling of words in a document, query, and browsers, among other contexts. The goals of developing human language technology applications can only be achieved if localization, and basic language tools or utilities like spell checkers are made publicly available for a language. This paper describes such an effort for the Kenyan language of Gĩkũyũ.

Gĩkũyũ can be classified as a resource scarce language with respect to language technology resources, tools and applications. This situation can be attributed to different factors. First, Kiswahili and English are the dominant languages in Kenya, meaning that most of all textual communication is in these languages. Second, Gĩkũyũ orthography contains two diacritically marked characters (ĩ and ũ) that require extra keystrokes to generate, a situation which often makes users opt for the diacritically unmarked equivalents, resulting in non-standard Gĩkũyũ texts. These extra characters also pose a challenge for automated corpus collection methods, such as those using optical character recognition (OCR).

However, despite such an unfavorable backdrop, Gĩkũyũ, together with a few other Kenyan languages, are steadily becoming commercial languages as evidenced by their increased use in broadcast media, publishing and advertising. These developments point to a need for language technology support for such languages to boost their use, growth and viability. The work described here makes a positive contribution to this need, by providing a word processing utility that will encourage and enhance creation of correctly spelled Gĩkũyũ texts.

1.1. Gĩkũyũ Language Technology

There exists a number of language technology research efforts on Gĩkũyũ. These include a grapheme-based approach to diacritic restoration (Wagacha et al., 2006b; De Pauw et al., 2007), morphological analysis using a maximum entropy approach (De Pauw and Wagacha, 2007) and finite-state techniques (Chege, 2009), machine translation (Chege, 2009) and speech synthesis (Maina, 2009).

The closest effort towards Gĩkũyũ spell checking is a dictionary-based system, incorporated in a Gĩkũyũ text editor (Chege, 2007). This system works well, but for only a limited number of words, as contained in the dictionary. The work described herein overcomes this limitation by using a rule-based approach for determining the correct spelling of any Gĩkũyũ word.

2. Gĩkũyũ

Gĩkũyũ is a Bantu language belonging to the Kamba-Kikuyu subgroup of the Niger-Congo family, with over seven million speakers living in Central Kenya. The language has six dialects and is lexically similar to closely related languages such as Gĩchuka, Kĩembu, Kĩmerũ, and Kĩkamba.

Gĩkũyũ is highly inflectional and is characterized by a complex word structure and phonemics. It is also highly agglutinative, with words being formed from a battery of prefixes and suffixes. Like many Bantu languages, Gĩkũyũ has fifteen noun classes and two additional locative classes. It also has a concord system formed around the noun classes. Gĩkũyũ is a tonal language, a feature that introduces ambiguity at different grammatical levels, although tonality is not represented in the standard orthography.

2.1. Noun Morphology

Gĩkũyũ nouns can be grouped into two categories, namely, derived and underived nouns. Underived nouns consist of named entities, while derived nouns can be formed in one of

two ways: by affixing diminutive, augmentative or collective prefixes to an underived noun, or through verbal nominalization. The following examples illustrate these processes:

nyũmba ⇒ *kĩ-nyũmba*	(a big house)
imondo ⇒ *tũ-mondo*	(many nice handbags)
thaaka ⇒ *mũ-thaak-i*	(player)
hooya ⇒ *i-ho-ero*	(Place of prayer)
getha ⇒ *i-geth-a*	(harvesting occasion)
thooma ⇒ *ga-thom-i*	(the small one who reads)

Membership to a noun class is determined by a concord system with agreement enforced on other sentence components, such as adjectives and verbs. All Gĩkũyũ nouns, derived or underived, can also be optionally affixed with the locative suffix *-inĩ*, which changes the meaning from a referential entity to a location, as shown in the following examples:

metha-inĩ	(on the table)
mũtĩ- inĩ	(on the tree)

2.2. Verb Morphology

A typical Gĩkũyũ verb consists of a combination of zero or more dependent morphemes, a mandatory dependent morpheme and a mandatory final vowel. The simplest verb consists of a verb root and a final vowel. These are usually commands or directives. Subjunctive verb formations, i.e. commands, can optionally take a plural marker *-i* or *–ni*. Examples include:

ma-thaak-e	(so that they play)
in-a-i	(sing)
nĩ-ci-mũ-hat-agĩr-a	(they usually sweep for him/her)
reh-e-ni	(bring)

Gĩkũyũ verbs can also undergo full or partial reduplication, depending on the number of syllables in a word. Words with two syllables or less undergo full reduplication, while words with more than two syllables undergo partial reduplication, with only the first two syllables being reduplicated. Examples are:

negena ⇒ *nega-negena*	(make noise a little more)
tiga ⇒ *tiga-tiga*	(leave a little more)

Gĩkũyũ verbs are also affected by consonantal and vowel phonemics. Meinhof's Law involves consonants *b, c, r, t, g, k* being replaced with NC composites in verbs obeying first person singular, noun classes 1, 8 and 9 concord systems. Dahl's Law is a consonantal mutation that involves the cause sound *k* appearing before trigger voiceless sounds *c,k,t*, being replaced with its equivalent voiced sound *g*. Examples include:

rathima ⇒ *ndathima*
uma ⇒ *nyumia*
kũ-theka ⇒ *gũtheka*
kĩ-ka-thira ⇒ *gĩgathira*

Vowel mutation includes vowel lengthening before prenasalized stops and vowel assimilation when some vowel combinations appear in the neighborhood of each other (Mugane, 1997).

3. Methodology

Developing a spell checker requires a method of determining the set of valid words in a given language, against which the words to be checked are compared. In addition, if a word is not valid, a list of possible suggestions or alternatives need to be returned, in some prioritized order. In this work, we use the Hunspell tools (Németh, 2010), which facilitate the definition of the valid words in a language, as well as the likely suggestions.

We relied on a Gĩkũyũ corpus to obtain valid words, as well as to determine the typical misspellings that could occur in this language. The following subsections describe the Gĩkũyũ corpus and the pre-processing that was done prior to building the spell checker. A description of how to customize the Hunspell environment for Gĩkũyũ is also given. The latter subsections discuss the actual definition of the spell checker, categorized into nouns and verbs.

3.1. Corpus Collection

The primary development corpus is from a collection of a set of 19,000 words from previous works on Gĩkũyũ at the School of Computing and Informatics, University of Nairobi (Wagacha et al., 2006a; Wagacha et al., 2006b; De Pauw et al., 2007; De Pauw and Wagacha, 2007). This corpus has a bias on religious material but also includes poems, short stories, novels and Internet sources. The corpus was pre-processed manually to eliminate non-Gĩkũyũ words, and to correct diacritics, where necessary. The corpus was then manually annotated where words were categorized into corresponding parts of speech, in line with Hunspell's defined continuation classes. Perl scripts were used for generic annotation and marking.

The test corpus was acquired from two sources: a popular Gĩkũyũ blog "Maitũ nĩ ma itũ (Our Mother is our truth)" was chosen as it contains diacritically-marked texts on a variety of contemporary topics and Ngũgĩ wa Thiong'o's novel "Mũrogi wa Kagogo", which is not diacritically marked and represents how a normal user would type on a standard keyboard.

3.2. Hunspell Language Tools

Hunspell is a set of open source tools and utilities used for development and testing of spellcheckers and morphological analyzers. The main goal of Hunspell and its predecessors is to compress the lexicon of a language into a manageable size. Hunspell is an enhancement of its predecessors *Ispell* and *MySpell*, with the latter being the official spellchecker for OpenOffice and Mozilla products.

Hunspell was built to support languages with rich morphology, including complex prefixes and compounding. In addition, Hunspell supports circumfixation where a certain suffix can only co-occur with a given prefix or set of prefixes. Hunspell makes use of two files: the affix file which defines the morphological rules of the language, and the dictionary file which contains the actual word stems. Applicable affix rules must be specified for each stem.

3.3. Defining the Gĩkũyũ Spell checker

The spellchecker is implemented using the concept of continuation classes, where a word is represented as a compo-

Foc + Subj + Neg + Cond + Tense + Obj + Redup + Verb + DvbExt + Asp + FVwl

Figure 1: Verbal Affixation in Gĩkũyũ.

sition of one or more morphemes.

3.3.1. Hunspell Language Setup for Gĩkũyũ

To handle Gĩkũyũ diacritics, it is important to set character support to Unicode (UTF-8). In addition, since Gĩkũyũ verbs generate many affix rules, Flag is set to a number so as to handle the numerous affix rules. The Gĩkũyũ alphabet includes the apostrophe and hyphen, as in *ng'ombe* and *iria-inĩ*, and the orthography set is therefore extended with these characters. This is important as it helps Hunspell determine word stops. Since Gĩkũyũ has more than one level of prefixes and suffixes, support for complex prefixes, as well as circumfixation, has to be enabled in Hunspell.

3.3.2. Suggestions Component

The suggestions component is used to generate probable suggestions for a misspelled word. It is implemented in the affix file. Hunspell uses two sections in the affix file when generating suggestions for misspelled words. The first is the TRY command. This lists the language's orthography set in order of frequency. A more frequently used character has more weight during suggestions. The TRY command is shown below:

TRY eiĩanrtocduũgmhbykw'jNRTCGDMHBEAUŨYOIĨKWJ

The second command used in the suggestion component is the REPLACE command. The command lists the most commonly misspelled n-grams and their replacements. The major n-grams are a result of influence from different dialects, foreign languages and also by differences in spoken versus written Gĩkũyũ. Examples of Gĩkũyũ replace suggestions include:

```
REP 35
REP s c
REP sh c
REP sh ch
REP c ch
REP f b
REP v b
REP l r
REP i ĩ
```

3.3.3. Noun Component

The noun component is implemented in two parts, namely the underived nouns and the derived nouns, taking into account that both noun types can take the optional locative suffix.

```
SFX 251 Y 1
SFX 251 –inĩ .
```

Underived nouns have a continuation class leading to the common locative class. The class consists of optional diminutive, augmentative and collective prefixes. Class selectors are influenced by prenasalization and Dahl's Law.

```
PFX 201 Y 25
PFX 201 mb kab mb
PFX 201 mb tũb mb
PFX 201 ng rũg ng
PFX 201 ng' tũg ng'
PFX 202 Y 16
PFX 202 0 tũ [abmngrw][^bdngj]
PFX 202 i ka i[^cktĩ]

# Inside the Dictionary File
mbiira/201,251
icungwa/202,251
```

Derived nouns are formed through circumfixation. The CIRCUMFIX and NEEDAFFIX are used to enforce circumfixation. Examples of circumfixation include:

```
PFX 211 Y 20
PFX 211 0 mũ/002 .
PFX 211 0 tũ/002 .
PFX 211 0 kĩ/002 [^ckt]
PFX 211 0 gĩ/002 [ckt]
SFX 261 Y 1
SFX 261 0 i/211,002

#Inside the Dictionary File
thaaka/261
ruga/261
```

Example generated words are: *mũrugi, kĩrugi, karugi*

3.3.4. Verb Component

The continuation classes for verbs cater for the focus marker, concord subject, object classes, conditional, negation, tense, deverbal extensions, aspectual markers and final vowels. Since Hunspell only supports a maximum of three prefixes and three suffixes, it was impossible to directly implement the continuation classes as described, since verbs can have up to seven prefixes and four suffixes, as illustrated in Figure 1.

To handle this challenge, we combined all prefixes into one complex prefix. While it is possible to identify other prefix combinations that can simplify implementation, such combinations are subject to other problems. For instance, following from Dahl's Law, it is necessary to determine where the trigger (c, k, t) and the cause (k) appear together.

Similar problems are evident with prenasalization and Meinhof's Law. The singular complex prefix approach that was adopted for this study is shown in the following snapshot of the verb component definition. Given that the simplest verb form consists of a final vowel appended to the verb stem, the only mandatory continuation class in the verb component is therefore the final vowel continuation class.

PFX 611 Y 27705
PFX 611 0 ndĩha [ˆeĩoũ]
PFX 611 e ndĩhe e
PFX 611 ĩ ndĩhe ĩ
PFX 611 0 matingĩrama [ˆeĩoũ]
PFX 611 e matingĩrame e
PFX 611 ĩ matingĩrame ĩ

 # Negation
PFX 711 Y 1495
PFX 711 0 nda [ˆaeĩoũ]
PFX 711 a nda a
PFX 711 u gũtigu u[ckt]
PFX 711 e hatikwe e[ˆckt]

#Aspectual/Modal Suffixes, Voiced Consonants
SFX 482 Y 3
SFX 482 a ete [ˆi]a
SFX 482 ia etie ia
SFX 482 wo etwo wo

#Deverbal Extensions, All Verbs
SFX 450 Y 30
SFX 450 a anga [ˆi]a
SFX 450 a angwo [ˆi]a

 # Deverbal Extensions, Verbs beginning
 # with Voiced Consonants
SFX 451 Y 12
SFX 451 a ĩka [ˆi]a
SFX 451 ia ĩka ia
SFX 451 a ĩra [ˆi]a
SFX 451 a ĩrwo [ˆi]

 # Deverbal Extensions, Verbs beginning
 # with Voiceless Consonants
SFX 452 Y 12
SFX 452 a eka [ˆi]a
SFX 452 ia eka ia
SFX 452 a era [ˆi]a
SFX 452 a erwo [ˆi]a

 # In the Dictionary File
arama/611,711,612,712,613,713,614,714,450,451,480,481
etha/611,711,612,712,613,713,614,714,450,452,480,482

Verbs in the dictionary file are categorized on the basis of applicable continuation classes. As seen in the examples above, these categories (groupings) are also influenced by whether the verb ends in a mid-low or mid-high voice, and whether it can take a causative extension, among others.

4. Results and Discussion

The developed Gĩkũyũ spellchecker and "suggester" engine was incorporated into OpenOffice Writer[1] and evaluated using the test corpus.

The results obtained are shown in Table 1 and are based on typical evaluation measures. In this study, True Positives (TP) represent those correctly spelled words that are recognized as such by the spell checker. False Positives (FP) represent misspelled words that are not flagged as such. True Negatives (TN) represent misspelled words that are flagged as such, while False Negatives (FN) represent correctly spelled Gĩkũyũ words that are flagged as misspelled.

Results	TP	FP	TN	FN	TOTAL
No. of Instances	3351	756	854	618	5579
Precision	$TP/(TP+FP) = 0.82$				
Recall	$TP/(TP+FN) = 0.84$				
Accuracy	$(TP+TN)/Total = 0.75$				

Table 1: Evaluation Results for Test Corpus

On analysing the results, it was noted that the major reason for unrecognized Gĩkũyũ words (False Negatives) is word stem missing in the dictionary file, as well as proper names, especially of people and places. Having a richer corpus from which to draw stems in addition to the inclusion of a named entity recognition heuristic would be one strategy to reduce the false negatives, thereby improving spell checking accuracy.
It was observed that, when spell checking diacritically marked texts, many misspellings (True Negatives) are generated. Subsequently, these words were easy to correct using the suggestions generated. However, the suggestion component degraded when the misspelling was a combination of a diacritic and other (one or more) characters.
An issue that is not evident in the results, but which was a major challenge, is over-generation, where the uncontrolled combination of prefixes and suffixes especially on verb morphology, generated a large number of non-Gĩkũyũ words due to semantic/meaning violations.

5. Conclusion

In this work, we have reviewed the development of an open-source spellchecker for Gĩkũyũ language using the Hunspell language tools. Results obtained in applying the developed spellchecker in OpenOffice Writer, have shown an acceptable performance with an accuracy of 75%, a precision of 82% and a recall of 84%.
The developed spell checker clearly has practical value in the spell checking of Gĩkũyũ texts. The developed tool can also provide utility in the collation and correction of additional Gĩkũyũ texts during corpus compilation. The methodology employed in this work can be easily ported to other Bantu languages that share a large percentage of similarity stems, as one approach in bootstrapping the development of spell checkers for these languages.

Availability and Acknowledgments

A beta version of the spellchecker is available from http://extensions.services.openoffice.org/en/project/Gikuyu.
The research presented in this paper was made possible through a generous grant by the Acacia Program in IDRC,

[1]For OpenOffice 2.x and below, integrating the spell checker requires copying files to appropriate locations and editing the dictionary.lst file accordingly, while for OpenOffice 3.x, dictionaries are added as extensions.

Canada, through the African Localization Network - ANLoc.

6. References

ANLoc. (2010). *The African Network for Localization*. [Online]. Available: http://www.africanlocalisation.net (accessed March 2010).

Chege, K. (2007). *Language-Independent Spellchecker: Gĩkũyũ Word Processor*. Nairobi, Kenya: School of Computing and Informatics, University of Nairobi. Unpublished Paper.

Chege, K. (2009). *Morphological Analysis of Gĩkũyũ: Towards Machine Translation*. Nairobi, Kenya: School of Computing and Informatics, University of Nairobi. Unpublished Paper.

De Pauw, G. & Wagacha, P.W. (2007). Bootstrapping morphological analysis of Gĩkũyũ using unsupervised maximum entropy learning. In H. Van hamme & R. van Son (Eds.), *Proceedings of the Eighth Annual Conference of the International Speech Communication Association*. Antwerp, Belgium.

De Pauw, G., Wagacha, P.W. & de Schryver, G-M. (2007). Automatic diacritic restoration for resource-scarce languages. In Václav Matoušek & Pavel Mautner (Eds.), *Proceedings of Text, Speech and Dialogue, Tenth International Conference*. Heidelberg, Germany: Springer Berlin / Heidelberg, pp. 170–179.

Maina, L.W. (2009). *Text-to-Speech System for Gĩkũyũ with Interactive Voice Response System*. Nairobi, Kenya: School of Computing and Informatics, University of Nairobi. Unpublished Paper.

Mugane, R. (1997). *A Paradigmatic Grammar of Gĩkũyũ*. Stanford, USA: CLSI Publications.

Németh, L. (2010). *Hunspell*. [Online]. Available: http://hunspell.sourceforge.net (accessed: March 2010).

Wa Mbugwa, G. (2010). *Maitu ni ma itu (Our Mother is our truth)*. [Online]. Available: http://gigikuyu.blogspot.com (accessed March 2010).

wa Thiong'o, N. (2007). *Mũrogi wa Kagogo*. Nairobi, Kenya: East African Educational Publishers Ltd.

Wagacha, P.W., De Pauw, G. & Getao, K. (2006)a. Development of a corpus for Gĩkũyũ using machine learning techniques. In J.C. Roux (Ed.), *Proceedings of LREC workshop - Networking the development of language resources for African languages*. Genoa, Italy: European Language Resources Association, pp. 27–30.

Wagacha, P.W., De Pauw, G. & Githinji, P.W. (2006)b. A grapheme-based approach for accent restoration in Gĩkũyũ. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*. Genoa, Italy: European Language Resources Association, pp. 1937–1940.

The verbal phrase of Northern Sotho: A morpho-syntactic perspective

Gertrud Faaß

Universität Stuttgart
Institut für maschinelle Sprachverarbeitung
Azenbergstraße 12, 70174 Stuttgart, Germany
gertrud.faasz@ims.uni-stuttgart.de

Abstract

So far, comprehensive grammar descriptions of Northern Sotho have only been available in the form of prescriptive books aiming at teaching the language. This paper describes parts of the first morpho-syntactic description of Northern Sotho from a computational perspective (Faaß, 2010a). Such a description is necessary for implementing rule based, operational grammars. It is also essential for the annotation of training data to be utilised by statistical parsers. The work that we partially present here may hence provide a resource for computational processing of the language in order to proceed with producing linguistic representations beyond tagging, may it be chunking or parsing. The paper begins with describing significant Northern Sotho verbal morpho-syntactics (section 2). It is shown that the topology of the verb can be depicted as a slot system which may form the basis for computational processing (section 3). Note that the implementation of the described rules (section 4) and also coverage tests are ongoing processes upon that we will report in more detail at a later stage.

1. Language introduction and overview

Northern Sotho is one of the eleven national languages of South Africa and one of the four Sotho languages of the South-eastern Language Zone group (S.30 in the classification of Guthrie (1971)): Northern and Southern Sotho, Tswana ("Western Sotho") and Lozi (Silozi, Rozi). The Sotho languages are written disjunctively (Van Wyk, 1958), i.e. a number of (mostly inflectional) affixes (henceforth "morphemes") are written separately instead of being merged with the stems that they belong to. Especially Northern Sotho verbs differ significantly from that of other languages, not only in respect to their moods, but also in the many different morpheme constellations that they may appear in.

Concerning its computational processing, Northern Sotho language resources and tools have been the subject of a few publications during recent years of which, for space reasons, we can only list a few: a methodology for tagging corpora of Northern Sotho has been designed by Prinsloo and Heid (2005), a number of finite-state tokenizers have been reported upon, concerning its verbs e.g. by Anderson and Kotzé (2006), a first tagset and a tagging experiment have been described by De Schryver and De Pauw (2007). Taljard et al. (2008) defined a more finely granulated tagset of which we partially make use in this paper. Lastly, Faaß et al. (2009) reported on results of tagging Northern Sotho with an approach of disambiguation of polysemous items.[1] This paper aims at designing a first morpho-syntactic framework for a number of Northern Sotho verbal constellations.[2] The paper does not claim to present a complete description, it is however shown here that a significant fragment of the verbal grammar can be depicted with a basic distinction made between elements syntactically subcategorised by the verb stem on the basis of its lexical semantic properties on the one hand, and its inflectional elements on the other.

The rules defined here are already in the process of being implemented in a rule-based parser (cf. (Faaß, 2010b) for a first report), however, they may also be used for other purposes, e.g. to produce training data for a statistical parser. Coverage tests are currently in process.

2. Northern Sotho verbs: An Introduction

Northern Sotho verbs appear in a variety of moods; for space reasons, we however focus on the independent predicative moods in this paper: indicative, situative and relative (for an overview of the Northern Sotho moods, see e.g. (Lombard, 1985, p. 144)). These moods all appear in the three tenses future, present and past and all may be negated. To mark tense or polarity, each of the moods makes use of specific morphemes to appear in front of the verb stem. Note that except for the missing tense marking, the dependent grammatical moods (consecutive, habitual and subjunctive) basically make use of the same morphemes as described here. The verb stem itself may show specific affixes fused to it, e.g. indicating some of the past constellations (allomorphs of the past tense morpheme *-il-*), and it appears with certain endings, inter alia *-a* or *-e*, that each are predefined by the constellation they must occur in. The ending *-a*, for example, usually appears in the positive constellations, while some negative ones require *-e* to appear. Other verbal affixes indicate certain semantic changes (e.g. passive: *-w-*, applicative voice: *-el-*, or reciprocal: *-an-*), of which most change its argument structure, however, these are not relevant for mood detection and thus not furtherly described here.

The indicative mood differs from the situative mood insofar as it forms a matrix sentence, cf. (1-a), while the situative

[1] For an overview of NLP resources of the African languages of South Africa, cf. (Roux and Bosch, 2006).

[2] This paper focuses on main verbs, it is developed on the grounds of (Faaß, 2010a).

mood usually does not, cf. (1-b). The relative mood depicts relative clauses, e.g. as in (1-c).

(1) a. *monna* *o* *reka dipuku*
man$_{\text{N01}}$ `subj-3rd-01` buy books$_{\text{N10}}$
'(a) man buys books'

b. *ge* *monna* *a* *reka dipuku*
when man$_{\text{N01}}$ `subj-3rd-01` buy books$_{\text{N10}}$
'when (a) man buys books'

c. *monna* *yo* *a*
man$_{\text{N01}}$ `dem-3rd-01` `subj-3rd-01`
rekago *dipuku*
buy-`rel` books$_{\text{N10}}$
(a) man who buys books'

2.1. Elements of the Northern Sotho verb

This section is focusing on the following contents of the Northern Sotho verbs: The subject concord (relevant for subject-verb agreement, the verb stem and its objects, some negation clusters, and tense markers. These present the elements present in all grammatical moods of Northern Sotho.

2.1.1. Subject-verb agreement

Northern Sotho predicative verbs[3] all have to agree with their subject (in terms of either noun class[4] or person/number). A class specific subject concord marks this (usually anaphoric) relationship to the referent; if the relationship cannot be established, the neutral concord *e* is used.

In (1-a), $o_{\text{subj-3rd-01}}$ appears, a class 1 subject concord which is to be used for the positive indicative. In (1-a) (indicative) and (1-b) (situative), the verb stem (V) and the subject concord (CS$_{\text{categ}}$[5]) together form the linguistic verb. In the relative mood, cf. (1-c), a demonstrative concord CDEM$_{\text{categ}}$ is added. This concord – from a morpho-syntactic perspective – is not part of the verb itself, but precedes it (forming a CP containing a VP[6]). It is thus not described furtherly in this article. The ending *-go* internally marks the relative mood (see also paragraph 2.1.4.).

Like other Bantu languages, Northern Sotho is a null-subject language: in the case that the subject noun is omitted (e.g. because it is known in the discourse), the subject concord will acquire its grammatical function, cf. (2). For indefinite cases, similar to the English expletive 'it', the indefinite subject concord *go* appears with the grammatical function of a subject.

(2) *o* *reka dipuku*
`subj-3rd-01` buy books$_{\text{N10}}$
'(s)he buys books'

[3]Non-predicative verbs of Northern Sotho are infinitives and imperatives, cf. e.g. Lombard's modal system (Lombard, 1985, p. 144 table 7.5.4).

[4]The noun classes of Northern Sotho are described in detail by e.g. (Lombard, 1985, p. 29 et seq.) and (Faaß, 2010a).

[5]$_{\text{categ}}$ stands for classes as defined by (Taljard et al., 2008): 01 - 10, 14, 15, LOC and the persons PERS_1sg to PERS_3pl respectively.

[6]The demonstrative concord also heads other constellations, like, e.g. adjectival phrases.

Each mood makes use of a specific type of subject concord. It is thus mandatory to define several sets of these so that they can be depicted as morpho-syntactic rules. Poulos and Louwrens (1994, p. 168 et seq.) define two sets, ignoring the fact that *o* and *a* (both described as being in set 1) occur in specific moods and are therefore not interchangable. We therefore opt for the definition of three sets, hence extending the labels, where set 1 contains *o* for class 1, and set 2 containing *a* instead (the subject concords of all other classes remain identical), and set 3 containing what is described as the "consecutive" (e.g. by Lombard (1985, p. 152 et seq.)). This set of subject concords also appears in non-consecutive moods, e.g. in one of the negated past tense forms of the relative mood, therefore we decide for a more neutral naming of the category, "3", instead.

Out three sets of subject concords are labeled accordingly as 1CS⟨categ⟩, 2CS⟨categ⟩ and 3CS⟨categ⟩; cf. examples in (3-a), where we repeat (1-a), (3-b), where we repeat (1-c), and (3-c).

(3) a. *monna* $o_{\text{1CS01(set1)}}$ *reka dipuku*
man$_{\text{N01}}$ `subj-3rd-01` buy books$_{\text{N10}}$
'(a) man buys books'

b. *monna* *yo* $a_{\text{2CS01(set2)}}$
man$_{\text{N01}}$ `dem-3rd-01` `subj-3rd-01`
rekago *dipuku*
buy-`rel` books$_{\text{N10}}$
'(a) man who buys books'

c. *lesogana* le_{CDEM05} $le_{\text{1CS05(set1)}}$
young man `dem-3rd-05` `subj-3rd-05`
sego $la_{\text{3CS05(set3)}}$ *go toropong*
`neg` `subj-3rd-05` go town-`loc`
(a) young man who did not go to town.

2.1.2. The verb stem and its arguments

The verb stem *reka* '[to] buy' in the examples above is transitive and followed by its object, *dipuku* 'books', hence one could indeed assume that Northern Sotho is an SVO language. However, this order of functional elements is only valid for overt objects and independent of the moods the verb appears in: in the case of an object being omitted, like e.g. in a discourse where it is already known, a pronominal object concord (CO$_{\text{categ}}$) has to appear in front of the verb stem, as shown in (4). Note that in such a case (i.e. where a positive indicative present tense ends in the verb stem), a tense marker is to be inserted after the subject concord (cf. paragraph 2.1.4.) .

Concerning double transitives, both objects may not be simultaneously substituted by means of a pronominal object concord. The Northern Sotho verb, unlike, e.g. Setswana verbs, may contain only one object concord (usually, the indirect object is pronominalized), an example of a pronominalisation process is shown in (5).

(4) *monna* *o* *a* *di* *reka*
man$_{\text{N01}}$ `subj-3rd-01` `pres` `obj-3rd-10` buy
'(a) man buys them'

(5) a. *ke* *diretše* *monna*
`subj-1st-sg` make-`appl-perf` man$_{\text{N01}}$
kofi
coffee$_{\text{N09}}$
'I made coffee for (the) man'

b. *ke* *diretše* *monna*
`subj-1st-sg` make-`appl-perf` man$_{\text{N01}}$

yona
emp-3rd-09
'I made it for the man'

c. *ke* *mo*
subj-1st-sg obj-3rd-01
diretše *yona*
make-appl-perf emp-3rd-09
'I made it for him/her'

The subject concord preceding this constellation can be set separately from the VBP, not only because it is an inflectional element that is relevant for the subject-verb agreement, but also because it can appear with any other verb stem+object(s) constellation. We therefore define an optional "Verbal Inflectional Element" (henceforth VIE) containing it. Note that the positive imperative VP of Northern Sotho contains a VBP only.

Though Northern Sotho is classified as a disjunctively written language, there are some object concords that are merged to the verb stem, like *N–*, referring to the first person singular. The verb stem *bona*$_{\text{V\_tr}}$ '[to] see', merges with this object concord to *mpona* '[to] see me', as in (6). From a syntactic perspective, such verbs have their argument structure saturated, therefore we use respective labels when tagging such tokens: "V_sat-tr" to indicate a saturated transitive verb. Half saturated double transitive verbs are labelled "V_hsat-dtr", cf. (7) containing the double transitive verb stem *fa* '[to] give', again merged with the object concord of the first person singular, *N–*, forming *mpha* 'give me'. Such specific annotations for fused forms, however, may become obsolete in the future, as a morphological analyser can split them, annotate them with their respective part of speech and report the two elements separately to the parser. Such an analyser already exists, cf. e.g. Anderson and Kotzé (2006) and some of its implemented routines could thus be utilised as a pre-processing element to parsing.

(6) *monna* *o* *a* ***mpona***$_{\text{V\_sat-tr}}$
man$_{\text{N01}}$ subj-3rd-01 pres obj-1st-sg-see
'(a) man sees me'

(7) *monna* *o* ***mpha***$_{\text{V\_hsat-dtr}}$
man$_{\text{N01}}$ subj-3rd-01 obj-1st-sg-give
dipuku
books$_{\text{N10}}$
'(a) man gives me books.'

2.1.3. Negation

The kinds of negation affixes (or clusters thereof) that are to appear in a verbal constellation of Northern Sotho vary significantly in the different moods (and tenses). Two examples are shown here: a negated indicative (8) and a negated situative in (9).

(8) *monna* ***ga*** ***a*** *reke dipuku*
man$_{\text{N01}}$ neg subj-3rd-01 buy books$_{\text{N10}}$
'(a) man does not buy books.'

(9) *ge* *monna* ***a*** ***sa*** *reke dipuku*
when man$_{\text{N01}}$ subj-3rd-01 neg buy books$_{\text{N10}}$
'when (a) man does not buy books'

(8) and (9) demonstrate that the order in which subject concord and negation morpheme(s) appear depends on the mood of the verb. Some negated forms, e.g. one of the negated forms of the relative, as shown in (3-c) above, even contain two subject concords surrounding the negation morpheme. Thus we need to define one slot containing both. Secondly, these examples show that the verb stem in a negated phrase often ends in *–e* instead of *–a*. The verbal ending in general, however, may itself be seen as underspecified information, because it does not determine a specific mood or tense, but a set of them. Together with certain contents of the VIE, it may be distinctive when identifying a certain mood/tense. We therefore add an attribute to the VBP: "Vend" describing the verbal ending that the verb stem has to appear with (cf. paragraph 2.). This attribute is to be stored with the correct value for each verb stem in the lexicon of the parser. Any morpho-syntactic rule defined thus can utilise the attribute "Vend" and assign a default value to it (cf. table 4 demonstrating the morpho-syntactic rules describing the indicative VP).

Such a method makes also sense in the case of irregular verbs: the so-called "defective" verb stem *re* '[to] say', for example, behaves like any regular verbs ending in *–a*. Following this method, we can define *re* as ending in *–a*.

2.1.4. Tense marking

The present tense morpheme *a* appearing in (4) and (6) above, is explicitly defined for the positive indicative verbal phrases that end in the verb stem (the so-called "long" form of the verb, cf. e.g. (Poulos and Louwrens, 1994, p. 208 et seq.)). All other present tense constellations do not contain a specific tense marker. The past tense is often marked with an allomorph of the verbal ending *-il–e*, which may appear inter alia as *-etš–e, -itš–e, -tš–e, -š–e* etc. Again, we can make use of the attribute "Vend" in describing all past tense forms as ending in *-ile*.

Consider examples (10) and (11), where (11) contains the subject concord of class 2 (usually referring to humans in the plural) and the past tense form of *sepela* 'walk', *sepetše*.

(10) *monna* *o* *rek**ile*** *dipuku*
man$_{\text{N01}}$ subj-3rd-01 buy-past books$_{\text{N10}}$
'(a) man bought books'

(11) *ba* *sepe**tše***.
subj-3rd-02 walk-past
'they walked'

For other past tense forms, like the negated past relative shown in example (3-c), however, the constellation of subject concords and negation morpheme in the VIE is specific enough to mark the past tense. Here, the verb stem appears without any specific tense infix, and with the verbal ending *a*.

There are basically two interchangable future tense affixes found in Northern Sotho texts: *tlo* and *tla*, the relative mood of the future tense makes additional use of *tlogo* and *tlago* which appear whenever the verb stem does not have the relative suffix *-go* added. Any positive predicative mood can use one of these affixes which appear after the subject concord (or the cluster containing negation and subject concord) and before the object concord, as e.g. in (12).

(12) *monna* *o* ***tlo*** *di*
man$_{\text{N01}}$ subj-3rd-01 fut obj-3rd-10
reka
books$_{\text{N10}}$
'(a) man will buy books.'

The slot system					
VIE		**VBP**			
zero-2	**zero-1**	**slot zero verb stem and its object(s)**			
pos-n to pos-3 subject conc. and/or negation	**pos-2** tense marker	**pos-1** object concord	**pos-0**$_{\text{Vend}}$ **verb stem**	**pos+1** object 1	**pos+2** object 2

Table 1: A schematic representation of the slot system

3. A slot system to describe the Northern Sotho verbal constellations

Table 1 demonstrates our knowledge of the Northern Sotho topology of the verb in a slot system: the "core" of the Northern Sotho verb contains the verb stem and its object(s), we call this part of the verb "Verbal Basic Phrase" (VBP). It is split into four positions: pos-1 can contain a separate object concord (or is empty), pos-0 contains the verb stem (possibly merged with an object concord). The optional pos+1 and pos+2 finally may contain overt objects.

To cater for the present and future tense affixes which always occur between the subject concord/negation cluster and VBP, the slot VIE contains two positions which we call zero-1 and zero-2 (while the four positions of the VBP slot are summarised as zero-0), cf. table 1. Slot zero-2 can contain several morphemes, zero-1 has only one position defined.

4. Computational processing

As stated above, inflectional (tense/negation) and subject-verb-agreement information are provided by subject concords and morphemes that precede the verb stem with the exception of the past tense and passive affixes which are merged to the verb stem, changing its ending. Such information can be described as attribute/value pairs in the lexicon. Concerning the examples stated above, the lexicon entries will appear as shown in table 2, where the parameter "Vend" appears in column 3. Such handling makes sense moreover in the case of irregular verbs: *re* '[to] say', for example, behaves like any other regular verb ending in *–a*. Any morpho-syntactic rule defined for a parser can utilise this parameter. Table 3 shows a simplified Lexical Functional Grammar (LFG) lexicon, containing some sample entries of morphemes and nouns. Their parts of speech and noun class are also encoded as attribute/value pairs. Additionally, number and person attribute/value pairs are listed for sake of information.

A typical morpho-syntactic rule describing the indicative mood fills the slot system as shown in table 4. The parentheses in *(–w)* indicate optionality, i.e. all verbs contained in these constellations may appear in their passive form.

In the framework of LFG, we can define e.g. a VIE-rule for the positive indicative (the "short" and the "long"' forms) as follows (the following shows a simplified version of the so far implemented operational grammar (Faaß, 2010b) for the sake of demonstration):

```
VP   →   VIE VBP.
         "General Verbal Inflectional Elements : VIE"
VIE  →   "short present tense form:"
         "subject concord of the first set"
     {   1CS : (↑ TNS-ASP FORM) = short;
         "no specific tense element"
         e : (↑ TNS-ASP TENSE) = pres
         "constraint: verbal ending in lexicon to be"
         "defined as a"
         (↑ VEND) =c a

         "long present tense form:"
         "subject concord of the first set"
     |   1CS
         "present tense morpheme indicates the long form"
         MORPH: (↑ TNS-ASP FORM) = long
         (↑ TNS-ASP TENSE) = pres;
         "constraint: verbal ending in lexicon must"
         "be defined as a"
         e : (↑VEND) =c a
         ...
     }
```

A fragment of these morpho-syntactic rules has been processed in the framework of LFG[7], and implemented in the Xerox Linguistic Environment provided by the Xerox Palo Alto Research Centre (PARC, `http://www2.parc.com/isl/groups/nltt/xle`) in the framework of a research license. It is planned to make this grammar available as a free web service.

5. Conclusions and future work

Concerning computational processing of Northern Sotho text, so far only linguistic "essentials" like tagsets (De Schryver and De Pauw, 2007; Taljard et al., 2008) and tagging procedures have been developed. Furthermore, finite-state machinery approaches that describe Northern Sotho's linguistic words have been delineated. As the next logical step in producing high level computational linguistic representations, morpho-syntactic descriptions of part-of-speech constellations forming verbal phrases are provided by this paper.

[7] See e.g. `http://www-lfg.stanford.edu/lfg`.

verb stem	label (transitivity)	verbal ending	comments
reka	V_tr	–a	'[to] buy'
reke	V_tr	–e	'[to] buy'
rekile	V_tr	–ile	'bought'
rekago	V_tr	–a-rel	'who buy(s)'
bona	V_tr	–a	'[to] see'
mpona	V_sat-tr	–a	'[to] see me', Obj=1st-sg
mphe	V_hsat-dtr	–a	'[to] give me', Obj=1st-sg
sepetše	V_itr	–il–e	'walked'
longwa	V_itr	-w–a	'[to] bite' passive form
ya	V_itr	-a	'walk"
ile	V_itr	–il–e	'walked'
re	V_tr	–a	'[to] say'

Table 2: Lexicon entries of verb stems

monna	n(oun)	class = 1, pers = 3, num = sg
lesogana	n(oun)	class = 5, pers = 3, num = sg
a	2CS	class = 1
	3CS	class = 1
	...	
MORPH	(↑ TNS-ASP TENSE) = pres, (↑ TNS-ASP FORM) = long	
le	1CS	class = 5
	CDEM	class = 5
	...	
tlo	MORPH	(↑ TNS-ASP TENSE) = fut.

Table 3: Sample lexicon entries of other parts of speech

^{INDPRES}VP				
	VIE		**VBP**	
descr.	**zero-2**	**zero-1**	**zero**	**Vstem ends in**
pres.pos.long	$1CS_{categ}$	MORPH_pres	VBP	*(–w)–a*
pres.pos.short	$1CS_{categ}$		VBP	*(–w)–a*
pres.neg.	ga_{MORPH_neg} $2CS_{categ}$		VBP	*(–w) –e*
perf.pos.	$1CS_{categ}$		VBP	*–il(–w)–e*
perf.neg. 1	ga_{MORPH_neg} se_{MORPH_neg} $3CS_{categ}$		VBP	*(–w) –a*
perf.neg. 2	ga_{MORPH_neg} se_{MORPH_neg} $2CS_{categ}$		VBP	*(–w)–e*
perf.neg. 3	ga_{MORPH_neg} $3CS_{categ}$		VBP	*(–w)–a*
perf.neg. 4	ga_{MORPH_neg} $1CS_{categ}$ $–a_{MORPH\_past}$		VBP	*(–w) –a*
fut.pos	$1CS_{categ}$	*tlo/tla* $_{MORPH\_fut}$	VBP	*(–w)–a*
fut.neg	$2CS_{categ}$ ka_{MORPH_pot} se_{MORPH_neg}		VBP	*(–w) –e*

Table 4: A summary of the independent indicative forms

Here, the verb's topology, i.e. the rather fixed order of the tokens that form the Northern Sotho verb allows for the definition of a slot system, fulfilling the following conditions:

- The central slot contains the verb and its object(s) or object concord (and second object in the case of a double transitive verb). Its contents solely depend on the lexical semantics of the verb stem in question. Standing alone, it describes an imperative VP.
- In the case of a negated imperative or a predicative VP[8], one to two preceding slots are to be defined of which
  - the first slot contains a subject concord of a predefined set and/or a specific negation morpheme (cluster)
  - the second slot is optional, however, if it occurs, it contains at maximum one element indicating tense.

Information on the verbal ending is often crucial when identifying certain moods, tenses or negated forms. There-

[8]We exclude the infinitive here, though it as well can be described in the slot system. The infinitive will be described in a separate publication, cf. (Faaß and Prinsloo, forthcoming)

fore, the lexical attribute "Vend" is introduced, which may be used as constraints by morpho-syntactic rules. It caters however not only for "regular" forms (like, e.g. "Vend" = *-a* for the verb rek*a*), but also for irregular forms ("Vend" = *-a* for the verb stem r*e*) and for allomorphs of the past tense morpheme *-il-*, e.g. *-etš-)*, as in *sepetše*.

Different moods make use of different sets of subject concords. So far, only two such sets have been described (e.g. by Poulos and Louwrens (1994)). The first of these sets contains two different subject concords for noun class 1, *o* and *a*. In order to be able to identify the precise mood/tense/polarity of a verb, we define two sets (1CS and 2CS) instead. The third set, so far beeing called the "consecutive" set, is renamed to the more neutral name "3CS" as it also occurs in moods other than the consecutive. Morpho-syntactic rules can now make use of these more precisely defined sets of parts of speech.

The design of a slot system for verbal phrases described in this paper is part of a project developing an operational LFG grammar of Northern Sotho (Faaß, 2010a) in the framework of XLE. Current work entails the extension of the grammar, and an implementation of a Northern Sotho noun phrase chunker making use of the CASS parser (Abney, 1996) and doing coverage tests. For utilising the descriptions for statistical parsing, morpho-syntactic analyses of these may be taken as a start point for the development of training data.

6. References

Abney, S. (1996). Partial parsing via finite-state cascades. *Natural Language Engineering*, 2(4), pp. 337–344.

Anderson, W. & Kotzé, P.M. (2006). Finite state tokenisation of an orthographical disjunctive agglutinative language: The verbal segment of Northern Sotho. In *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC-2006*. Genova, Italia: ELRA. [CD-ROM].

De Schryver, G-M. & De Pauw, G. (2007). Dictionary writing systems (DWS) + corpus query package (CQP): The case of Tshwanelex. *Lexikos*, 17, pp. 226 – 246.

Faaß, G. & Prinsloo, D.J. (forthcoming). A morpho-syntactic view on the infinitive of Northern Sotho.

Faaß, G., Heid, U., Taljard, E. & Prinsloo, D.J. (2009). Part-of-Speech tagging in Northern Sotho: disambiguating polysemous function words. In *Proceedings of the EACL2009 Workshop on Language Technologies for African Languages–AfLaT 2009*. Athens, Greece: Association for Computational Linguistics, pp. 38–45.

Faaß, G. (2010a). *A morphosyntactic description of Northern Sotho as a basis for an automated translation from Northern Sotho into English*. Ph.D. thesis, University of Pretoria, Pretoria, South Africa. (submitted).

Faaß, G. (2010b). A toy LFG grammar describing some of the morpho-syntactics of Northern Sotho verbs. Presented at Spring Pargram Meeting, University of Constance. [Online]. Available: http://www.ims.uni-stuttgart.de/~faaszgd/spring-pargram-2010-NSotho-handout.pdf (accessed March 2010).

Guthrie, M. (1971). *Comparative Bantu: an introduction to the comparative linguistics and prehistory of the Bantu languages, vol 2*. Farnborough: Gregg International.

Lombard, D.P. (1985). *Introduction to the Grammar of Northern Sotho*. Pretoria, South Africa: J.L. van Schaik.

Poulos, G. & Louwrens, L.J. (1994). *A Linguistic Analysis of Northern Sotho*. Pretoria, South Africa: Via Afrika.

Prinsloo, D.J. & Heid, U. (2005). Creating word class tagged corpora for Northern Sotho by linguistically informed bootstrapping. In I. Ties (Ed.), *LULCL Lesser used languages and computational linguistics*. Bozen, Italy: Eurac, pp. 97–113.

Roux, J. & Bosch, S.E. (2006). Language resources and tools in Southern Africa. In *Proceedings of the Workshop on Networking the Development of Language Resources for African Languages. 5th International Conference on Language Resources and Evaluation*. Genova, Italy: ELRA, pp. 11–15.

Taljard, E., Faaß, G., Heid, U. & Prinsloo, D.J. (2008). On the development of a tagset for Northern Sotho with special reference to the issue of standardization. *Literator – special edition on Human Language Technologies*, 29(1), pp. 111–137.

Van Wyk, E.B. (1958). *Woordverdeling in Noord-Sotho en Zulu (Word division in Northern Sotho and Zulu)*. Ph.D. thesis, DLitt thesis. University of Pretoria, Pretoria, South Africa.

Capturing Word-level Dependencies in Morpheme-based Language Modeling

Martha Yifiru Tachbelie, Wolfgang Menzel

University of Hamburg
Natural Language System Group
tachbeli,menzel@informatik.uni-hamburg.de

Abstract

Morphologically rich languages suffer from data sparsity and out-of-vocabulary words problems. As a result, researchers use morphemes (sub-words) as units in language modeling instead of full-word forms. The use of morphemes in language modeling, however, might lead to a loss of word level dependency since a word can be segmented into 3 or more morphemes and the scope of the morpheme n-gram might be limited to a single word. In this paper we propose the use of roots to capture word-level dependencies in Amharic language modeling. Our experiment shows that root-based language models are better than the word based and other factored language models when compared on the basis of the probability they assign for the test set. However, no benefit has been obtained (in terms of word recognition accuracy) as a result of using root-based language models in a speech recognition task.

1. Introduction

1.1. Language modeling

Language models (LM) are fundamental to many natural language applications such as automatic speech recognition (ASR) and statistical machine translation (SMT).

The most widely used language models are statistical language models. They provide an estimate of the probability of a word sequence W for a given task. The probability distribution depends on the available training data and how the context has been defined (Junqua and Haton, 1995). Large amounts of training data are, therefore, required in statistical language modeling so as to ensure statistical significance (Young et al., 2006).

Even if we have a large training corpus, there may be still many possible word sequences which will not be encountered at all, or which appear with a statistically insignificant frequency (data sparseness problem) (Young et al., 2006). Even individual words might not be encountered in the training data irrespective of its size (Out-of-Vocabulary words problem).

The data sparseness problem in statistical language modeling is more serious for languages with a rich morphology. These languages have a high vocabulary growth rate which results in high perplexity and a large number of out of vocabulary words (Vergyri et al., 2004). As a solution, sub-word units are used in language modeling to improve the quality of language models and consequently the performance of the applications that use the language models (Geutner, 1995; Whittaker and Woodland, 2000; Byrne et al., 2001; Kirchhoff et al., 2003; Hirsimäki et al., 2005).

1.2. The morphology of Amharic

Amharic is one of the morphologically rich languages. It is a major language spoken mainly in Ethiopia and belongs to the Semitic branch of the Afro-Asiatic super family. Amharic is related to Hebrew, Arabic and Syrian.

Like other Semitic languages such as Arabic, Amharic exhibits a root-pattern morphological phenomenon. A root is a set of consonants (called radicals) which has a basic 'lexical' meaning. A pattern consists of a set of vowels which are inserted (intercalated) among the consonants of a root to form a stem. The pattern is combined with a particular prefix or suffix to create a single grammatical form (Bender et al., 1976) or another stem (Yimam, 2000). For example, the Amharic root sbr means 'break', when we intercalate the pattern ä_ä and attach the suffix ä we get säbbärä 'he broke' which is the first form of a verb (3rd person masculine singular in past tense as in other semitic languages) (Bender et al., 1976). In addition to this non-concatenative morphological feature, Amharic uses different affixes to create inflectional and derivational word forms.

Some adverbs can be derived from adjectives. Nouns are derived from other basic nouns, adjectives, stems, roots, and the infinitive form of a verb by affixation and intercalation. For example, from the noun lIğğ 'child' another noun lIğnät 'childhood'; from the adjective däg 'generous' the noun dägnät 'generosity'; from the stem sInIf, the noun sInIfna 'laziness'; from root qld, the noun qälId 'joke'; from infinitive verb mäsIbär 'to break' the noun mäsIbäriya 'an instrument used for breaking' can be derived. Case, number, definiteness, and gender marker affixes inflect nouns.

Adjectives are derived from nouns, stems or verbal roots by adding a prefix or a suffix. For example, it is possible to derive dInIgayama 'stony' from the noun dInIgay 'stone'; zInIgu 'forgetful' from the stem zInIg; sänäf 'lazy' from the root snf by suffixation and intercalation. Adjectives can also be formed through compounding. For instance, hodäsäfi 'tolerant, patient', is derived by compounding the noun hod 'stomach' and the adjective säfi 'wide'. Like nouns, adjectives are inflected for gender, number, and case (Yimam, 2000).

Unlike the other word categories such as noun and adjectives, the derivation of verbs from other parts of speech is not common. The conversion of a root to a basic verb stem requires both intercalation and affixation. For instance, from the root gdl 'kill' we obtain the perfective verb stem gäddäl- by intercalating pattern ä_ä. From this perfective stem, it is possible to derive a passive (tägäddäl-) and a causative stem (asgäddäl-) using prefixes tä- and as-, respectively. Other verb forms are also derived from roots in a similar fashion.

Verbs are inflected for person, gender, number, aspect, tense and mood (Yimam, 2000). Other elements like negative markers also inflect verbs in Amharic.
From the above brief description of Amharic morphology it can be seen that Amharic is a morphologically rich language.

1.3. Language modeling for Amharic

Since Amharic is a morphologically rich language, it suffers from data sparseness and out of vocabulary words problems. The negative effect of Amharic morphology on language modeling has been reported by Abate (2006), who, therefore, recommended the development of sub-word based language models for Amharic.
To this end, Tachbelie and Menzel (2007) and Tachbelie and Menzel (2009) have developed various morpheme-based language models for Amharic and gained a substantial reduction in the out-of-vocabulary rate. They have concluded that, in this regard, using sub-word units is preferable for the development of language models for Amharic. In their experiment, Tachbelie and Menzel (2007) and Tachbelie and Menzel (2009) considered individual morphemes as units of a language model. This, however, might result in a loss of word level dependencies since a word might be segmented into 3 or more morphemes and the span of the n-grams might be limited to a single word. The easiest way of handling this problem might be using higher order n-grams which, however, highly increases the complexity of language models. Therefore, approaches that capture word level dependencies are required to model the Amharic language. Kirchhoff et al. (2003) introduced factored language models that can capture word level dependency while using morphemes as units in language modeling.
In this paper, we present how we captured word-level dependency in morpheme-based language modeling (in the framework of factored language modeling) by taking advantage of the nature of the Amharic language that root consonants represent the lexical meaning of the words derived from them. Section 2. gives a description of our approach for handling word-level dependencies in morpheme-based language modeling and Section 3. presents detail of the language modeling experiment and the results. The language models have been applied to a speech recognition task. Section 4. deals with the speech recognition experiments we have conducted. Conclusions and future research directions are given in Section 5. But before that we give a brief description of factored language modeling.

1.4. Factored language modeling

Factored language models (FLM) have first been introduced in Kirchhoff et al. (2002) for incorporating various morphological information in Arabic language modeling. In FLM a word is viewed as a bundle or vector of K parallel factors, that is, $w_n \equiv f_n^1, f_n^2, ..., f_n^k$. The factors of a given word can be the word itself, stem, root, pattern, morphological classes, or any other linguistic element into which a word can be decomposed. The idea is that some of the feature bundles (for example: roots, patterns and morphological class) can uniquely define the words i.e. $(W = w_i) \equiv (R = r_i, P = p_i, M = m_i)$. Therefore, the word n-gram probabilities can be defined in terms of these features/factors as follows (Kirchhoff et al., 2003):

$$\begin{aligned} &P(w_i|w_{i-1}, w_{i-2}) \\ &\quad = P(r_i, p_i, m_i|r_{i-1}, p_{i-1}, m_{i-1}, r_{i-2}, p_{i-2}, m_{i-2}) \\ &\quad = P(r_i|p_i, m_i, r_{i-1}, p_{i-1}, m_{i-1}, r_{i-2}, p_{i-2}, m_{i-2}) \\ &\qquad P(p_i|m_i, r_{i-1}, p_{i-1}, m_{i-1}, r_{i-2}, p_{i-2}, m_{i-2}) \\ &\qquad\quad P(m_i|r_{i-1}, p_{i-1}, m_{i-1}, r_{i-2}, p_{i-2}, m_{i-2}) \end{aligned} \tag{1}$$

2. Capturing word-level dependency

As it has been indicated above, the idea in factored language modeling is that some of the features might uniquely define the words and language models can be estimated on the basis of these features. Since in Amharic the root consonants from which a word is derived represent the basic lexical meaning of the word, we considered the root consonants as features that uniquely define the word i.e.

$$w_i \equiv r_i \tag{2}$$

Where w_i is the ith word and r_i a root from which the word is derived. The word n-gram probabilities can, therefore, be defined (according to equation 1) in terms of the consonantal roots as follows.

$$P(w_i|w_{i-1}w_{i-2}) = P(r_i|r_{i-1}r_{i-2}) \tag{3}$$

As it is clear from formula 3, we actually developed a root-based n-gram model to capture word-level dependencies. One can also consider the model as a skipping one since we skip other morphemes during language model estimation.

3. Language modeling experiment

3.1. Morphological analysis

To use morphemes in language modeling a word parser, which splits word forms into their morpheme constituents, is needed. Different attempts (Bayou, 2000; Bayu, 2002; Amsalu and Gibbon, 2005) have been made to develop a morphological analyzer for Amharic using different methods. However, none of the systems can be directly used for our purpose. The systems developed by Bayou (2000) and Bayu (2002) suffer from lack of data. The morphological analyzer developed by Amsalu and Gibbon (2005) seems to suffer from a too small lexicon. It has been tested on 207 words and analyzed less than 50% (75 words) of them.
An alternative approach might be unsupervised corpus-based methods (such as Morfessor (Creutz and Lagus, 2005)) that do not need annotated data. These methods are particularly interesting for resource scarce languages like Amharic. However, the corpus-based morphology learning tools try to find only the concatenative morphemes in a given word and are not applicable for the current study in which the roots (the non-concatenative morphemes) are required. Therefore, we manually segmented 72,428 word types found in a corpus of 21,338 sentences.
To do the segmentation, we used two books as manuals: Yimam (2000) and Bender and Fulass (1978). Baye's book describes how morphemes can be combined to form words.

The list of roots in Bender and Fulass (1978) helped us to cross check the roots that we suggest during the segmentation.

However, we do not claim that the segmentation is comprehensive. Since a word type list has been used, there is only one entry for polysemous or homonymous words. For example, the word t'Iru might be an adjective which means 'good' or it might be an imperative verb which has a meaning 'call' (for second person plural). Consequently, the word has two different segmentations. Nevertheless, we provided only one segmentation based on the most frequent meaning of the word in our text. In other words we disambiguated based on the frequency of use in the text. Because the transcription system does not indicate geminated consonants, the geminated and non-geminated word forms, which might have distinct meanings and segmentations, have also been treated in the same manner as the polysemous or homonymous ones. For instance, the word ?at'ägäbu can be treated as an adverb which means 'next to him' or as a verb with a meaning 'they made somebody else full or they satisfied somebody else' based on the gemination of the consonant g. Consequently, this word could have, therefore, been segmented in two different ways: [?at'ägäb + u] if it is an adverb or [?a + t'gb + aa + u] if it is a verb derived from the root t'gb.

3.2. Factored data preparation

As our aim is to handle word level dependencies in the framework of factored language modeling we need a corpus in which each word is represented as a vector of factors or features. Although only the root consonants are required for the present work, we prepared the data in such a way that it can also be used for other experiments. In our experiment each word is considered as a bundle of features including the word itself, part-of-speech tag of the word, prefix, root, pattern and suffix. Each feature in the feature vector is separated by a colon (:) and consists of a tag-value pair. In our case the tags are: W for word, POS for Part-of-Speech, PR for prefix, R for root, PA for pattern and SU for suffix. A given tag-value pair may be missing from the feature bundle. In this case, the tag takes a special value 'null'.

The manually segmented data that include 21,338 sentences or 72,428 word types or 419,660 tokens has been used to prepare the factored version of the corpus. The manually segmented word list has been converted to a factored format and the words in the corpus have been automatically substituted with their factored representation. The resulting corpus has then been used to train and test the root-based language models presented below.

3.3. The language models

Although all the verbs are derived from root consonants, there are words in other part-of-speech class which are not derivations of root consonants. Normally, these words have the value 'null' for the root feature (the R tag). If we consider the word as being equivalent with its root, these words will be excluded from our model which in turn will have a negative impact on the quality of the language models. Preliminary investigation also revealed the fact that the perplexities of the models has been influenced by the null values for the root tag in the data. Therefore, we modified equation 2 as follows.

$$w_i \equiv \begin{cases} r_i, & if\ root \neq null \\ stem_i, & otherwise \end{cases} \tag{4}$$

Where $stem_i$ is the stem of a word (that is not derived from root consonants) after removing all the prefixes and suffixes. We did not introduce a new feature called stem to our factored data representation. But when the word is not a derivation of consonantal root, we consider the stem as a root instead of assigning a null value.

We divided the corpus into training, development and evaluation test sets in the proportion of 80:10:10. We have trained root-based models of order 2 to 5. Since previous experiments (Tachbelie and Menzel, 2007) revealed the fact that Kneser-Ney smoothing outperforms all other smoothing methods, we smoothed the root-based language models with this technique. Table 1 shows the perplexity of these models on the development test set. A higher improvement in perplexity (278.57 to 223.26) has been observed when we move from bigram to trigram. However, as n increases above 3, the level of improvement declined. This might be due to the small set of training data used. As n increases more and more n-grams might not appear in the training corpus and therefore the probabilities are computed on the basis of the lower order n-grams. The pentagram model is the best model compared to the others. This model has a perplexity of 204.95 on the evaluation test set.

Root ngram	Perplexity	Word ngram	Perplexity
Bigram	278.57	Bigram	1148.76
Trigram	223.26	Trigram	989.95
Quadrogram	213.14	Quadrogram	975.41
Pentagram	211.93	Pentagram	972.58

Table 1: Perplexity of root- and word-based models on development test set

In order to measure the benefit gained from the root-based model, we have developed word based models with the same training data. These models have also been tested on the same test data (consisting of 2,134 sentences or 20,989 words) and the same smoothing technique has also been applied. The difference from the root-based models is that complete words are used as units instead of roots. The pentagram model has a perplexity of 972.58 (as shown in table 1) on the development test set. Moreover, the number of out-of-vocabulary words is much lower (295) in the root-based models than in the word based ones (2,672). Although the test set used to test the word- and root-based models has the same number of tokens, direct comparison of the perplexities of these models is still impossible since they have a different number of out-of-vocabulary words. Therefore, comparison of the best root- and word based models has been performed on the basis of the probability they assign to the test set. The log probability of the best root based model is higher (-53102.3) than that of the word based model (-61106.0). That means the root-based

models are better than the word based ones with respect to the probability they assign to the test set.
Word based language models that use one additional word-dependent feature in the ngram history have also been developed since integrating features, such as part-of-speech, into language models might improve their quality. In these models, a word trigram probability is estimated, for example, as $w_n|w_{n-2}pos_{n-2}w_{n-1}pos_{n-1}$ instead of $w_n|w_{n-2}w_{n-1}$. Table 2 gives the perplexities of these models. Although they are better than the word only models (in terms of the probability they assign to the test set), none of the models outperformed the root-based ones. However, as the levels of detail modeled in root-based and other language models are different, the root-based models have been applied in a speech recognition system to prove that they are really better than the other models.

Language models	Perplexity
W/W2,POS2,W1,POS1	885.81
W/W2,PR2,W1,PR1	857.61
W/W2,R2,W1,R1	896.59
W/W2,PA2,W1,PA1	958.31
W/W2,SU2,W1,SU1	898.89

Table 2: Perplexity of models with different factors

4. Speech recognition experiment

In order to analyse the contribution (in terms of performance improvement) of the root-based and other factored language models in a speech recognition task, the speech recognition system which has been developed by Abate (2006) has been used. Section 4.1. presents the speech recognition system. To make our results comparable, we have developed root-based and other factored language models that are equivalent with the ones elucidated in section 3.3. They differ from the models described in the preceding section by having been trained on the text which has been used to develop the bigram word based language model that was originally used in the speech recognition system. These language models, interpolated with the bigram word based language model, have then been used to rescore lattices generated with the speech recognition system. We applied a lattice rescoring framework, because it is problematic to use factored language models in standard word decoders. Section 4.2. presents the new set of language models and 4.3. deals with the result of the lattice rescoring experiment.

4.1. The speech recognition system

4.1.1. The speech and text corpus

The speech corpus used to develop the speech recognition system is a read speech corpus (Abate et al., 2005). It contains 20 hours of training speech collected from 100 speakers who read a total of 10,850 sentences (28,666 tokens). Compared to other speech corpora that contain hundreds of hours of speech data for training, this corpus is obviously small in size and accordingly the models will suffer from a lack of training data.
Although the corpus includes four different test sets (5k and 20k both for development and evaluation), for the purpose of the current investigation we have generated the lattices only for the 5k development test set, which includes 360 sentences read by 20 speakers.
The text corpus used to train the backoff bigram language model consists of 77,844 sentences (868,929 tokens or 108,523 types).

4.1.2. The acoustic, lexical and language models

The acoustic model is a set of intra-word triphone HMMs with 3 emitting states and 12 Gaussian mixtures that resulted in a total of 33,702 physically saved Gaussian mixtures. The states of these models are tied, using decision-tree based state-clustering that reduced the number of triphone models from 5,092 logical models to 4,099 physical ones.
Encoding the pronunciation dictionary can range from very simple and achievable with automatic procedures to very complex and time-consuming that requires manual work with high linguistic expertise. The Amharic pronunciation dictionary has been encoded by means of a simple procedure that takes advantage of the orthographic representation (a consonant vowel syllable) which is fairly close to the pronunciation in many cases. There are, however, notable differences especially in the area of gemination and insertion of the epenthetic vowel.
The language model is a closed vocabulary (for 5k) backoff bigram model developed using the HTK toolkit. The absolute discounting method has been used to reserve some probabilities for unseen bigrams where the discounting factor, D, has been set to 0.5, which is the default value in the HLStats module. The perplexity of this language model on a test set that consists of 727 sentences (8,337 tokens) is 91.28.

4.1.3. Performance of the system

We generated lattices from the 100 best alternatives for each sentence of the 5k development test set using the HTK tool and decoded the best path transcriptions for each sentence using the lattice processing tool of SRILM (Stolcke, 2002). Word recognition accuracy (WRA) of this system was 91.67% with a language model scale of 15.0 and a word insertion penalty of 6.0. The better performance (compared to the one reported by Abate (2006), 90.94%, using the same models and on the same test set) is due to the tuning of the language model and word insertion penalty factors.

4.2. Root-based and factored models

The manually segmented data has also been used to obtain a factored version of the corpus that was used to develop the backoff bigram word based language model. The factored version of the corpus has been prepared in a way similar to the one described in Section 3.2. This corpus has then been used to train closed vocabulary root-based and factored language models. All the factored language models have been tested on the factored version of the test set used to test the bigram word based language model.
We have developed root-based n-gram language models of order 2 to 5. The perplexity of these models on the development test set is presented in Table 3. The highest perplexity

improvement has been obtained when the n-gram order has been changed from bigram to trigram.

Language models	Perplexity	Logprob
Root bigram	113.57	-18628.9
Root trigram	24.63	-12611.8
Root quadrogram	11.20	-9510.29
Root pentagram	8.72	-8525.42

Table 3: Perplexity of root-based models

Other factored language models that take one word feature (besides the words) in the n-gram history have been developed. The additional features used are part-of-speech (POS), prefix (PR), root (R), pattern (PA) and suffix (SU). The models are equivalent in structure with the factored language models described in Section 3.3. The perplexity and log-probability of these models are presented in Table 4. The models are almost similar in perplexity and probability.

Language models	Perplexity	Logprob
W/W2,POS2,W1,POS1	10.614	-9298.57
W/W2,PR2,W1,PR1	10.67	-9322.02
W/W2,R2,W1,R1	10.36	-9204.7
W/W2,PA2,W1,PA1	10.89	-9401.08
W/W2,SU2,W1,SU1	10.70	-9330.96

Table 4: Perplexity of other factored language models

4.3. Lattice rescoring

Since it is problematic to use factored language models in standard word decoders, we substituted each word in the lattices with its factored representation. A word bigram model that is equivalent to the one originally used in the speech recognition system has been trained on the factored data and used for factored representations. This language model has a perplexity of 63.59. The best path transcription decoded using this language model has a WRA of 91.60%, which is slightly lower than the performance of the normal speech recognition system (91.67%). This might be due to the smoothing technique applied in the development of the language models. Although absolute discounting with the same discounting factor has been applied to both bigram models, the unigram models have been discounted differently. While in the word based language model the unigram models have not been discounted at all, in the equivalent factored model the unigrams have been discounted using Good-Turing discounting technique which is the default discounting technique in SRILM.

The root-based and the other factored language models (described in Section 4.2.) have been used to rescore the lattices.

All the factored language models that integrate an additional word feature in the n-gram history brought an improvement in WRA. Models with four parents did not bring much improvement when the maximal n-gram order to be used for transition weight assignment was set to 2. However, when trigrams are used, all the models brought notable improvement (see Table 5).

Language models	Word recognition accuracy in %
Factored word bigram (FBL)	91.60
FBL + W/W2,POS2,W1,POS1	93.60
FBL + W/W2,PR2,W1,PR1	93.82
FBL + W/W2,R2,W1,R1	93.65
FBL + W/W2,PA2,W1,PA1	93.68
FBL + W/W2,SU2,W1,SU1	93.53

Table 5: WRA with other factored language models

Unlike the other factored language models, root based language models led to a reduced word recognition accuracy as Table 6 shows. Although the higher order root-based model, namely the penta-gram, assigned the highest probability to the test set compared to all the other factored language models, it resulted in a WRA which is below that of the original speech recognition system.

Language models	Word recognition accuracy in %
Factored word bigram (FBL)	91.60
FBL + Root bigram	90.77
FBL + Root trigram	90.87
FBL + Root quadrogram	90.99
FBL + Root pentagram	91.14

Table 6: WRA with root-based models

5. Conclusion and future work

In Amharic, the consonantal roots from which a word is derived represent the basic lexical meaning of the words. Taking advantage of this feature, root-based language models have been developed as a solution to the problem of loss of word-level dependencies in Amharic morpheme-based language modeling. Our experiment shows that root-based language models are better than the word based and other factored language models when they are compared on the basis of the probability the models assign to a test set.

Since the best way of comparing language models is applying them to the target application for which they are developed and see whether they bring improvement in the performance of the application or not, the root-based models have been applied to a speech recognition task in a lattice rescoring framework. However, the speech recognition system did not benefit from these models. Thus, other ways of integrating the root-based models to a speech recognition system might be worth exploring.

6. Acknowledgment

We would like to thank the reviewers for their constructive and helpful comments.

7. References

Abate, S.T., Menzel, W. & Tafila, B. (2005). An Amharic speech corpus for large vocabulary continuous speech recognition. In *Proceedings of Ninth European Conference on Speech Communication and Technology, Interspeech-2005*.

Abate, S.T. (2006). Automatic speech recognition for Amharic. Ph.D. thesis, University of Hamburg.

Amsalu, S. & Gibbon, D. (2005). Finite state morphology of Amharic. In *Proceedings of International Conference on Recent Advances in Natural Language Processing*, pp. 47–51.

Bayou, A. (2000). Developing automatic word parser for Amharic verbs and their derivation. Master's Thesis, Addis Ababa University.

Bayu, T. (2002). Automatic morphological analyzer for Amharic: An experiment employing unsupervised learning and autosegmental analysis approaches. Master's Thesis, Addis Ababa University.

Bender, M.L. & Fulass, H. (1978). *Amharic Verb Morphology: A Generative Approach*. Michigan, USA: Michigan State University.

Bender, M.L., Bowen, J.D., Cooper, R.L. & Ferguson, C.A. (1976). *Languages in Ethiopia*. London, UK: Oxford University Press.

Byrne, W., Hajič, J., Ircing, P., Jelinek, F., Khudanpur, S., Krebc, P. & Psutka, J. (2001). On large vocabulary continuous speech recognition of highly inflectional language - Czech. In *Proceedings of the European Conference on Speech Communication and Technology*, pp. 487–489.

Creutz, M. & Lagus, K. (2005). Unsupervised morpheme segmentation and morphology induction from text corpora using morfessor 1.1. Technical Report A81. Neural Networks Research Center, Helsinki University of Technology.

Geutner, P. (1995). Using morphology towards better large-vocabulary speech recognition systems. In *Proceedings of IEEE International on Acoustics, Speech and Signal Processing*. Los Alamitos, CA, USA: IEEE Computer Society, pp. 445–448.

Hirsimäki, T., Creutz, M., Siivola, V. & Kurimo, M. (2005). Morphologically motivated language models in speech recognition. In *Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning*, pp. 121–126.

Junqua, J-C. & Haton, J-P. (1995). *Robustness in Automatic Speech Recognition: Fundamentals and Applications*. Norwell, MA, USA: Kluwer Academic Publishers.

Kirchhoff, K., Bilmes, J., Henderson, J., Schwartz, R., Noamany, M., Schone, P., Ji, G., Das, S., Egan, M., He, F., Vergyri, D., Liu, D. & Duta, N. (2002). Novel speech recognition models for Arabic. Technical report, Johns-Hopkins University Summer Research Workshop.

Kirchhoff, K., Bilmes, J., Das, S., Duta, N., Egan, M., Ji, G., He, F., Henderson, J., Liu, D., Noamany, M., Schone, P., Schwartz, R. & Vergyri, D. (2003). Novel approaches to Arabic speech recognition: Report from the 2002 Johns-Hopkins summer workshop. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pp. 344–347.

Stolcke, A. (2002). SRILM - an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*. Denver, USA: International Speech Communication Association, pp. 901–904.

Tachbelie, M.Y. & Menzel, W. (2007). Subword based language modeling for Amharic. In *Proceedings of International Conference on Recent Advances in Natural Language Processing*, pp. 564–571.

Tachbelie, M.Y. & Menzel, W. (2009). *Morpheme-based Language Modeling for Inflectional Language - Amharic*. Amsterdam and Philadelphia: John Benjamin's Publishing.

Vergyri, D., Kirchhoff, K., Duh, K. & Stolcke, A. (2004). Morphology-based language modeling for Arabic speech recognition. In *Proceedings of International Conference on Spoken Language Processing*, pp. 2245–2248.

Whittaker, E. & Woodland, P. (2000). Particle-based language modeling. In *Proceedings of International Conference on Spoken Language Processing*, pp. 170–173.

Yimam, B. (2000). *YäamarIña Säwasäw*. Addis Ababa: EMPDE.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V. & Woodland, Ph. (2006). *The HTK Book*. Cambridge, UK: Cambridge University Engineering Department.

Towards the Implementation of a Refined Data Model for a Zulu Machine-Readable Lexicon

Ronell van der Merwe, Laurette Pretorius, Sonja Bosch
University of South Africa
PO Box 392, UNISA 0003, South Africa
vdmerwer@unisa.ac.za, pretol@unisa.ac.za, boschse@unisa.ac.za

Abstract

The development of a machine-readable lexicon for Zulu requires a comprehensive data model for representing lexical information. This article focuses on the modelling of the verb structure in Zulu, and more specifically verbal extensions and deverbatives due to their complex recursive morphological structure. Moreover, we also show how the machine-readable lexicon, based on the proposed data model, may be embedded in an update framework. This framework also includes a finite-state morphological analyser, its guesser variant and electronically available Zulu corpora as a potential source of new lexical information. Finally implementation issues are considered in order to ensure accurate modelling and efficiency in the lexical repository.

1. Introduction

Comprehensive machine-readable (MR) lexicons remain one of the most basic language resources in human language technologies and natural language processing applications. The quest for comprehensiveness entails among others, the inclusion of all the lexical information and structure contained in paper dictionaries, and the sourcing of lexical information from electronically available Zulu corpora via an update framework. In particular, we aim at clarifying the role of the MR lexicon as a component in the update framework, as well as its relation with the other components, viz. the finite-state morphological analyser (ZulMorph), its guesser variant and electronically available corpora (cf. Pretorius and Bosch, 2003, Bosch et al., 2008). For lesser-resourced languages the development of such a resource is challenging and time consuming, and requires the optimal use of available corpora, enabling technologies and implementation approaches.

Firstly, we briefly discuss the components and processes that constitute the update framework for building and maintaining a comprehensive MR repository of lexical information for Zulu.

Secondly, we refine a comprehensive data model for such a resource, first proposed by Bosch et al. (2007:142), where it was argued that "well-defined recursion is the correct and intuitive way to capture certain linguistic information such as verbal extensions …" The importance of allowing recursion within entries in MR lexicons for the South African Bantu languages, in order to facilitate accurate modelling, is supported by Weber (2002:8). He states that "lexical databases should accommodate (1) derived forms having multiple senses and (2) derived forms ... of the bases from which they are derived". In the original data model, verbal extensions were treated as the main source of recursion. In this paper, the data model is extended by including deverbatives, the modelling of which is also based on recursion.

Finally, we consider the suitability of various implementation approaches with specific focus on the recursive nature of the data model.

2. Update Framework

The framework for developing and updating the comprehensive MR lexicon for Zulu is shown in Figure 1. Apart from the MR lexicon, the framework includes a finite-state morphological analyser, its guesser variant, new language data in the form of electronic corpora and linguistic expertise for moderating additions derived from the Guesser.

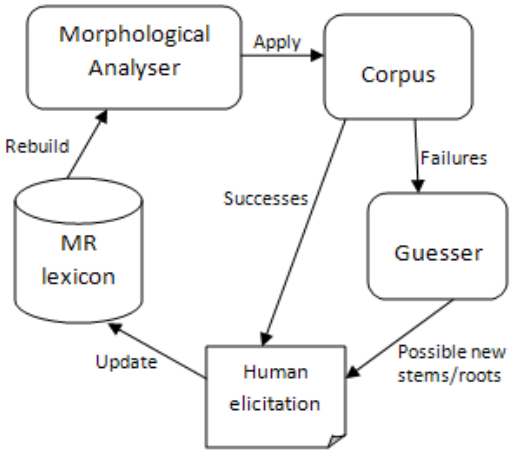


Figure 1: The Lexicon update framework

The main purpose of a comprehensive MR lexicon is to serve as a machine-readable repository of all lexical information. This resource may then be used and reused in various applications. One of its intended applications is to support the development and maintenance of a finite-state morphological analyser, e.g. ZulMorph. The morphological analyser represents only morphosyntactic information in the form of Zulu morphotactics, morphophonological alternation rules and an embedded stem/root lexicon, as automatically extracted from the MR lexicon. The guesser variant of the morphological analyser is designed to identify all phonologically possible stems/roots. This newly found candidate stems/roots are considered for linguistic validity and for inclusion in the MR lexicon.

3. Data model

The development of a comprehensive MR lexicon relies on an underlying data model for the Bantu languages that provides for storage of and access to all the units of information, both rule-based (regular) and idiosyncratic, associated with the attested words in a language. The lexicon is conceptualised as consisting of a collection of entries at the highest level. Each entry then has a tree-like structure for accommodating the relevant lexical information. The top four levels of this tree are shown in Figure 2. Each entry consists of a head (the stem) and a body, which represents general information about the stem such as its phonetic transcription, tone, morphosyntactic information (including part of speech), sense, dialect information, etymology, etc. (cf. Bosch et al., 2007) for a detailed discussion).

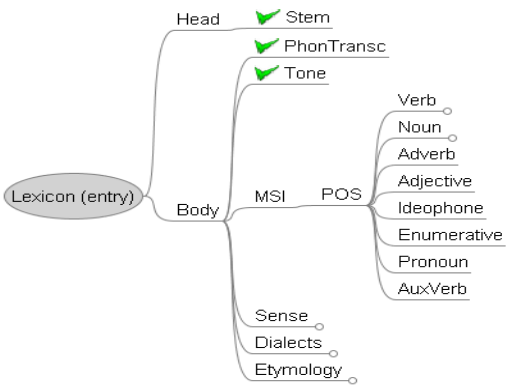


Figure 2: Top levels of the data model.

Figures 3a and 3b provide details of the modelling of the verb, exhibiting two sources of recurrence, viz. verbal extensions and nominal suffixes of deverbatives (more details in 3.1 and 3.2), which results in the generation of complex data such as nested derivational forms and their sense information. These (recursive) characteristics present specific challenges in developing a MR lexicon for the Zulu language. The Kleene * indicates 0 or more occurrences, the Kleene + indicates 1 or more occurrences, | optionality, and √ a leaf node.

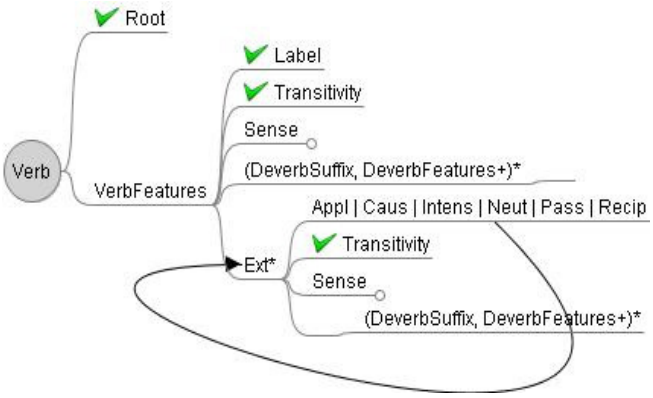


Figure 3a: Verb structure fragment of the data model

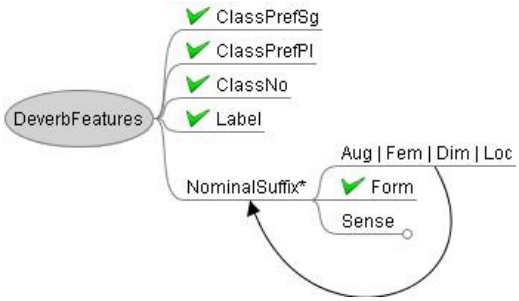


Figure 3b: Deverbative features fragment

The relevant fragment of the data model in the form of a DTD is as follows:

```
<!ELEMENT Entry (Head,Body)>
<!ELEMENT Head (Stem)>
<!ELEMENT Body (PhonTransc*, Tone*, MSI,
Sense+, Dialects*, Etymology?)>
...
<!ELEMENT MSI (POS)>
<!ELEMENT POS (Verb | Noun | Adverb |
Adjective | Ideophone | Enumerative | Pronoun
| AuxVerb)>
...
<!ELEMENT Verb (Root,VerbFeatures)>
<!ELEMENT VerbFeatures (Label, Transitivity,
Sense,(DeverbSuffix,DeverbFeatures+)*,Ext*
)>
<!ELEMENT Ext ((Appl | Caus | Intens | Neut
| Pass | Recip), Transitivity,Sense,
(DeverbSuffix,DeverbFeatures+)*,Ext*)>
...
<!ELEMENT DeverbFeatures  (ClassPrefSg?,
ClassPrefPl?, ClassNo, Label,
NominalSuffix*)>
<!ELEMENT NominalSuffix ((Aug | Fem | Dim |
Loc),Form, Sense, NominalSuffix*)>
...
```

We discuss and exemplify the modelling of two sources of recurrence, viz. verbal extensions and deverbatives.

3.1 Verbal extensions

In Zulu morphology, the basic meaning of a verb root may be modified by suffixing so-called extensions to the verb root. Examples of such extensions are the causative, reciprocal, passive, applied and neuter. We use the verb root *-bon-* 'see' to illustrate the sequencing of these suffixes as well as the associated modification of the basic meaning of the verb root.

(1) *-bon-is-* 'show, cause to see'
-verb.root-caus
(2) *-bon-an-* 'see each other'
-verb.root-recip
(3) *-bon-w-* 'be seen'
-verb.root-pass
(4) *-bon-el-* 'see for'

-verb.root-appl
(5) *-bon-akal-* 'be visible'
-verb.root-neut
(6) *-bon-is-an-* 'show each other'
-verb.root-caus-recip

Verbal extension suffixes modify the basic meaning of the verb root as illustrated in (1) to (6). In certain instances, as illustrated in (6), more than one extension may even be suffixed to the verb root. This structure is modelled in Figure 3 in the `Ext` substructure where recursion is denoted by the arc.
The sequencing of extensions is largely idiosyncratic and attested sequences need to be obtained from corpora and then stored in the MR lexicon. For instance, in a sequence of two or more extensions, the passive is usually last in the sequence. However, in the case of certain verb roots, the reciprocal extension follows the passive, while both sequences are possible in other cases, for example:
(7) *-bon-an-w-* 'seen by each other'
-verb.root-recip-pass
(8) *-bon-w-an-* 'seen by each other'
-verb.root-pass-recip

There are also instances where the applied extension follows the passive (cf. Van Eeden, 1956: 657):
(9) *ya-bulal-w-el-a* > *yabulawela* 'he was killed for'
subj.conc-verb.root-pass-appl-suffix

Similarly, Doke and Vilakazi (1964) list examples (10) and (11) which exemplify the causative and applied extensions in converse sequences, each expressing a slightly different meaning:
(10) *-bon-is-el-* 'look after for'
-verb.root-caus-appl
(11) *-bon-el-is-* 'cause to take care of'
-verb.root-appl-caus-

Existing paper dictionaries of Zulu do not contain exhaustive information on the combinations and sequences of extensions with verb roots. This type of information could, however, be extracted semi-automatically from language corpus resources.
Further examples of verbal extension sequences and how the meaning of the basic root is changed, are given in (12) to (16). This is modelled in Figure 3 by the `Sense` substructure associated with each `Ext`.
(12) *-bon-akal-* "be visible"
-verb.root-neut
(13) *-bon-akal-is-* "make visible"
-verb.root-neut-caus
(14) *-bon-el-an-* "see for/perceive for/take care of each other"
-verb.root-appl
(15) *-bon-el-el-* "treat with consideration"
-verb.root-appl-appl
(16) *-bon-el-el-w-* "be treated with consideration"
-verb.root-appl-appl-pass

For a detailed exposition of Zulu grammar and linguistic terminology, cf. Poulos and Msimang (1998).

3.2 Deverbatives

Deverbatives are formed when a verb (or extended verb) root is suffixed with a deverbative suffix (*o-*, *i-*, *a-*, *e-* and *u-*)[1] and takes a noun class prefix before the verb root. Such derivational affixes cannot combine randomly with any verb root, since they are restricted by semantic considerations.
The data model provides for the capturing of class information, the deverbative suffix, the sequence (recursion) of nominal suffixes, and associated semantic information. This structure is modelled in Figure 3 in the `Nominal Suffix` substructure where recursion is again denoted by the arc. The example in (21) represents the extended root *-bon-is-el-o* with two verbal extensions *-is-* and *-el-* as well as a nominal suffix *-ana*.
Deverbative nouns cannot arbitrarily be formed from any verb root (Van Eeden, 1956:712). At present the resolution of this issue, that is the valid combinations of noun prefixes and (extended) verb roots in the formation of deverbative nouns, is mainly determined from entries in existing dictionaries (attested forms) and occurrences of such combinations in corpora (as yet unlisted forms). The following are examples, based on the root *-bon-* "see":
(17) *isi-bon-i* "mourner (lit. one who looks on at a funeral)"
Class.pref.cl 6/7-verb.root-deverb.suffix
(18) *um-bon-i* "one who sees"
Class.pref.cl 1/2-verb.root-deverb.suffix
(19) *um-bon-o* "apparition, vision"
Class.pref.cl 3/4-verb.root-deverb.suffix
(20) *isi-bon-is-o* "signpost, signal"
Class.pref.cl 6/7-verb.root-caus-deverb.suffix
(21) *isi-bon-is-el-o-ana* "small example"
Class.pref.cl 6/7-verb.root-caus-appl-deverb.suffix-dim

4. Towards implementation

The capturing of data in a rigorous, systematic and appropriately structured way is of utmost importance to ensure that data exchange, in this case between the machine-readable lexical database (MR lexicon) as the source and other applications (for example the morphological analyser), is consistent.
Important considerations in the choice of database implementation include the *type* of data that has to be stored, the different *views* of and *access* to the data that may be required by applications, and the *amount* of data that needs to be stored.
Type: Lexical data is semi-structured (Manning and Parton, 2001) and a chosen database environment for the

[1] *o-* and *i-* are used productively for the formation of impersonal and personal deverbatives respectively, while *a-*, *e-* and *u-* seldom occur and are non-productive.

capturing of lexical information should allow for recursion. In particular, it should allow for multiple occurrences (recursion) of certain substructures such as `Ext` and `NominalSuffix`, resulting in n-depth structures that are reused throughout the representation of such information.

Views and access: For the purposes of rebuilding the morphological analyser (see Figure 1) the preferred view would be the morphological structure of all the word roots in the database while access would be sequential. It should be clear that for other applications different views may be required and access may even be random.

Amount of data: It is estimated that the Zulu MR lexicon should make provision for between 40 000 and 50 000 entries, each of which would have its own specific associated lexical information (see Figure 2).

XML and Unicode are *de facto* standards for mark-up and encoding. Therefore, only Native XML and XML-enabled databases are considered.

An XML-enabled database is a database with extensions for transferring data between XML documents and its own data structures. XML-enabled databases that are tuned for data storage, such as a relational or object-oriented database, are normally used for highly structured data.

A native XML database is one that treats XML documents and elements as the fundamental structures rather than tables, records, and fields (Harold, 2005). Native XML is more suitable for unstructured data or data with irregular structure and mixed content.

Since lexical data are *semi-structured* there are *two* choices: Either endeavour to fit the data into a well-structured database, or store it in a native XML database, designed to handle semi-structured data (Bourret, 2010). In subsequent sections the various options are briefly discussed.

4.1 Native XML databases

Native XML databases (NXD) allow for user-defined schemas. For Zulu such a schema will be based on the proposed data model (given in DTD notation) (Bosch et al., 2007) and its extension (Section 3). This approach to lexical database development was, for instance, also successfully used for Warlpiri, an Indigenous Australian language (Manning and Parton, 2001). Query languages such as XQuery (a W3C standard) and XUpdate, particularly suitable for database-oriented XML, may then be used to query and update the database.

A fragment of the XML code for example (21), according to the DTD in section 3, is as follows:

```
<Verb>
  <Root>bon</Root>
  <VerbFeatures>
    <Label>v</Label>
    <Transitivity>t</Transitivity>
    <Sense>see</Sense>
    <Ext> <!-- Orth. form: bonisa -->
      <Caus>is</Caus>
      <Transitivity>t</Transitivity>
      <Sense>show; cause to see</Sense>
      <Ext> <!-- Orth.form:bonisela -->
        <Appl>el</Appl
        <Transitivity>t</Transitivity>
        <Sense>look after for
        </Sense>
        <DeverbSuffix>o</DeverbSuffix>
        <DeverbFeatures>
          <ClassPrefSg>
            isi
          </ClassPrefSg>
          <ClassPrefPl>
            izi
          </ClassPrefPl>
          <ClassNo>6-7</ClassNo>
          <Label>n:dev</Label>
          <NominalSuffix>
            <Dim>
              <Form>ana</From>
              <Sense>small example</Sense>
            </Dim>
          </NominalSuffix>
        </DeverbFeatures>
      </Ext>
    <Ext>
  </VerbFeatures>
</Verb>
...
```

Using XQuery to return the English translation for *–bon-* in example (21):

```
for $x in doc("verb.xml")/verb
where $x/Root="-bon-"
return $x/VerbFeatures/Sense
```

Result:

```
<?xml version="1.0" encoding="UTF-8"?>
<Sense> see </Sense>
```

4.2 XML-enabled databases

Alternatives to NXDs are XML-enabled databases such as relational and object-orientated databases.

4.2.1 Relational databases

Relational databases are a *de facto* standard for operational and analytical applications (Lin 2008; Naser et al., 2007). Standard query languages (such as SQL) can be used to query the relational database and both commercial and open source software is available for parsing of XML. However, representing the recursive structure of the verbal extensions and the nominal suffixes in a relational database is problematic since the traversal of n-depth structures results in a large number of small tables with "artificially" generated pointers. Arenas and Libkin (2008) investigated the theoretical issues of data exchange in XML documents and indicated that it is not always a clear-cut situation to assume that a set of data in a database shall always produce consistent data, due to the restrictions that exist in a relational database.

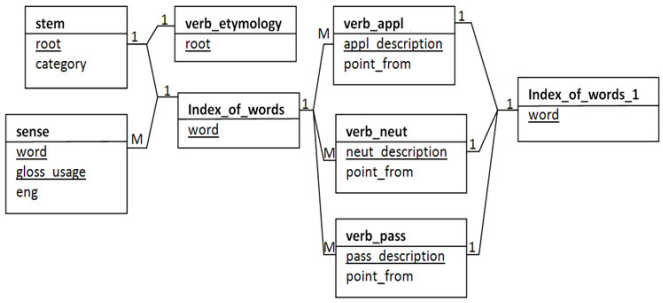


Figure 4: Entity-relationship diagram for `Ext`-structure

Figure 4 shows a fragment of the entity-relationship diagram (ERD) that was created to represent the `Ext`-structure in a relational database. Each individual representation is captured in the `Index_of_words` table and then this instance is represented in the relevant table, e.g. `verb_appl` or `verb_neut`. The field `point_from` will be used as a "pointer" that refers back to the previous element from where the new element is derived. The sense of each new element is captured in the `sense` table. The ERD is a possible solution for `Ext`, but rendering DTD compliant XML from this structure and producing consistent data may be time consuming since each individual table needs to be converted to XML and then merged to generate the final XML to be used for archiving and inter-system usage.

4.2.2 Object-oriented databases

An object oriented database supports recursion in the modelling of the verb structure and deverbatives in Zulu. An exposition of the details and full complexity of this approach for the implementation of the data model falls outside the scope of this article. We show only the basic idea by applying it to the verb root and its verbal extensions as follows:

The class `VRoot` occupies the highest level in the class hierarchy for extended verb roots, followed by six subclasses, viz. `VRoot+Appl`, `VRoot+Caus`, `VRoot+Intens`, `VRoot+Neut`, `VRoot+Pass` and `VRoot+Recip`. In turn, each subclass again has its own six subclasses, viz. `VRoot+Appl+Appl`, `VRoot+Appl+Caus`, …, and so on. This hierarchy is dynamically expanded as new extended roots are identified. Unlike an entity in a relational database, an object includes not only information about the relationships between facts within an object, but also information about its relationship with other objects.

A particular verb root (say *-bon-*) and its extensions are then individual objects (instantiations) in the appropriate classes. Each such instantiation has its own unique identity (OID) and inherits specific attributes and methods from its super class/parent object, making it quite intuitive to present the lexical data in an object-oriented database.

In example (6) more than one extension is suffixed to the verb root. The classes of interest in this example are `VRoot`, `VRoot+Caus`, and `VRoot+Caus+Recip` and the instantiated objects are for *-bon-*, *-bon-is-* and *-bon-is-an-,* as shown in Figure 5, where the arrow indicates inheritance of information via the class hierarchy. An example of inherited information is the `Etymology` of the basic root, while lexical information that is associated with each unique object is the `Sense` information. Figure 5 illustrates how such recursion is captured in an object-oriented approach.

Naser et al. (2007) describe the "forward engineering" process whereby the object oriented database environment (including inheritance and nesting) is used as input and a corresponding XML schema and document(s) are produced as output. The "two-way mapping" algorithm (Naser et al., 2009) may be used to transform the data in the object-oriented approach, firstly to flat XML, then to the nested XML schema and lastly to the final XML document.

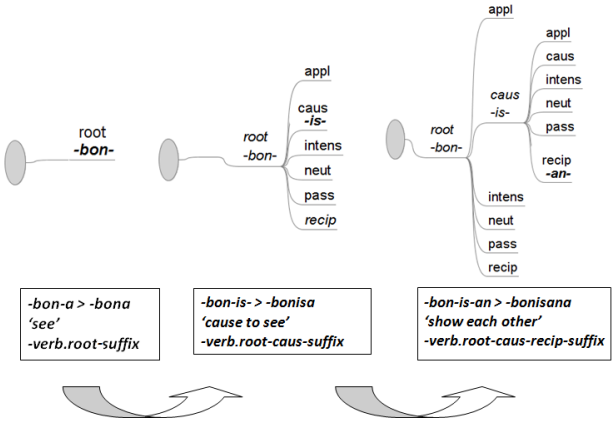


Figure 5: Recursion as objects

5. Conclusion and future work

In this article we described the refinement of a data model for a Zulu MR lexicon, focusing on the modelling of two sources of recurrence in the verb structure, viz. verbal extensions and deverbatives. We also showed how the MR lexicon, based on the proposed data model, may be embedded in an update framework. Finally we proposed Native XML and object-orientated databases as two possible approaches to implementation. Reasons for this may be summarized as follows:

Type of data: Object oriented databases are designed to work well with state of the art object oriented programming languages. Object oriented databases use the same model as these programming languages as they store and index theoretical objects. "Object databases are generally recommended when there is a business need for high performance processing on complex data" (Foster, 2010).

XML Databases offer the same functionality as Object oriented databases. The data is structured in a hierarchical manner except that Native XML databases store XML documents instead of theoretical objects. While this is conceptually the same data storage, XML databases have

the added benefit of being able to exchange the data in its native format – no overheads are incurred by transformation and mapping of the data to and from other database structures or representation.

Views and access: Where Object Databases have Object Query Language (OQL), XML Databases have XQuery, which is a W3C standard. Multiple views of the data, as well as appropriate access, are supported via these powerful query languages.

Amount of data: Both native XML and object oriented databases are able to process arbitrarily large documents. Moreover, these approaches are claimed to offer high performance since queries over a well-designed, well-implemented native XML or XML-enabled object oriented database are faster than queries over documents stored in a file system – particularly in cases where queries are significantly more frequent than insertions and updates, which is expected to be the case for a comprehensive, mature and stable MR lexicon.

Work planned for the near future includes

- the development of prototypes for the MR lexicon for Zulu in order to investigate, demonstrate, evaluate and compare the two possible approaches to implementation;
- the development of software support for semi-automating the lexicon update framework;
- the bootstrapping of MR lexicons for various other Bantu languages from the Zulu lexicon.

6. Acknowledgements

This material is based upon work supported by the National Research Foundation under grant number 2053403. Any opinion, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Research Foundation.

7. References

Arenas, M., Libkin, L. (2008). XML data exchange: consistency and query answering, *Journal of the ACM*, 55(2). [Online]. Available: http://portal.acm.org/citation.cfm?id=1346332 (accessed February 2010).

Bosch, S.E., Pretorius, L. & Jones, J. (2007). Towards machine-readable lexicons for South African Bantu languages. *Nordic Journal of African Studies* 16(2), pp.131–145.

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.

Bourret, R. (2009). XML Database Products:
XML-Enabled Databases. [Online]. Available: http://www.rpbourret.com/xml/ProdsXMLEnabled.htm (accessed March 2010).

Doke, C.M. & Vilakazi, B. (1964). *Zulu-English Dictionary*. Johannesburg: Witwatersrand University Press.

Foster, C. (2010). XML databases – The business case. *CFoster.net*. [Online]. Available: http://www.cfoster.net/articles/xmldb-business-case/ (accessed March 2010).

Harold, E.R. (2005). Managing XML data: Native XML databases - Theory and reality. IBM developerWorks. [Online]. Available: http://www.ibm.com/developerworks/xml/library/x-mxd4.html (accessed March 2010).

Lin, C.Y. (2008). Migrating to relational systems: Problems, methods, and strategies. *Contemporary Management Research*, 4(4), pp.369–380.

Manning, C. & Parton, K. (2001). What's needed for lexical databases? Experiences with Kirrkirr. *Proceedings of the IRCS Workshop on Linguistic Databases*, pp. 167-173. University of Pennsylvania, Philadelphia.

Naser, T., Alhajj, R. & Ridley, M.J. (2009). Two-way mapping between object-oriented databases and XML. *Special Issue: Information Reuse and Integration*. R. Alhajj (Ed.), 33, pp.297–308.

Naser, T., Kianmehr, K., Alhajjb, R. & Ridley, M.J. (2007). Transforming object-oriented databases into XML. *Proceedings of the IEEE International Conference on Information Reuse and Integration*, IRI 2007. pp.600–605.

Poulos, G. & Msimang, C.T. (1998). *A Linguistic Analysis of Zulu*. Pretoria: Via Afrika.

Pretorius, L. & Bosch, SE. (2003). Finite-state computational morphology: An analyzer prototype for Zulu. *Machine Translation – Special issue on finite-state language resources and language processing*, 18, pp.195–216.

Van Eeden, B.I.C. (1956). *Zoeloe-Grammatika*. Stellenbosch: Universiteits-uitgewers.

Weber, D.J. (2002). Reflections on the Huallaga Quechua dictionary: derived forms as subentries. [Online]. Available: http://emeld.org/workshop/2002/presentations/weber/emeld.pdf (accessed February 2010).

Improving Orthographic Transcriptions of Speech Corpora

Nico Laurens Oosthuizen, Martin Johannes Puttkammer, Martin Schlemmer
Centre for Text Technology (CTexT™),
North-West University (NWU), Potchefstroom Campus, South Africa
{nico.oosthuizen, martin.puttkammer, martin.schlemmer}@nwu.ac.za

Abstract

Employing multiple transcribers for the task of orthographic transcription can cause inconsistent transcriptions on various levels. We describe a method for comparing different orthographic transcriptions of similar utterances in order to ensure consistency. In order to ensure that the transcriptions are as error free as possible, confusables, splits, insertions, deletions and non-words are identified on the orthographic level. Comparison of transcriptions is performed by mapping erroneous transcriptions to the intended transcription using Levenshtein distance. Errors are annotated using HTML mark-up to provide a visual representation of the differences. Since it is difficult to automate the correction of these errors, it is performed manually, while each change is checked in conjunction with the relevant speech file. It is indicated that this method decreases the number of transcription errors.

1. Introduction

The Lwazi (meaning "knowledge") project is a telephone-based, speech-driven information system that was commissioned by the South African Department of Arts and Culture to provide South Africans with access to government information and services in any of the 11 official languages. The Lwazi project was conducted by the Human Language Technology (HLT) Research Group of the Meraka Institute of the CSIR. See http://www.meraka.org.za/lwazi for more information.

The Centre for Text Technology (CTexT™) was sub-contracted by the Meraka Institute for the collection and transcription of the automatic speech recognition (ASR) data for the eleven official languages of South Africa.

For each language a total of 200 mother tongue speakers were recorded via landline or mobile telephone and each recording was transcribed manually. Prompt sheets consisting of 30 utterances were compiled, divided into two sections; 14 open questions (e.g. How old are you?) and 16 phoneme-rich sentences.

Roughly 350 phoneme-rich sentences were selected from various corpora and randomly used to populate the prompt sheets. This resulted in each sentence being recorded 6-10 times, totalling 3, 200 phonetic rich sentences for each language.

The project spanned over approximately 2 years and 4-6 transcribers were employed per language. Despite various phases of quality control, differences within similar sentences were discovered. As the ASR corpora developed were relatively small (479 speech minutes being the maximum in the corpora for one of the languages), the quality of the transcriptions needed to be extremely accurate to ensure a usable ASR system. Introducing yet another quality control phase yielded no significant results and another method was needed to identify and correct errors.

2. Identified Differences

After evaluating some transcriptions, we found that the differences could be grouped into 5 categories:

English examples:

- Confusables
  - has it been tried on <too> small a scale
  - has it been tried on <to> small a scale
- Splits
  - there's <nowhere> else for it to go
  - there's <no_where> else for it to go
- Insertions
  - so we took our way toward the palace
  - so we <we_>took our way toward the palace
- Deletions
  - as <to_>the first the answer is simple
  - as the first the answer is simple
- Non-words
  - there is no <arbitrator> except a legislature fifteen thousand miles off
  - there is no <abritator> except a legislature fifteen thousand miles off

isiXhosa examples:

- Confusables
  - andingomntu <othanda> kufunda
  (I'm not a person <who loves> to read)
  - andingomntu <uthanda> kufunda
  (I'm not a person <you love> to read)
- Splits
  - alwela phi na <loo_madabi>
  - alwelwa phi na <loomadabi>
  (Where is it taking place, those challenges)

- Insertions
  - ibiyini ukuba unga mbambi wakumbona
  (Why didn't <you catch> him or her when you saw him or her)
  - ibiyini <na_>ukuba unga mbambi wakumbona
  (Why didn't < you caught> him or her when you saw him or her)

- Deletions
  - yagaleleka impi <ke_>xa kuthi qheke ukusa
  - yagaleleka impi xa kuthi qheke ukusa
  (It started the battle at the beginning of the morning)

- Non-words
  - yile <venkile> yayikhethwe ngabathembu le
  - yile <venkeli> yayikhethwe ngabathembu le
  (It is this shop that was selected by the Bathembu) <venkeli> in the second example is a spelling mistake.

Setswana examples:

- Confusables
  - bosa bo <jang> ko engelane ka nako e
  - bosa bo <yang> ko engelane ka nako e
  (How is the weather in England at this time?) <yang> in the second example is slang for "how".

- Splits
  - le fa e le <gone> re ratanang tota
  - le fa e le <go_ne> re ratanang tota
  (Even though we have started dating)

- Insertions
  - ba mmatlela mapai ba mo alela a robala
  - ba mmatlela mapai <li-> ba mo alela a robala
  (They have looked for blankets and made a bed for themselves to sleep)

- Deletions
  - ke eng gape se seng<we> se o se lemogang
  - ke eng gape se seng se o se lemogang
  (What else have you noticed?)

- Non-words
  - lefapha la dimenerale le <eneji>
  - lefapha la dimenerale le <energy>
  (Department of minerals and energy)
  < energy > in the second example is a spelling mistake.

These errors were mainly attributed to the inexperience of the transcribers as they were usually students with no formal linguistic training and no experience in transcribing data.
Implementing a method to identify and correct these differences would decrease the number of transcription errors, thus increasing the overall quality of the transcriptions completed per language.

3. Methodology

Figure 1 shows a flowchart that illustrates the method used to improve the transcriptions. It is important to note that the comparison of transcriptions to the original sentence was only performed on the 16 phoneme rich sentences in each recording as the other 14 utterances were answers to open questions and thus did not have a relevant original sentence to be compared to.
This method is implemented in four steps, which are discussed in more detail in the following subsections.

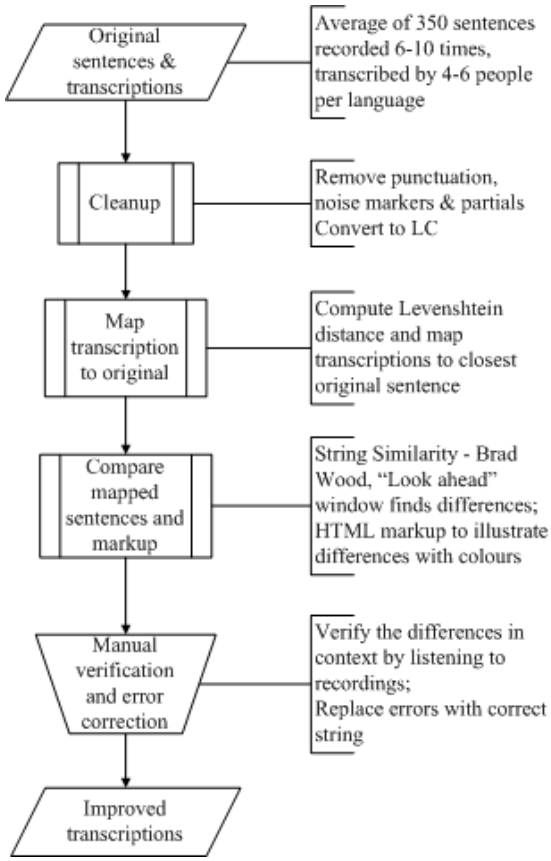


Figure 1 : Process Flowchart

3.1 Cleanup

The first step is to remove any punctuation, noise markers and partials present in both the original and transcribed sentences. This was done as even perfectly transcribed sentences could still differ from the original due to added information such as the indication of noises and partials. Commas were also used to indicate pauses apart from normal usage and needed to be removed as well.
For example, the original sentence "My life in Georgetown was uneventful." was transcribed as "my life in [n] Georgetown was, uneventful." with the noise marker [n] indicating an external noise and the comma

after “was” indicating a long pause. All sentences were also converted to lowercase as the annotation protocol stipulated that sentences should not start with a capital letter.

3.2 Transcription mapping

A Levenshtein distance can be used to determine similarities between two different strings. Thus, after cleanup was performed, the Levenshtein distance was computed between each original (O) and transcribed (T) sentence and then ranked according to the Levenshtein distance between the sentences.

The Levenshtein-distance is computed by allocating a cost of 1 for insertions and deletions and a cost of 2 for substitutions. The minimum cost is calculated between each character of the original and each character of the transcribed sentence. This cost is then used to construct an edit distance matrix using the following distance formula: (Jurafsky, & Martin, 2009; Levenshtein, 1966).

$$dis\tan ce[i,j] = \min \begin{cases} dis\tan ce[i-1,j] + ins\cos t(t\arg et_{i-1}) \\ dis\tan ce[i,j-1] + ins\cos t(source_{j-1}, t\arg et_{i-1}) \\ dis\tan ce[i,j-1] + del\cos t(source_{j-1}) \end{cases}$$

The resulting minimum edit distance is then normalised as a percentage of the maximum sentence length of the two given sentences, which is then used to map each transcribed sentence to the closest original sentence.

In cases where there is no difference between the original and transcribed sentence (DIFF (O, T) = 0), nothing further is done with the sentences, and the following steps are only applicable if DIFF (O, T) = 1. This results in a mapping of transcriptions containing differences to an original sentence:

Original sentence (O):

- slighter faults of substance are numerous

Transcriptions (T):

- slighter fault substance are numerous 90.20%
- slighter faults of substances are numerous 97.60%

In this example the first transcribed sentence matches the original sentence with a percentage of 90.20 as the word “of” was omitted as well as the character “s” in “faults”. The second transcribed sentence is a closer match at 97.60% as it only contains 1 insertion, the “s” in “substances”.

3.3 Comparing mapped sentences and mark-up

Next we use a string comparison algorithm developed by Brad Wood (2008) to identify the differences between the original sentence and each transcribed sentence mapped in the previous step. This is to be done by finding the Longest Common String (LCS). The original code was written by Brad Wood in CFScript and ported to Perl for our purpose.

The operation of Wood's algorithm is as follows: Two strings are read into the function and a windowing method is implemented to compare parts of each string to one another. The window over the first string is kept in a stationary position while a second window over the other string is moved until the maximum search distance is reached. Each character in the second window is compared to each character in the first window. The offset adjusts for extra characters in either window.

Any differences found are annotated with HTML tags with different background colours to highlight the differences visually. For example, in the first transcribed sentence “slighter fault substance are numerous” the “s”, a space and the “of” was omitted. This will be represented as follows in HTML:

- slighter faults of substance are numerous

The process is then reversed by keeping the first window stationary and moving the other window so that matches can be found in both directions. This is repeated until both windows reach the end of the relevant string. The result is a HTML document that illustrates all differences which may occur, in various colours.

3.4 Manual Verification

For the purpose of manual verification, all sentences with DIFF (O, T) = 1, are examined in two ways. The spoken utterance (U) is compared to the original sentence (O) and to the transcription (T). If DIFF (O, U) = 1 and DIFF (T, U) = 0, then U = T and no change is needed. If DIFF (O, U) = 1 and DIFF (T, U) = 1, the transcription is incorrect and needs to be manually checked.

This verification was done for each language by one of the transcribers initially used for the transcriptions that proofed to be the most competent. Verification by a mother tongue speaker is needed as a difference does not necessarily indicate an error and should be verified by listening to the spoken utterance and not just by correcting the transcription according to the original sentence. This is illustrated in the following example:

Original sentence (O):

- “I told him to make the charge at once.”

Spoken utterance (U)

- “I told him to make the change at once.”

Transcriptions (T):

- “I told him to make the cha<n>ge at once.”

In this example the speaker deviated from the prompt sheet and said “change” instead of “charge” and no correction was needed.

Human intervention is required to ensure that the identified differences are indeed errors and that accurate transcriptions are not replaced with the original transcription in error.

In the following example the sentence was read as it appeared on the prompt sheet, but transcribed incorrectly

and had to be replaced:

Original sentence (O):

- "a heavy word intervened, between..."

Spoken utterance (U)

- "a heavy word intervened, between..."

Transcriptions (T):

- "a heavy wo<o>d intervened, between..."

After each error is verified, replacements are done to the transcriptions automatically.
It is possible that DIFF (O, U) = 1 and DIFF (O, T) = 0, indicating that DIFF (T, U) = 1, but the comparison was only performed if DIFF (O, T) = 1. This comparison should be included in future work.

4. Results

We evaluate our method in terms of improvement of transcription accuracy. All evaluations, results and findings are listed below:

Language	Differences found	Actual errors
Afrikaans	776	152
English	1143	337
isiNdebele	958	291
isiXhosa	1484	1081
isiZulu	1854	1228
Sepedi	1596	736
Sesotho	739	261
Setswana	1479	828
Siswati	1558	351
Tshivenda	814	191
Xitsonga	1586	456

Table 1 : Number of differences Found and Actual Errors

Table 1 shows the number of differences found (DIFF (O, T) = 1) as well as the number of actual errors (DIFF (T, U) = 1). These differences were found in the 3, 200 phonetic rich sentences for each language. Although the actual errors are low compared to the differences identified, this can mainly be contributed to speakers deviating from the prompt sheet (DIFF (T, U) = 0). Afrikaans had the least amount of errors, but still contained 152 errors. The errors found in isiXhosa and isiZulu indicate that about one in three sentences contained an error. Identifying and correcting these errors made a significant improvement in the transcription accuracy, ensuring a usable and valuable resource for these languages.

5. Conclusion

In this paper we introduced a method for identifying differences in ASR data. Correcting these differences decreased the number of transcription errors, thus increasing the overall quality of the transcriptions, resulting in a usable ASR system. The final data distributed to the Meraka Institute for the Lwazi project had an average transcription accuracy of 98%.
This method can be used in the compilation of ASR data for other resource-scarce languages, hopefully resulting in usable ASR systems. Even though the cause of the differences was attributed to (initially) inexperienced transcribers, the results show that it is still possible to achieve high accuracy. This is especially relevant in cases where trained linguists with experience in orthographic transcriptions are difficult to obtain. A further advantage is that employment opportunities are providing to people that would otherwise not have been employed for the task due to their lack of training and/or experience. Providing training to these transcribers during the project also resulted in competent transcribers by the end of the project and these transcribers could prove invaluable in future projects involving orthographic transcriptions.

6. References

HTL-Group, Meraka Institute, CSIR (2010). Lwazi, *A telephone-based, speech-driven information system*. [Online]. Available: http://www.meraka.org.za/lwazi (accessed February 2010).

Jurafsky, D. & Martin, J.H. (2009). Speech and Language Processing: *An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics*. 2nd edition. Prentice-Hall, pp. 74-77.

Levenshtein, V.I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Cybernetics and Control Theory*, pp. 707-710.

Wood, B. (2008). ColdFusion Levenshtein Distance: *String comparison and highlighting*. [Online]. Available: http://www.bradwood.com/string_compare (accessed February 2010).

Towards The Manding Corpus: Texts Selection Principles and Metatext Markup

Artem Davydov
St. Petersburg State University, Faculty of Oriental and African Studies
199 034, Universitetskaya nab. 9/11, St. Petersburg, Russia
artdavydov@mail.ru

Abstract

In the present paper I discuss some issues related to the development of the Manding (Mande family, Niger-Congo) corpus. In the first part of the paper I give a brief survey of the potential text sources for the corpus, emphasize the problems related to their usage and make an attempt to formulate some basic principles of texts selection. In the second part I propose a typology of texts to be used for the metatext markup in the corpus.

1. Introduction

Manding is the branch of closely related languages inside the Mande family, Niger-Congo. In the present paper I will discuss some basic principles of texts selection for the electronic corpus of Manding, which is now being built, and give some preliminary remarks on the metatext markup in this corpus. The major language in question is Bamana (spoken in Mali), although the parsing tool we develop will be potentially able to work with some other Manding languages (Maninika, Dyula). The potential range of users of the corpus, as we see it, is relatively wide and includes Mandeists (both grammarians and lexicographers), typologists, specialists in language development and native speakers of Manding who are interested in their own language (school teachers, publishers, etc.).

The development of the corpus is only in its initial stage, therefore the present paper is rather a material for discussion, than an account of how our corpus is already arranged.

2. Texts selection principles

In (Sinclair 2005) two types of criteria of text selection are discussed: the *external* criteria, which take into consideration the communicative function of a text, and the *internal* criteria, which reflect details of the language. Sinclair emphasizes that corpora should be constructed exclusively on external criteria, i.e. the contents of a corpus should be selected disregarding the language they contain. I share Sinclair's opinion, as the scholar's goal is to describe the linguistic reality rather than to alter it; but accepting the external criterion as the only one for building the Manding corpus we face certain difficulties.

Below I give a brief survey of the available text sources for the corpus. As I will attempt to show in this paragraph, building a corpus of a language recently put into writing, we cannot but move towards the language standardization. We have no other choice than to use the internal criteria along with the external criteria, at least at the initial stage of the corpus building.

2.1. Oral texts

The Malian society is not monolingual. Bamana-French code-switching is a common practice, especially among the intellectual elite; unadapted French loanwords are frequent in Bamana oral texts. Thus, the Bamana corpus, if we want it to reflect the real everyday languaging, requires the parsing tool not only for Bamana, but also for French. Still, there are some sources which represent the oral form of Bamana in its "pure" form, such as transcriptions of speeches and interviews made by linguists and ethnologists in villages, where the French language is not in use. Transcriptions of TV- and radio programs should only be used with great prudence for building the oral speech sub-corpus. Two groups of texts must be distinguished here:

- spontaneously generated texts;
- texts which had been first written and then pronounced.

The texts of the second group (movie scripts, pre-written radio programs, etc.) can not be considered as oral, or, at least, they should be attributed to a separate category of oral texts. This point may seem self-evident, but, surprisingly, in many linguistic corpora (in the Russian National Corpus, for example) these two groups of texts are undistinguished.

2.2. Written texts

As concerns written texts, the following types are relevant:

2.1.1. Published folklore texts

Such texts as popular tales, anecdotes, epic poetry, etc. are of great importance, as far as they represent, in fact, oral speech. Fairy-tale publications are numerous, anecdotes are regularly published in the *Kibaru* newspaper (Bamako) and there are a lot of scientific publications of epic Manding texts, so there are no problems with the availability of texts.

2.1.2. Fiction books

Like transcriptions of TV- and radio programs, fiction

books must be used with great caution, for the language it may contain is often highly unnatural, as it represents a translation of a mediocre quality from French. For instance, a Bamana novel "Kanuya Wale" was first written by Samba Niaré's in French (under the title "Acte d'amour") and then translated by the author into Bamana. The author himself mentioned that he first wrote the novel in French, because he could not express his ideas directly in his mother tongue. Thus, the novel contains syntactic structures, word usages and even morphological features untypical for the spoken Bamana. As a result, the novel is sometimes almost incomprehensible without the support of the French original text.

Writing first in French, then translating the texts into Bamana is rather common for Malian writers. The question is, should we disregard translated texts at all (it was the decision of the developers of the Russian National Corpus) or should we consider French calques to be a characteristic trait of the modern Bamana literature and include texts like "Kanuya Wale" into the corpus? When a corpus for a language with a long written tradition is created, it is easy to avoid translations from other languages, since original texts are readily available. But can we afford the same working with Bamana?

2.1.3. The press

The press is another potential source of corpus materials, though it raises the same questions as the fiction literature.

2.1.4. Educational and religious literature

The other possible sources are: school textbooks, educational booklets, religious literature, etc. Judging by the bibliography compiled by G. Dumestre (1993), the diversity of genres of the Bamana literature is rather rich. Of course, this diversity must be adequately reflected in the corpus.

2.1.5. Nko publications

Finally, there is another type of texts we should not forget about – the Nko publications. The original N'ko script was invented in 1940s for the Maninka language by a Guinean self-educated scholar Suleymane Kanté. Since then N'ko has evolved to a rich written tradition and a strong cultural movement which has many followers in Manding-speaking communities all over the West Africa, especially in Guinea. The followers of the N'ko movement believe that various Manding languages (Bamana, Maninka, Mandinka, etc.) are the dialects of the same language called "N'ko" (*ń kó* means 'I speak' in the majority of the Manding languages). Some steps towards the official recognition of N'ko have been made in Guinea, where it is used for formal schooling.

Using N'ko texts for building a corpus has both advantages and disadvantages. The most advantageous point is that in N'ko publications all tones are indicated. Unfortunately, disadvantages are even more numerous. First of all, using the N'ko script in the corpus will diminish the range of its potential users. Of course, this problem is solvable as it is possible to automatically convert N'ko into Roman letters. Secondly, N'ko is not just an alphabet, but a whole written tradition. The N'ko written language has significantly evolved from the spoken Maninka. The texts in N'ko contain numerous neologisms created by the followers of the N'ko movement, and even syntactical constructions untypical of the spoken Maninka. If N'ko texts constitute the main bulk of the corpus, we run the risk to have a corpus of a variety which has only a limited use among the followers of the N'ko movement. Finally, despite the pan-Manding ideology of the N'ko movement, the absolute majority of N'ko publications are in Maninka. However, the popularity of N'ko in Mali is growing; some field work needs to be done in order to find out if any N'ko texts in Bamana (published or unpublished) are already disposable.

2.3. Conclusion

As it has been shown above, certain types of texts raise doubts. The question is, should they be included into the corpus or not. In my opinion, it must be answered positively, for, under existing conditions of lack of resources, it would be too wasteful to reject any text data. On the other hand, every text included into the corpus must be provided with detailed metadata in order to allow the user to single out the sub-corpora he is interested in.

At the same time, at the initial stage of the corpus building the preference must be given to the texts which represent the "pure" Bamana language (folklore, village recordings, etc.) in order to develop an operational parser/tokenizer. At the next stage, more complex and "artificial" texts can be added.

3. Metatext Markup

The Manding texts classification proposed below takes into account the EAGLES recommendations on text typology (Sinclair 1996) and the metatext markup of the Russian National Corpus, as described by Savchuk (2005), but is accommodated to the Manding data specifics. To describe each text in the corpus, I propose various parameters which can be united into several groups:

- text metadata;
- data about the author;
- linguistic metadata;
- subject matters;
- technical metadata.

In the subsequent paragraphs these groups of parameters are discussed in more detail.

3.1. Text metadata

1. The *title* of the text (if any).
2. The *date* of creation of the text (exact or approximate), which is not to be confused with the date of publication.
3. The *size* of the text (in words).
4. *Original text* or *translation*. If the text is a

	Bamana		Maninka	
TAM	**Affirmative**	**Negative**	**Affirmative**	**Negative**
Perfective (transitive)	*-ra*	*má*	*-da, -ra*	*má*
Perfective (intransitive)	*yé*		*kà*	
Present	*bɛ́*	*tɛ́*	*yé…-la*	*tɛ́…-la*
Future	*bɛ́nà*	*tɛ́nà*	*dí, dínà*	*tɛ́*
Optative	*ká*	*kána*	*yé*	*ká, kána*

Table 1: Bamana and Maninka TAM-systems

translation, then the language of the original must be indicated.

5. *Channel: written* or *oral*.
6. Bibliographical data (for published texts): the place and the year of publishing; the name of the publishing house; the number of pages. For periodicals: issue number.
7. For manuscripts: the place and the date of creation (if known); the form of the text (handwritten, typewritten, electronic).
8. For oral speech events: the place and the date of recording.

3.2. Data about the author

This point includes the following parameters:

1. *Author's full name* (if known). If the author is using an alias, his real name must also be indicated (if known). If the text has several authors, all of them must be listed. Note: In the Manding corpus, information concerning the author's name is more useful, than, for example, in the corpora of European languages, as it indicates ethnic and caste origin of the author.
2. *Author's age* at the moment of creation of the text (exact or approximate).
3. *Author's gender* (male, female or unknown).
4. *Author's fluency in language* (native- or non-native speaker). Texts written in Bamana (or in any other Manding language) by non-native speakers are few, but the quantity of texts generated by non-native speakers will significantly increase when it comes to the sub-corpus of oral speech, the Manding languages being broadly used as lingua franca in West Africa.

3.3. Linguistic metadata

3.3.1. Manding variety

The Manding corpus is intended to be multilingual, although on the initial state, only Bamana data will be included. It is potentially possible to use the same morphological parser for three out of four major Manding varieties (Bamana, Maninka and Dyula), but not Mandinka, as its morphology differs significantly from the other varieties. Thus, the main options for the *Manding variety* parameter are: Bamana, Maninka, Dyula. Here, we face certain difficulties connected to the fact that borders between Manding varieties are only vague. In this point, only a conventional decision is possible. The most convenient criterion for distinguishing the major Manding languages is the system of verbal TAM (tense, aspect, modality) markers. Table 1 shows the TAM markers in Bamana and Maninka of Guinea.

In addition, the linguistic diversity inside the major Manding varieties is considerable, although "normative", or "standard" variants are available ("standard Bamana", based on the dialect of Bamako; Maninka-Mori, based on the dialect of Kankan). Dialect tags must be assigned to the texts to the extent possible. At the same time, texts written or pronounced in standard Bamana may have some characteristics of other dialects, depending on the origin of the writer/speaker. Two decisions are possible here: to assign only one dialect tag to each text depending on a quantitative criterion, or to indicate both dialects. The second decision seems preferable, as the presence of two dialect tags itself will tell the user that the text represents a "mixed" idiolect.

Dialectal tags can be assigned according to the external criterion (if the origin of the speaker/writer is known) or to the internal criterion (if the dialectal origin of the data can be identified reliably).

3.3.2. Writing system

Manding uses various writing systems. The main options for the *Writing system* parameter are:

1. *Malian Roman-based orthography*, which exists in two variants, old and new. The absolute majority of texts published in Mali use this system of orthography.
2. *French-based spontaneous orthography* used by people literate in French, but illiterate in their mother tongue. This system is not standardized, but texts written in this "orthography" have much in common: digraphs *ou* and *gn* for /u/ and /ɲ/, respectively, acute accents on the final vowel, etc.
3. The *Adjami* writing system based on the Arabic alphabet. Like the French-based spontaneous orthography, it is not standardized and exists in many local and individual variants.
4. The *N'ko* writing system, which has already been mentioned in p. 2.1.5.

5. Other Roman-based orthographies used for various Manding idioms in different countries of West Africa.

As shown by Vydrine (2008, 19-20), it is technically possible to establish correlations between different graphical forms of the same word in the Manding corpus. Each text in the corpus is stored in two graphical forms: in its original form and in "standard" transcription. The latter is very much alike the new Malian Roman-based orthography, but as opposed to it, has tonal markers.

3.4. Subject matters

The *subject matters* parameter is used for the metatext markup in every major European language corpus I know. The majority of these corpora are based on the EAGLES text typology recommendations. For the Manding corpus I propose a more concise list of subject matters for two reasons: firstly, the variety of texts in Manding is inferior in comparison with any European language; secondly, a comparatively small corpus that we are building would not need such a detailed subject matters classification as a large one. For the same reasons I do not propose to introduce the second dimension to the subject matters classification, although some corpus developers do that. For instance, in the Russian National Corpus, two dimensions are used: *subject matters* and so-called *functional spheres* (journalism, technical, academic, official and business, day-to-day life, advertising, theological, electronic communication). On the other hand, in many large corpora (such as British National Corpus), only one dimension is successfully used.

The subject matters classification I propose (which does not pretend to be exhaustive) reflects the real functioning of the Manding languages. So far, it consists of twenty six tags, and some of them are united into groups:

1. *Folklore: Fairy-tales, Anecdotes, Epics, Proverbs, Traditional Songs Lyrics.*
2. *Fiction: Prose, Movie Scripts, Theatre, Poetry, Popular Songs Lyrics.*
3. *Religious: Christian, Islamic, Other.*
4. *Educational: School Textbooks, Language and Literacy, Health, Agriculture, Science.*
5. *Personal: Personal Records, Correspondence.*
6. The other tags are: *Journalism, Sports, History, Advertising, Business, Everyday Life.*

Tags can be combined without any limits. For example, a school textbook in chemistry will acquire the following tags: *Educational: School Textbooks* and *Educational: Science*. For a merchant's trade log the tags will be: *Personal Records* and *Business*, etc.

3.5. Technical metadata

Finally, to trace the corpus updating each text is to be provided with the following information:

- the name of the project member who added the text to the corpus;
- the date of the adding the text to the corpus.

4. Conclusion

The suggested metatext markup system will provide a user with the ability to create sub-corpora with the specified parameters. It will also help to control the process of filling the corpus with new text data and to estimate the balance of the corpus.

5. Acknowledgements

The Manding Corpus project is supported by Russian Foundation for Basic Research (RFBR) under Grant № 10-06-00219-a: Elaboration of the model of electronic corpus of texts in Manding languages (Maninka, Bamana).

6. References

Dumestre, G. (1993). *Bibliographie des ouvrages parus en bambara*. In: Mandenkan nº 26, pp. 67–90. [Online]. Available: http://llacan.vjf.cnrs.fr/PDF/Mandenkan26/26Dumestre.pdf (accessed March 2010)

Савчук С.О. (2005). Метатекстовая разметка в Национальном корпусе русского языка: базовые принципы и основные функции. In: Национальный корпус русского языка: 2003-2005. Результаты и перспективы. М., с. 62–88. [Savchuk, S.O. Metatext *Markup in the National Corpus of the Russian Language: Basic Principles and Functions.* In The National Corpus of the Russian Language: 2003-2005. Results and perspectives. Moscow, pp. 62–88.] [Online]. Available: http://ruscorpora.ru/sbornik2005/05savchuk.pdf

Sinclair, J. (2005). *Corpus and Text - Basic Principles.* In Developing Linguistic Corpora: a Guide to Good Practice, ed. M. Wynne. Oxford: Oxbow Books: 1–16. [Online]. Available: http://ahds.ac.uk/linguistic-corpora/

Sinclair, J. (1996). *Preliminary recommendations on text typology*. EAGLES Document EAG-TCWG-TTYP/P. [Online]. Available: http://www.ilc.cnr.it/EAGLES96/texttyp/texttyp.html

Vydrine, V. (2008). Glossed electronic corpora of Mande languages : A perspective that we cannot avoid. In *Mande languages and linguistics. 2nd International Conference: Abstracts and Papers*. St. Petersburg, pp. 16–23.

Towards a Comprehensive, Machine-readable Dialectal Dictionary of Igbo

Wanjiku Ng'ang'a

School of Computing and Informatics
University of Nairobi
Kenya
wanjiku.nganga@uonbi.ac.ke

Abstract

Availability of electronic resources, textual or otherwise, is a first step towards language technology research and development. This paper describes the acquisition and processing of a multi-dialectal speech and text corpora for the Igbo language. The compiled corpus provides the key resource for the definition of a machine-readable dialectal dictionary for Igbo. The work centres around an online portal that facilitates collaborative acquisition, definition and editing of the dialectal dictionary. The complete dictionary, which includes features such as phonetic pronounciation, syllabification, synthesized pronounciation as well as GIS locations, is then made available via a website, which will be accessible publicly once sufficient entries have been defined.

1. Background

The Igbo language is one of Africa's great indigenous languages, with over 30 million speakers around the world. Igbo belongs to the Benue-Congo branch of the Niger-Congo language family of Africa and is spoken in seven states in Nigeria: Abia, Anambra, Delta, Ebonyi, Enugu, Imo and Rivers, as well as among large and growing émigré populations in the United States and Europe. Linguists recognize more than 5 major dialect "clusters" and more than 15 main Igbo dialects in existence, but studies in Igbo dialectology are ongoing and the final number is likely to be higher (Achebe and Ng'ang'a, 2007).

The history of dictionary-making in Igbo may be said to have begun in the 18th century with the production of bilingual wordlists and glossaries by European missionaries. In 1777 for instance, the Moravian mission agent G. C. A. Oldendorp published "Geschichte der Mission der evangelischen Brüder…", which contained a number of Igbo words and numerals. The production of bilingual wordlists and vocabularies by European explorers and missionaries in Africa continued in the first half of the 19th century and were published either in separate volumes or included as appendices or glossaries at the end of grammar books. Samuel Crowther, a native agent of the Church Missionary Society (CMS), produced the Isoama-Ibo primer in 1857 (Crowther, 1882); and a number of translations of the Gospel into Igbo followed Crowther's primer during the second half of the 19th century (Norris, 1841; Schon, 1861; Koelle, 1854). The documentation of Igbo vocabulary at this time owed a great deal to the pragmatic concerns of the compilers, who were invariably Christian missionaries actuated by evangelical zeal; and their works lacked semantic coverage and basic phonological, morphological and syntactic information (Emenanjo and Oweleke, 2007). These pioneers were more concerned with translating Western religious texts into Igbo, than in documenting Igbo as it was spoken.

With the 20th century came the publication of dictionaries (Ganot, 1904), as well as a number of scriptural texts written in what was then called 'Union Ibo', an expedient and ultimately unworkable hybrid which was invented by the CMS under the British colonial administration for the evangelical mission. In 1913 for instance, the Union Ibo Bible was published, produced by Archdeacon T. J. Dennis. Since then, several other dictionaries and word lists have been compiled (Williamson, 1972; Echeruo, 1998; Igwe, 1999). These wordlists and dictionaries were produced single-handedly, by individual scholars or enthusiasts, without collaboration among lexicographers, linguists and others. Consequently, one of the major inadequacies of these dictionaries for humanities studies has been that, like all previous efforts at documenting the language, these dictionaries are limited in the scope and coverage of the vocabulary of Igbo language (Emenanjo and Oweleke, 2007).

1.1. The Igbo Archival Dictionary Project

Against this backdrop, there was an urgent need for the collaboration of lexicographers, linguists, interdisciplinary scholars, language planners and other stakeholders to bring the development of Igbo lexicography in line with international standards. In 1999 the Nigerian writer, Chinua Achebe, delivered a lecture in which he called on linguists and other scholars to check the declining fortunes of the Igbo language by working towards its comprehensive documentation. In his speech, he noted that "*Language is not a piece of iron that the blacksmith takes and puts into the fire, takes out and knocks into shape or moulds as it pleases him. Language is sacred, it is mysterious, it is fearsome. It is living and breathing. It is what separates humans from animals. It separates people from their counterparts, towns from their neighbors*", emphasizing the need to develop a comprehensive dictionary of Igbo and all its dialects.

Following from this, Ike Achebe founded the Igbo Archival Dictionary Project (IADP) in 2001 as an association for leading linguists, anthropologists, historians and other scholars on Igbo language and culture with the purpose of conducting research and salvage work to preserve and develop the Igbo language. The task of creating a comprehensive dictionary of the Igbo language is a large one, because it entails an inquiry not into a single dialect, but into the

entire complex of Igbo dialects. In 2002, IADP began the first significant effort in the history of the Igbo language, for its scientific documentation on a large scale. Over the subsequent years, IADP has been working at seven Nigerian Universities laying the ground work for a massive effort at Igbo documentation via the recording of spoken Igbo in all its dialects. IADP has trained and deployed more than 50 fieldworkers, linguists and consultants to work in Nigeria on the project. The project has to-date conducted fieldwork and recorded more than 1000 hours of local language speech in Igbo villages and towns, making it one of the largest salvage projects ever undertaken for documenting an African language from speech. Since 2002 the IADP has been building its Igbo Corpus from transcriptions of the audio recordings in its archives, with an objective of building the Igbo Corpus from local language speech to at least 25 million words by the end of 2012 (Achebe and Ng'ang'a, 2007).

1.2. The Igbo Online Resources Project

The Igbo Online Resources Project (IORP) is a sub-project within the greater IADP whose aim is to develop electronic resources for Igbo. The ultimate goal of the IORP is to develop language technology resources and tools that will facilitate the development of end-user applications for Igbo such as Word processors, Spell-checkers, Speech-mediated interactive voice response systems, and Machine translation of Igbo texts to name a few. To achieve this, we have embarked on the creation and development of varied electronic resources which include textual and audio corpora, a comprehensive machine-readable dictionary, speech synthesis software and software tools to perform miscellaneous linguistic tasks such as tone-marking and word syllabification. It is envisaged that as this repertoire of language technology tools and resources continues to grow, the IORP will be on course to achieve her overall objectives, as stated earlier. This paper describes the methodology and computational efforts employed by the IORP in delivering a web-based Igbo dictionary.

2. Creating the Dialectal Dictionary

IORP proceeds from a fundamental lexicographical principle: that speech occupies a much greater role in language-use than writing; and that for African languages, with relatively short literary histories, the significance of the spoken word and the oral tradition as repositories of vocabulary, greatly outweighs the value of printed texts for lexicography (Achebe and Ng'ang'a, 2007). Citation files have traditionally been based on writing; but it is now clear to lexicographers that the traditional citation file cannot adequately represent the various uses of language (Landau, 2001), especially for a language such as Igbo that has a predominantly non-literary past. This is because the vast proportion of the Igbo language is still undocumented, and Igbo continues to be a severely under-resourced language. One of the problems associated with lexicography for many African languages is the absence of a large body of print-texts created over a substantially long period of time from which items and their contextual meanings can be derived or deduced. Lexicography is virtually impossible without a linguistic corpus. For many European languages, for example, printed texts have been the basis for such a corpus. From such corpora, it was possible to make a census of a language's discrete word-forms. In addition, the history of the meanings and morphological status of such forms can be easily derived from such period-specific textual records. Scholars for languages like Igbo that do not have a long record of written texts, will have to base analysis of language-use on an actual and verifiable body of spoken evidence established and recorded in advance of specific citations[1] (Achebe and Ng'ang'a, 2007). The transcriptions of the audio recordings collected by the IADP form the key resource for dictionary definition for the IORP, as explained in subsequent sections. Delivering a web-accessible Igbo dictionary involves a 6-step process that manipulates/produces the following resources:

1. An Audio record of an interview is transcribed to produce
2. An XML-formatted transcript of the interview. This is then uploaded to a central server. Several such transcriptions form the
3. Electronic Igbo Corpus. Various pre-processing modules are applied to the corpus yielding
4. Unique word lists and concordances. These are used during the lexicographic work that produces
5. Unmoderated dictionary entries. Once these are moderated by the editorial committee, a list of
6. Moderated dictionary entries are then available for release via the online dictionary.

2.1. Creating the electronic Igbo Corpus

To avail the audio corpus of the IADP for lexicographic work, the first task is that of creating electronic transcriptions of the audio recordings. The first challenge that was encountered was that of orthography since Igbo contains many diacritical marks for tone-marking of vowels and nasals (high, low and downstep) as well as special characters that are not directly accessible from a standard keyboard. To facilitate easy and accurate typing of fully tone-marked Igbo texts, we developed Igbo software-based keyboards that enable a transcriber to insert all Igbo characters (both lower and upper case) with requisite tone-marking diacritics, with a single keystroke. In addition, we have created the Igbo Corpus Builder (ICB) software

[1] Members of the Igbo Archival Dictionary Project continue to produce scholarship on aspects of the work they have done for the project. In 2003, Ozo-Mekuri Ndimele, IADP's coordinator for the Echie dialect in 2002, published "A Concise Grammar and Lexicon of Echie" (Aba: National Institute for Nigerian Languages). Similarly, E. N. Oweleke, a foundation member of the IADP, who has followed closely the development of Igbo lexicography in IADP submitted a PhD dissertation titled, "Some Issues in Lexicography and the Problems of Affixation and Verb-Noun Selectional Restrictions in the Igbo Dictionary": (Unpublished Ph.D. Dissertation, 2007), University of Port Harcourt, Nigeria.

which is a special editor for creating interview transcriptions encoded in XML. The ICB facilitates insertion of interview metadata which includes the interview date, location (State, Local government area, town, village), interviewee details (age, gender, occupation), dialect information and discourse/topic classification.

2.2. The Lexicographical Task

Once an interview transcription is created via the ICB, it is added to the electronic textual Igbo corpus. The next step is that of pre-processing the interview texts to facilitate lexicograhical work, for which we developed a suite of pre-processing software tools that perform a myriad of natural language processing tasks. Text pre-processing mainly involves obtaining an alphabetically sorted list of all unique words in the text. A concordance for each unique word is also created as this provides the context for meaning identification and verification. Using this information, the lexicographers are then able to define dictionary entries using the headwords identified from the texts. Lexicographical work on a multi-dialectal language such as Igbo presents challenges at the macro-structure level, and dictionary compilation needs to address the question of how to represent headwords; that is, whether dialect forms should be selected as headwords and how variant forms should be represented. For example, if we take the Igbo variants in Table 1 which are glossed as 'body' in English, the question is, should these be listed as variants or headwords or should the lexicographer decide to use one dialect form? (Emenanjo and Oweleke, 2007).

àshụ (in Enuanị)
àrụ (in Onịcha)
ẹhụ (in Ika)
ẹsụ (in Ụkwụani)
ẹshụ (in Nsụka)
àhụ (in Owere)

Table 1: Dialect Variants for body

In addressing this problem, we take the view that a dialect dictionary basically works with three core data types: form, sense (meaning) and location, since it aims to document and classify dialectal form variants that are used to talk about specific senses in specific locations. For many dialect dictionary projects, the choice of how to organize these core data types has largely been influenced by the publication medium - printed media are one-dimensional and linear, and therefore dialect dictionaries invariably present the data sequentially, according to some ordering principle. In practice, this means that the editors have to choose one of the core types as the most important organizing principle. However, it is clear that the choice of opting for one organization over the other is not based on fundamental differences in importance of one core data type over the others, but purely on practical reasons. Different uses of the data are better catered for by one or the other organization, but the nature of the data does not have an intrinsic hierarchy such as "sense over form" or "form over sense".

We have adopted a dictionary organization model that allows us to abandon the distinction between macro- and micro-structure, opting instead to reduce micro-structure to the relation between the three core data types, and to broaden macro-structure to a dynamic, use-driven classification that is based on a combination of the basic tripartite units (Achebe and Ng'ang'a, 2007). To achieve this flexible data model, we ensured that the relationships inherent in the data were separated from the core data itself, in the design of the database schema. This flexibility in working with the data, makes it possible to organize the data in several different ways, thereby allowing the user to choose the viewpoint most suitable to their needs. For example, if the user wants to know the form variation for the sense "body", s/he will choose a sense-based view which would show all six variants given in Table 1. If s/he wants to know what the sense distribution of the form "àkwá" is, s/he will want to have the form-based view on the data. And finally, if s/he wants to view dictionary data for a particular dialect (location), for example to make a local dictionary, s/he will want to have a location-based view on the data. By adopting a database design that allows for multiple views, the resulting dictionary can be used for many different purposes.

The Igbo lexicographical task is a highly collaborative one, engaging different teams with different specialities: field workers, stenographers, transcribers, lexicographers and the editorial board. While most of the teams are within physical reach in Nigeria, the team of lexicographers and the editorial board are variously located across the globe. The IORP therefore developed an online portal to facilitate communication, data sharing and collaboration. The portal provides web-accessible tools to send and receive data files, browse the electronic corpus and generate corpus statistics, view the word lists and associated concordances, define words for the dictionary and moderate entries for release in the online dictionary. The portal creates a seamless collaboration platform and includes a wiki site to enhance communication and discussion amongst all the researchers.

2.3. Online Igbo Dictionary

After a lexicographer has completed defining a headword, the word must go through a moderation process, which is done by an editorial board, via the portal. The moderation ensures the accuracy of the entire entry. Once a word has passed the moderation stage, it becomes available for release in the online dictionary. Given that the overall objective is to develop a comprehensive, dialectal dictionary, the online dictionary provides features which enable a user to not only find the meaning(s) of a given word, but to obtain further information on any dialectal variations that may be associated with the current search term. By considering the variability of written Igbo texts with respect to tone-marking, we anticipate different orthographical representations for search terms, where users may type words with or without tone-marking. The latter scenario may be occasioned by several factors such as lack of an appro-

priate keyboard that allows efficient tone-marking of Igbo words, lack of knowledge on the exact tone-pattern associated with the search word, or an intentionally-unspecified (discovery) search. Since Igbo words only bear meaning when tone-marked, we provide two ways to enable a user accomplish the task of looking up meanings successfully: i) an Igbo character palette[2] that contains all Igbo characters with their tone-marked varieties and ii) automatic tone-marking, accomplished by a software module which automatically generates all tone-marked possibilities for a given canonical (non tone-marked) search term. The results of auto-tonemarking are filtered to reflect only those variations that have been defined in the dictionary database. These variations are provided as a list, from which the user can select the intended tone-marked search term. The user is however at liberty to search for any word, tone-marked or canonical, to find out if it is exists in the dictionary. Where a word does not exist, the website includes a "Suggest a Word" interface where users can suggest words for inclusion into the dictionary. Suggested words will go through an editorial and moderation process before being included into the dictionary.
For each search term, a wide range of information is provided:

1. English gloss, with a detailed English description where necessary.
2. The search word with appropriate tone-marking. This is important for cases where the user searches for a canonical word, and they are then able to see how it is tone-marked for different meanings.
3. Word syllabification and associated syllabic tones, since the syllable is the tone-bearing unit in Igbo.
4. The phonetic spelling of the search word, using the International Phonetic Alphabet.
5. A list of dialects where a given tone-marked variant bears the same meaning.
6. A word's geographical distribution which is availed interactively via a geographical information system (GIS) of Igbo land. Here, the user can see the exact villages, towns and states where the word is spoken, and with what variances across meanings and dialects.
7. An audio pronounciation of the search word which is generated by an Igbo speech synthesis engine which has been developed by the IORP[3]. This feature greatly enhances a user's experience as they can hear the tonal richness of Igbo.

2.4. Summary Statistics and Availability

To date, the IADP has collected over 1000 hours of Igbo speech. Of these, approximately 100 hours have been manually transcribed. Electronic transcription using the ICB is currently on-going. The project portal is complete with all dictionary definition modules having been tested and commissioned for use. The IORP is now in the electronic corpus aggregation phase where the ICB formatted, XML encoded interviews are being uploaded onto a common server-based data repository. Once a sizeable corpus has been obtained, the lexicographical work will begin in earnest. Currently, only a few entries have been defined, moderated and availed via the online dictionary.

All the digital resources developed by the IORP will be made available online - the online Igbo-English dictionary, the speech synthesis system, our GIS-based dialect maps and the machine translation engine will be launched as a suite of applications from a single web portal. IORP's preference for Open Source ware in the development of the products will ensure easy access, usability and wide dissemination via the Internet.

3. Conclusion

This paper has described a methodology for creating a machine-readable dictionary from an audio corpus. It highlights the challenges facing lexicographical work for African languages which are characterized by limited electronic resources, if any. By recognizing that African languages are largely oral with little or no literary past, the work described here undertakes a massive task of creating an audio databank for one of Africa's biggest languages. By devising language technology and software tools for language processing, the IORP implements a workable, efficient and cyclic workflow that transforms the audio corpus into an electronic textual corpus, adopting corpus encoding standards. With the corpus and requisite concordance tools in place, the lexicograhical task for Igbo is greatly enhanced if not simplified. In addition, by collecting spoken Igbo from across Igbo land, we are in a position to define the first, truly comprehensive dialectal dictionary of Igbo.

One of the major achievements of the work undertaken so far has been in solving the orthographic challenges that have plagued the creation of Igbo corpora - that of creating a corpus of fully tone-marked texts. This has been achieved by way of the software keyboards and the unicode-based ICB editor. This is a significant accomplishment since non tone-marked corpora are only useful to native Igbo speakers who can decipher the intended meaning (Uchechukwu, 2004a; Uchechukwu, 2004b). In addition, the lack of tone-marking results in an explosion of ambiguity at different levels of grammatical analysis which greatly complicates any language technology efforts. We have also successfully developed language technology tools and a natural language processing pipeline that support the compilation of a corpus of fully tone-marked texts, processing these texts for different linguistic analyses, and displaying fully tone-marked text via a web-browser, bringing the Igbo language into the internet domain without any representation limitations. With these tools, the task of creating a comprehensive dictionary for Igbo is now a feasible reality.

[2]This palette is accessible on the dictionary page as a selectable component.

[3]IORP commissioned Dr. Chinyere Ohiri-Aniche to document Igbo phonetic units in 2007 as part of IORP's endeavour to build linguistic resources for Igbo.

4. Acknowledgements

This paper has benefitted from ongoing collaborative research and discussion within the Igbo Archival Dictionary Project. The author wishes to acknowledge particularly the contributions of Professor E. Nolue Emenanjo, the Chief Editor of IADP; Professor Clara Ikekeonwu, the Chairman of the Editorial Board of IADP; and participants at the recent IADP Coordinator's Workshop held in Awka on March 4-7, 2010, for insights on Igbo dialectology. Late Professor Adiele Afigbo, National Coordinator of IADP from 2006-2009, helped develop the methodological approach to fieldwork discussed in this paper. Professor M. J. C. Echeruo coordinated the initial workshops at which issues of IADP's macro-structure were first raised and agreed upon. We are grateful for discussions with Professor M.C. Onukawa, Professor Ozo Mekuri Ndimele, Professor Ezikeojiaku, Dr. E. Oweleke, Dr. Ceclia Eme, Dr. Chris Agbedo, Dr. B. Mmadike, Jumbo Ugoji and George Iloene, which have enriched this paper.

The work accomplished by the IADP and IORP has been supported by the World Bank, the United Nations Foundation, The Ford Foundation and the Fulbright Program. We are grateful to the Vice-Chancellors of Nnamdi Azikiwe University Awka, Imo State University Owerri, Abia State University Uturu, Ebonyi State University Abakiliki, University of Port Harcourt, and University of Nigeria Nsukka, who have actively supported the goals of the IADP. We also acknowledge the technical support of teams situated at Bard College, USA and University of Nairobi, Kenya.

5. References

Achebe, I. & Ng'ang'a, W. (2007). The making of the Igbo online resources project. Technical report, IADP, New York.

Crowther, S.A. (1882). *Vocabulary of Ibo Language*. London, UK: Society for Promoting Christian Knowledge.

Echeruo, M.J.C. (1998). *Igbo-English Dictionary, with an English-Igbo Index*. London, UK: Yale University Press.

Emenanjo, E.N. & Oweleke, E.N. (2007). Compilation of the Igbo archival dictionary. Technical report, IADP, Port Harcourt.

Ganot, A. (1904). *English-Igbo-French Dictionary*. Rome, Italy: Sodality of St. Peter Claver.

Igwe, G.E. (1999). *Igbo-English Dictionary*. Ibadan, Nigeria: University Press Plc.

Koelle, S.W. (1854). *Polyglotta Africana or a Comparative Vocabulary of Nearly Three Hundred Words and Phrases in More Than One Hundred Distinct African Languages*. London UK: Church Missionary House.

Landau, S.I. (2001). *Dictionaries: The Art and Craft of Lexicography*. Cambridge, UK: Cambridge University Press.

Norris, E. (1841). *Outline of a Vocabulary of a Few of the Principal Languages of Western and Central Africa, Compiled for the Use of the Niger Expedition*. London, UK.

Schon, J.F. (1861). *Grammatical Elements of the Ibo Language*. London, UK.

Uchechukwu, C. (2004a). The Igbo corpus model. In *Proceedings of the 11th EURALEX International Congress*. Lorient, France: Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud.

Uchechukwu, C. (2004b). The representation of Igbo with the appropriate keyboard. In *International Workshop on Igbo Meta-Language*. Nsukka, Nigeria: University of Nigeria.

Williamson, K. (1972). *Igbo-English Dictionary*. Benin City, Nigeria: Ethiope Publishing.

Corpus Building in a Predominantly Oral Culture: Notes on the Development of a Multiple-Genre Tagged Corpus of Dagbani

Tristan Michael Purvis
University of Maryland, Box 25, College Park, MD 20742
tpurvis@umd.edu

Abstract

This paper describes the development of a multiple-genre tagged corpus of texts and recordings in Dagbani, a language spoken by a newly literate speech community in an African context with unique oral traditions. The paper begins with a brief review of the target speech community and corpus collection, followed by a discussion of the problems encountered and procedures conducted for tagging the corpus, and concludes with a presentation of example applications of and future plans for this corpus.

1. Introduction

Among the many challenges of developing African language technologies is the fact that language technology development highly favors the use of written resources both in terms of data management and specific applications, whereas a majority of African cultures place a higher value on oral traditions. This paper describes the development of a multiple-genre tagged corpus of texts and recordings for a language (Dagbani) spoken by a predominantly oral, yet also newly literate, speech community in an African context.

2. Background on Dagbani Language and Corpus Collection

Dagbani is a Gur (Niger-Congo, Atlantic-Congo) language spoken in northern Ghana by approximately 800,000 people (Lewis, 2009). As one of 12 nationally sponsored languages, the range of genres available for corpus collection is reinforced by government support for radio and television broadcasts, substantial funding and coordination of mother-tongue education in public schools and adult literacy programs, and government and NGO assistance for printing publications of various sorts. An attempt was made to balance the corpus with roughly equivalent written and spoken counterparts, as illustrated in Table 1. For example, the genre of oral history can be compared directly with a written counterpart; traditional (oral) fable-telling can be compared with written stories; scripted dialogue and personal letters to some extent are paired with spoken conversation; etc.

The Dagbani corpus, originally compiled and processed for use in a multi-dimensional analysis of register variation in Dagbani (Purvis, 2008; see Biber, 1988, 1995 on MD analysis methodology in general), comprises 163 texts (or text excerpts) and transcripts of recordings averaging 880 words each (roughly 140,000 words total) and representing approximately 32 genres. In addition to the primary corpus of naturally occurring data, there is a secondary corpus of 183 experimentally collected texts and recordings averaging 200 words (37,500 words total), made up of pairs of equivalent written and spoken narratives by the same speakers on the same topic.

Written	Spoken
Short stories, fables	Salima (fable-telling)
Novel excerpts	Oral interview with novelist
News articles	Oral recounting of articles
Written history	Oral history
Poetry	Oral intro/preface; Interview
Introduction	with novelist (monologue)
Personal letters	Natural conversation
Scripted conversation	Radio call-in discussion
Written plays	Broadcast talk show
	Improvisational drama
(see News article above)	*Radio & TV News
	*Broadcast announcement
Educ.materials (grammar)	College lecture
Legal texts	Courtroom proceedings
Medical pamphlets	Technical medical lecture
Agricultural pamphlets	Technical agric. discussion
Educational stories	Political speeches
Qur'an translations	Muslim sermons
Bible translations	Christian sermons
Written sermon etc.	*Christian prayers
*oral delivery, likely read from script	

Table 1. Genres collected for the Dagbani corpus

3. Tagging

Unlike corpus development for better resourced languages, which typically starts with a substantial base of electronically available texts and transcripts, the initial processing of texts and recordings for this corpus was almost entirely manual. First of all, there are hardly any online public texts in Dagbani apart from a translation of the Universal Declaration of Human Rights and selected translations of the Qur'an; so all texts for this research would have to be entered manually.

When working collaboratively with locals in a remote, semi-rural area in developing nation, simple technological assumptions such as Unicode font compatibility and keyboard set-up cannot be taken for granted. Therefore, to avoid problems in data processing and file transfer, a simplified transcription system was adopted based on a somewhat iconic use of Arabic numerals to represent the five non-Latin symbols used in Dagbani orthography: 3 = ɛ; 0 = ɔ; 6 = ɣ; 4 = ŋ; 7 = ʒ.

The option of scanning some of the written texts was ruled out due to poor print quality and frequent typographical errors or nonconventional orthography found in many Dagbani documents. A widespread number of homonyms/homographs in the Dagbani language would have posed another challenge for the automated tagging of the Dagbani texts, especially in combination with these other constraints. In speech, the number of homophones is reduced by tonal and vowel qualities, but notation of intonation and lexical tone are not represented in Dagbani orthography nor were these phonological distinctions within the scope of the research for which this corpus was first compiled. Unlike English, for example, where homonymy is rarely found among functional categories (e.g., there is only one *and*, only one *the*, only one *we*),[1] the homonymy in Dagbani does involve all sorts of major functional linguistic categories, which are often essential anchoring points in tagging programs or algorithms. Table 2 provides two different exemplary sets of such homographs and/or polysemes.

(a) *di*		(b) *ni*	
Grammatical Category	**Pre-tag Code**	**Grammatical Category**	**Pre-tag Code**
verb 'eat'	di	COMP; QUOtative	ni ni:
3S/INAM Subj pro; 3S/INAM Poss pro; and 3S/INAM Obj. as proclitic (rare)	di+ di$ di@	SUBordinator, used in relative clause constructions & temporal clause	ni*1 ni*2 ni-as
NEG. IMPERative	di*	FUTure marker	ni%
preverbal particle (time depth marker)	di%	NP conjunction; instrumental use	ni& ni&-w
		LOCative	ni#

Table 2: Examples of Dagbani homography/homophony

For example, the sequence *di di di zaa* could be read as 'don't eat it all' (NEG/IMPER eat it all) or 'it eats it all' (3S/INAM eat it all). As another example, what is typed as 'bɛ pan bo (ma)' in one written text and heard as [bɛ pa m bɔ (o)] in a separate recording ends up being transcribed in standard orthography as *bɛ pa ni bɔ* ('they now will look for . . .', 3P now FUT seek). 'bɛ pan bo' is meaningless, there being no word transcribed as 'pan' in Dagbani orthography. The phonetic representation [bɛ pa m bɔ] could have been interpreted literally (but incorrectly in this case) as 'many of them look for' (3P many EMPH seek). The correct transcription *bɛ pa ni bɔ* contains homographs that could lead to misinterpretation.

Consequently, a system of notation (exemplified in the "Pre-tag Code" column of Table 2) was devised in which each word was coded upon initial data entry for part of speech and other linguistic information relevant to a given word class (noun class, verb type, inflectional categories, syntactic role, etc.). In many cases the pretag symbols were either iconic or productively used with multiple lexical items; so the system could be learned easily and transcription could be completed rapidly. For example, all pronouns are tagged upon initial data entry for syntactic role of subject (+), object (@), or possessive ($). Or, compared to the example of ni& shown in Table 2 for nominal conjunction, the ampersand is also used to distinguish VP conjunction (ka&) and IP conjunction (ka&&) from homophonous forms such as the complement FRONting/focus particle ka! where we find the same exclamation point as used to distinguish the subject focus/EMPHasis particle n! from homophonous forms like the first person singular proclitic (n+, n$).[2]

Thus, the majority of features to be tagged for potential linguistic analysis were previously coded in each lexical item or somewhere in the text or transcript and only needed to be converted to a more formalized tagging system. Various discourse categories such as addressive and left-dislocated topic were also notated immediately as the text was entered into electronic form.

A sample of the formal tagging system is presented in Table 3. The first column of the tag indicates the basic part-of-speech category (**N**oun, **V**erb, ad**J**ective, adve**R**b, **D**eterminer, pronoun (**1-5**, **W**h), **C**onjunction, **P**article, or interjection/e**X**clamation). The remaining columns account for morphological, syntactic, or lexical information specific to a given part of speech, such as valence and aspect for verbs.

Col 1	Col 2	Col 3	Col 4	Col 5	Col 6
Noun	**S**ingular **P**lural #:denom **D**ummy	**A**nim. **I**nanim. **P**roper ***Bu*** ***Gu*** ***Lana*** **R**edup. **Q**uantity	po$s. **L**oc. **D**ir.	**C**ompound Lexicalized **K**ompound **V**erbal compound	**E**nglish **A**rabic **H**ausa **T**wi **M**ampruli **G**a
Verb	**T**rans. **I**ntrans. **D**itrans. **M**odal **S**entential **C**aus. **A**djectiv. **P**luraction #:deverbal e**Q**uative e**X**istential	**P**erf. **I**mperf. i**M**per. **R**edup. -***Gi*** form lexical **S**erial	**L**oc. **D**ir. **N**eg.	**1**st in serial **2**nd in serial **3**rd in serial ... **A**-1st in n-serial **B**-2nd in n-serial **C**-3rd in n-serial ...	**E**nglish **A**rabic **H**ausa **T**wi

[1] Cases do exist, e.g. *to* (preposition) versus *to* (infinitive), but nowhere near as much as in Dagbani, as illustrated in Table 2.

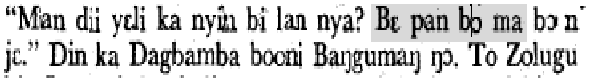
[2] The use of symbols such as +, @, $ of course requires adding escape characters (or other work-around strategies) in later stages of data processing, but this is outweighed by the rapid notation of linguistic information during data entry.

Col 1	Col 2	Col 3	Col 4	Col 5
Part-icle	**V**erbal	**S**entent. (*yi*,*ni*,*ti*)	**T**i-sequence marker	**P**urposive
			Ni	**E**vent
				1-herc
		Infinitive		**2**-hirc
		Negative		**3**-as/when
			Conditional	**C**ondition
				Result
		Time		
		e**M**phatic		
		Future	fut w/**N**eg	e**M**phatic
		adve**R**bial	**M**anner	
			Time	
			Purpose	**I**nf w/purp
	su**F**fix	**I**nfostatus	**I**mperfective	
			Focus	
		Nominal	**P**lural	
			Monger	

Table 3: Sample of formal tagging system

A few categories apply to multiple parts of speech - such as reduplication found with nouns, verbs, adjectives, adverbs, and interjections or language origin to document lexical borrowing or code-switching with nouns, verbs, adjectives, adverbs, determiners, conjunctions, and interjections. Conventional gloss categories and/or lemma gloss are combined with this tagging system. Perl script was used to convert pretagged texts and transcripts to fully tagged texts en masse.

As an illustration, the sentence presented earlier, *bɛ pa ni bɔ*, first encountered with nonconventional orthography as shown in Figure 1, would be entered as `b3+ pa* ni% b0VT` following the pretag data entry stage and later converted to `b3<3PSAW-3P> pa<PVRT-now> ni<PVF-FUT> b0<VT-seek>`.



Pretag: `b3+ pa* ni% b0VT`
Tagged: `b3<3PSAW-3P> pa<PVRT-now> ni<PVF-FUT> b0<VT-seek>`
Orthography: *bɛ pa ni bɔ*
Interlinear: 3P now FUT seek
Phonetic: [bɛ pa m bɔ]
Translation: 'they now will look for . . .'

Figure 1: Sample of nonconventional orthography (source: Sulemana, 1970)

In this case, *bɛ* (3P) is indicated as a third person plural inanimate weak personal pronoun acting as syntactic subject[3]; the word *pa* ('now') is indicated as a member of the large class of preverbal clitic particles (`<PV`) in general and more specifically as providing temporal adverbial information; the word *ni* (FUT), also of the preverbal clitic particle class, is indicated as the (affirmative) future marker; and *bɔ* ('seek') is indicated as a transitive verb with no further aspectual inflection.

Apart from time efficiency, a primary concern against using manual tagging as opposed to automated tagging algorithms is that of consistency and reliability. While automated tagging procedures are prone to mistags, at least the inaccuracy of the algorithm is consistent. Biber (1988) claims that spot checks reveal his tagging algorithms to be 90 percent accurate or better. Considering the relatively low frequency of certain key features such as "that-complements" and "that-relatives" which were reported as exemplary instances of mistagging in Biber's corpus, however, the overall accuracy rating of 90 percent may be misleading. A hand-tagged corpus as in the present study arguably benefits from a higher level of accuracy, with the potential for inaccuracy mainly dependent upon data entry errors and misjudgments on the part of the data analyst or native-speaker research assistants. The accuracy of coding linguistic features for this corpus was verified through periodic retagging of selected text excerpts and confirming that these matched up to the original tags.

4. Corpus Applications

The Dagbani corpus has been used in a number of research projects for theoretical and descriptive linguistic purposes. Immediate plans are to prepare the data for more advanced computational applications. A variety of completed, ongoing, and projected applications are reviewed below.

4.1 Factor analysis of register variation

As noted in Section 2, this corpus was originally collected for use in a study of register variation of Dagbani following what is known as the multi-dimensional analysis protocol (Purvis, 2008). For this study, factor analysis was conducted based on normalized frequency counts of targeted linguistic features in order to identity clusters of co-occurring variables and interpret the underlying communicative function. A Perl script was again used to compute frequencies of the targeted features for each text or transcript en masse. Some features were tallied directly (e.g. `FUT` for a unique future marker); while others involved variable expressions such as `<[1-4]...S` for all strong (emphatic/disjunctive) personal pronouns. Automated algorithms using Regular Expressions were also employed to track linguistic features involving multiple lexical items—e.g., `<VS*[A-Z1-3|]> ka<CSS2_ COMP2-and_COMP>` for identifying instances where *ka* as opposed to the more standard *ni* is used as a complementizer for verbs taking sentential complements. Some targeted linguistic features required additional manipulations of the texts and/or tags. For example, tallying of token-type ratio (TTR) for the first 300 words of text involved first running a Perl script to extract the

[3] Pronominal allomorphs are actually distinguished by clitic position not syntactic role, but syntactic role was tagged in the interest of theoretical and descriptive analysis (Purvis, 2007).

first 300 words and then a systematic trimming of lexical data to a subset of the tag categories and/or gloss tags so that the TTR was based on word lemmas as opposed to variably inflected word forms. Similarly, the tags were manipulated to ensure that a tally of lexical borrowings (of nouns and verbs mostly) treated inflected forms such as plural and possessive nouns as instances of the same lexical target.

4.2 Corpus analysis of pronoun allomorphs

The Dagbani corpus has also been exploited to efficiently track patterns of the use of allomorphs of weak personal pronouns (Purvis, 2007). These were formerly described as varying according to syntactic role (see, e.g., Olawsky, 1999; Wilson, 1972). Based on evidence from this corpus, however, they are better understood as clitics whose variable forms depend on their position in relation to their lexical host.

4.3 Corpus analysis of left dislocation

In another study inspired by seemingly anomalous patterns in the factor output of the original study on register variation, the corpus has provided a means to track the patterns of occurrence of and analyze the motivations behind the use of left dislocation constructions in Dagbani (Purvis, 2009).

4.4 Corpus analysis of relative constructions

Presently, the tagging of relative pronouns and other lexical items used in variable relative clause constructions in Dagbani is being expanded to indicate a number of contextual factors deemed useful in analyzing variation in the appearance of these constructions (e.g., placement in relation to verb of the main clause; placement in relation to verb of the relative clause, etc.)—in order to address unanswered questions and track previously undocumented variants (cf. Wilson, 1963). This pursuit has been facilitated by the legacy tagging system in combination with linguistic concordancing software.

4.5 Current developments

In its current state, the Dagbani corpus already constitutes a useful tool for computational applications. For example, with word by word tagging of part of speech and various categories of inflectional and derivational morphology and syntactic role in some cases, this modest corpus may be used as training data for morphological and syntactic parsing and machine translation of novel Dagbani texts and possibly adapted to other Gur languages. (The Gur cluster has been identified as one of 15 core languages and language clusters which in combination reportedly provide communicative coverage for up to 85% of Africa's population (Prah, 2002).)

Currently, we are in the process of aligning the tagged transcriptions with digitized recordings for the spoken portion of the Dagbani corpus to be available for use in applications such as speech recognition. As for the written texts, we are investigating methods for associating the tagged transcriptions with OCR-processed versions of the original documents which may contribute towards efforts to train software to correctly recognize nonstandard orthography and accurately deal with the widespread homography found in this language.

5. Summary

This paper has described the development of a multiple-genre tagged corpus of texts and recordings for a language (Dagbani) spoken by a newly literate speech community with an otherwise predominantly oral culture. Following a review of the target speech community and corpus collection, we have discussed problems encountered in data collection and presented the procedures followed for tagging the corpus. The corpus has benefitted a number of descriptive and theoretical research projects to date and is undergoing further processing for future computational applications.

6. Acknowledgements

The bulk of corpus collection and processing was funded by a U.S. Department of Education Fulbright-Hays doctoral dissertation grant in 2003-2004 and could not have been completed without the help of the numerous collaborators in the Dagbani-speaking community.

7. References

Biber, D. (1988). *Variation Across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, D. (1995). *Dimensions of Register Variation*. Cambridge: Cambridge University Press.

Lewis, M.P. (Ed.). (2009). *Ethnologue: Languages of the World, 15th edn*. Dallas, Tex.: SIL International.

Olawsky, K. J. (1999). *Aspects of Dagbani Grammar*. München: Lincom Europa.

Prah, K. (2002). Language, Neo-colonialism, and the African Development Challenge. TRIcontinental, No. 150. Cape Town: CASAS.

Purvis, T.M. (2007). A Reanalysis of Nonemphatic Pronouns in Dagbani. In F. Hoyt, N. Seifert, A. Teodorescu, & J. White (Eds.), *Proceedings of the Texas Linguistics Society IX Conference: The Morphosyntax of Underrepresented Languages*. Stanford: CSLI Publications, pp. 239–263.

Purvis, T.M. (2008). A Linguistic and Discursive Analysis of Register Variation in Dagbani. Doctoral dissertation. Indiana University.

Purvis, T.M. (2009). Left Dislocation in Dagbani. Paper presented at the Sixth World Conference on African Linguistics (WOCAL6), Cologne, Germany.

Sulemana, T. (1970). *Naa Luro*. Accra: Bureau of Ghana Languages.

Wilson, W.A.A. (1963). Relative constructions in Dagbani. *Journal of West African Languages*, 2 (Part 2).139–144.

Wilson, W.A.A. (1972). *Dagbani: An Introductory Course*. Tamale, Ghana: GILLBT.

Author Index