Bootstrapping Language Description: The case of Mpiemo (Bantu A, Central African Republic)

Harald Hammarström* **Christina Thornell†** **Malin Petzell†** **Torbjörn Westerlund‡**

*Dept. of Computing Science
Chalmers University
412 96 Gothenburg
Sweden
harald2@cs.chalmers.se

†Dept. of African Languages
Gothenburg University
405 30 Gothenburg
Sweden
{christina.thornell,malin.petzell}@african.gu.se

‡ Dept. of Linguistics and Philology
Uppsala University
751 26 Uppsala
Sweden
torbjorn.westerlund@lingfil.uu.se

Abstract

Linguists have long been producing grammatical decriptions of yet undescribed languages. This is a time-consuming process, which has already adapted to improved technology for recording and storage. We present here a novel application of NLP techniques to bootstrap analysis of collected data and speed-up manual selection work. To be more precise, we argue that unsupervised induction of morphology and part-of-speech analysis from raw text data is mature enough to produce useful results. Experiments with Latent Semantic Analysis were less fruitful. We exemplify this on Mpiemo, a so-far essentially undescribed Bantu language of the Central African Republic, for which raw text data was available.

1. Introduction

Descriptive linguistics, i.e., producing a grammatical description of a language (often previously unstudied or little-studied), is essential for the understanding of the language diversity in the world, for linguistic theory, for the historical study of populations and, last but not least, for the speakers themselves (van der Voort, 2007). It is even more a priority given the current state of language endangerment (Brenzinger, 2007).

Describing a language typically consists of producing a grammar, a dictionary and a collection of texts. In this paper, we suggest that this process can benefit from technology in the sense that it can speed up the human tasks of analysis and organisation. In particular, we show that techniques from computational linguistics are now mature enough that *morphological analysis*, *part-of-speech analysis* and potentially *lexical semantic analysis* can be bootstrapped from raw text. As an example language, we use Mpiemo (Bantu A, Central African Republic), for which some raw text data was available.

We focus here on motivation and proof-of-concept, leaving the linguistic details to a specialist northwest Bantu audience, and the technical details to a computational linguistics audience.

2. Motivation and Related Work

In language documentation and language description, one is bang-up-to-date with technology for recording, storage, annotation, modularization and presentation (Gippert et al., 2006)[1]. But technology can be further used to bootstrap analysis and speed-up manual work. In particular, we suggest that some analysis and organizing can be automatically extracted from *raw text data*.

Typically, a researcher works on grammar, texts and dictionary incrementally. A text is gathered first, which is then analysed and vacuumed for dictionary entries. Usually, texts can be gathered by a wider range of people, including people not schooled in linguistic theory, and there are many cases, old and new, where large text collections exist but there is no written down grammar/dictionary for the same language.[2] In other words, large text collections already exist for various undescribed languages, and for many others, text collections can be gathered relatively cheaply. This motivates our approach of bootstrapping from text.

There are also other, perfectly legitimate, ways to adapt grammar writing to enable technological exploitation. Nordhoff (2007a), Nordhoff (2007b) describes the grammar authoring system GALOES where the researcher writes the data in a format which allows harvesting, i.e., a computational tool can automatically select and collect data from grammars written in this way. Considerable flexibility in presentation, i.e., away from the strictly linear format of book grammars, also come with this grammar authoring system. Similarly, Beermann Hellan (2007) describes TypeCraft which is a support tool for glossing and annotation which helps researchers with consistency and sharing.

[1] Cf. the issues of Language Archives News `http://www.mpi.nl/LAN/`

[2] Three examples from three continents are Alsea (isolate; North America) has a text collection from 1920 (Frachtenberg, 1920), Uduk (Koman; Africa) has a New Testament translation from 1963 (Sudan Interior Mission, 1963) and Tabo (isolate; Oceania) has a New Testament translation from 2006 (Schlatter and Schlatter, 2006).

This enables more systematic searching and harvesting as well. These approaches are complimentary to the one suggested in this paper because the analysis itself is still fully the researchers burden, and use of the tools require linguistic training as well as computer familiarity.
Similar, unsupervised, techniques as we describe in this paper exist for further applications such as Information Retrieval, Spell-Checking etc. which are on the want list for low-density languages (Saxena and Borin, 2006), but this is not the focus of the present paper. Unfortunately, we are not aware of any Speech Technology tools equally suitable for facilitating work on language description.

3. Mpiemo Profile and Data

Mpiemo is spoken predominantly in the southwest of the Central African Republic (CAR) and in neighbouring Cameroon and Congo (= République du Congo, or Congo-Brazzaville). There are approximately 24 000 speakers in the Central African Republic, about 5 000 in Cameroon and an unknown, but presumably small, number of people in Congo (Gordon, 2005).
In the Central African Republic, almost all speakers are bilingual in Sango (the lingua franca of CAR), and knowledge of (varieties of) Gbaya, French, Lingala is also common. Mpiemo is losing ground but is still being transmitted to children. At present it is not an endangered language. Traditionally Mpiemo is not written but an orthography has been developed recently by missionaries (Thornell, 2004a). Mpiemo is placed in the Bantu A.80 (or 'Maka-Njem') group, but there is no detailed understanding of its proper classification (Maho, 2003).
The is no published grammatical description of Mpiemo but a text collection is scheduled to appear shortly (Thornell, 2008). There are also some papers on special topics (Thornell and Nagano-Madsen, 2004; Thornell, 2003; Thornell, 2004b) as well as some unpublished papers by SIL members in Cameroon. While the full morphosyntax of Mpiemo has yet to be described, some typological features are apparent. Like (almost) all Bantu languages, Mpiemo has a noun class system with alternating singular/plural prefixes. However, unlike Southern and Eastern Bantu, Mpiemo and other northwest Bantu languages tend not to have elaborate verb morphology. The language has SVO basic constituent order and has tones, but the tonal distinctions appear to have a low functional load.
At our disposal we had raw text data amounting to approximately 60 000 running words collected (1999-2008) by Christina Thornell in the Nola district of the Central African Republic. The texts are narrative descriptions of daily activities and local flora/fauna. We made use of all text data available. An example snippet is shown in Table 1.

4. Bootstrapping Experiments

4.1. Morphological Induction

As mentioned above, Mpiemo appears to have very little morphology. However, it is quite clear that there is a typical Bantu noun class system with alternating singular/plural prefixes, i.e., all nouns have two forms, one with a prefix to yielding singular meaning and one prefix yielding plural meaning. The Bantu descriptive tradition calls each prefix a 'class' and each class has a number. The goal is that classes which are cognate across Bantu languages should have the same class number in different languages (Maho, 2003). Our task is thus to unravel the Mpiemo specifics and relate them to the Bantu descriptive tradition.
Hammarström (2007) describes techniques for inducing concatenative morphology automatically, i.e., with no human intervention, from raw text data. In other words, if we input raw text data only, salient suffixes and prefixes can be extracted, and stems which take the same suffixes/prefixes systematically, can be listed. How this is done is explained elsewhere (Hammarström, 2007) including a full survey of work done on morphology induction.
The algorithm of Hammarström (2006a) was run on the approxiamtely 60 000 running words of Mpiemo text. The goal was to find the known prefixes correctly segmented and not to find any spurious prefixes or suffixes. As expected, the algorithm finds no salient suffixes for Mpiemo.[3] As for prefixes, the algorithm found the segmentations listed in Table 2. All of the segmentations turn out to be consistent with human analysis. (There is no point in a formal evaluation since the human analysis is not definitive, rather, the idea is to suggest segmentations that the researcher checks.)

Segmentation	Comment
a-	class prefix for 5
b-	class concord for 2
bi-	class prefix for 8
bo-	class prefix for 2
bì-	tonal allomorph for *bi-*?
bɛ-	class prefix for 2a?
bɛ̀-	allomorph for *bɛ-*
m-	concord for 6
mɛ-	class prefix for 6
mɛ̀-	tonal allomorph for *mɛ-*
y-	concord for 9 and others
yi-	concord for 9

Table 2: Outcome of affix extraction for Mpiemo.

Hammarström (2006b) is an unsupervised method to find stems which tend to appear with the same set of affixes, or, as one might call it, paradigm induction. Together with prefix extraction, we get a ranked list of <stem, prefix-set> pairs. The top pairs are shown in Table 3. The precision is excellent – fully conformant to human analysis – but recall is low. The paradigm of most stems cannot be inferred since they occur too sparsely, or, in other words, the corpus size is too small.
The value of these lists is that it speeds up the human analysis. Looking at the ranked lists, it is easy for a researcher to compare with other Bantu languages of the same region. The best described closely related language is Kol in

[3] There is actually at least one known suffix in Mpiemo, a plural imperative plural imperative suffix, but it does not occur in the (narrative) texts.

Bandi hɛ ri kɛ gwɔbi i ri bɛ dɛ gɔ: Hi nɔ mɛligi, hi kɛ bɛ sombi Mpanja, hi jɔ̀ɔ̀ pèà gɔ, ha nɛ Kamil hɔ́ ri kɛ́. Hí jɔ̀ɔ̀ pèà gɔ́, Kamil nɔ mɛlándi. Hí kɛ̀ jɔ̀ɔ̀ téri sómbi, a nɔ́ mɛ́lígí, à wá tí sómbi ya. Mɛ̀ rì yɛ́ nyɛ̀ mɛ̀kɔ̀gì. À láɳ mɛ́lígí má tí sombi yà gɔ́. Hi kwàn, hí sàà, hí kɛ́ bɛ́ mpàlà.
La pêche se passe comme ça: Nous prenons les filets, nous allons à la rivière Pandja, nous arrivons là-bas, Camille et moi, nous partons. En arrivant là-bas, Camille prend la pirogue. A peine arrivons-nous au beau milieu de la rivière, il prend les filets, les met à la rivière. Je lui passe des pierres. Il tend les filets dans la rivière.

Table 1: Sample snippet of Mpiemo text.

Prefix-Set	Stem	Translation
bi-	sani	"thing"
∅-		
mo-	ri	"person"
bo-		
...		

Table 3: Top pairs in paradigm induction.

Cameroon (Henson, 2007). With stems neatly categorized for prefixes, it is straightforward to compare and to see that, e.g., *bi-* must be class 8. Similarly, all of the above prefixes can be readily identified as inherited Bantu classes or subclasses (Maho, 2003). There appears to be some tonal allomorphy associated with the noun class prefixes. The morphology induction algorithm has no access to semantics, so it can not suggest which prefixes are allomorphic to each other, but the listings are handy for forming testable hypotheses.
In any case, whether human or machine analysed morphology, all stems and paradigms need to be double checked with speakers.

4.2. Part-of-speech Induction

Even a cursory inspection of the text data shows that Mpiemo distinguishes nominal and verbal classes distributionally. In addition, there are a number of particles whose position is unclear. Our task is therefore to get some headway in the understanding of these particles.
We have surveyed part-of-speech induction techniques. In general, there is very little work that is both truly unsupervised and aimed at a wide range of languanges. Biemann (2006) describes a mostly unsupervised part-of-speech tagger. The algorithm determines the number of different part-of-speech tags automatically, but there are a number of parameters that need to be tweaked.
The results are complicated by a number of parameter variations which are set ad hoc according to our existent but imperfect knowledge of Mpiemo. The exact settings and iterations are of little interest in this case – the point is whether the unsupervised computational analysis, allowing for a reasonable number of iterations, was of any help for the researcher. The results are that nominal and verbal classes emerge, but there is more than one nominal class and more than one verbal class. Impressionistically, also many infrequent words seem to end up in the right company. This is important, because most words of a running text are infrequent, and a good first guess at their part-of-speech can save a lot of time in dictionary making. 'gɔ' which may be a focal particle, is given a class of its own. Pronouns and what appears to be a pre-verbal particle for future marking always end up in the same class.
The results are good enough for some provisional assignments, but the distributional nature of particles need further study.

4.3. Semantic Grouping

Latent Semantic Indexing (Sahlgren, 2006) is a popular technique that can be used to infer semantic distances between words from raw text data. The intuition is that words that appear in the same "context" tend to be similar in meaning, once frequency discrepancies are discounted for. (Frequent words appear in all contexts, but they are not semantically similar to "everything".) Sometimes a one-word windows is used as the context, sometimes the sentence, but most commonly the document is used as a context (the raw text data used comes already divided into documents in these cases). When latent semantic analysis is successfully applied to major European languages, the raw data sources are typically huge, with at least millions of word tokens.
The goal of experimenting with latent semantic analysis on Mpiemo was to find semantically related words, such as animates, and because many of the texts were about plants, perhaps a category of plant names. In order for LSA techniques to operate on the minuscule size of the corpus, we had little choise but to use the sentence as context (anything bigger would have made the data set tiny, and anything smaller would reduce the semantic analysis towards part-of-speech analysis, i.e., syntactically legal contexts). We then tried simply to cluster on the LSA similarity measure. The result was that 'question words' was the cluster deemed most semantically related, presumably because of the question marks in sentences containing them. Little more of value came out of the attempt, presumably because the text corpus was simply too small.

4.4. Discussion

Bootstrapping from text data for grammar/dictionary writing is parallel to Machine Translation in that it will not replace humans in the foreseeable future. Its purpose is instead to save time for the same humans. Even small time saves are valuable. We have indicated that bootstrapping is worthwhile if the text collection is of moderate size. There are also some positive side-effects of the attempts that were unforeseen:

- Transcription consistency checking (almost like spell-checking) came out naturally from the morphological listings.

- The automatically annotated texts, which would otherwise just have gathered dust after analysis, could easily be ported to other formats, for example TEI/XML to be used in a pedagogical tool which teaches grammar to linguistics students (Borin and Saxena, 2004).

NLP bootstrapping techniques can be seen as a generalization of a corpus concordancer. A concordancer highlights and selects raw data and presents it in a manner suitable for a human analysis. As we argue, the same can be done at least for morphological analysis and part-of-speech analysis.

The usefulness hinges on the existence of a large body of raw text data. For some languages, division of labour allows such data to be gathered relatively cheaply. For many other languages, text collections already exist and can be made use of.

5. Conclusion

We have shown that language technology can be used to save time in language description. For the particular language Mpiemo, the morphology is quite simple, and morphology induction works very well for it. The usefulness of part-of-speech induction is harder to assess, and we were not successful in exploiting techniques for latent semantic analysis. Some positive side effects that may arise from the applying NLP technology to languages which traditionally were not treated computationally, are consistency checking and usage of tagged corpora for teaching purposes.

6. Acknowledgements

Funding support for this study was granted by the Centre for Language Technology, Gothenburg in the small project titled *Language Technology for Languages of the Central African Republic*.

7. References

Dorothee Beermann Hellan. 2007. Development of linguistic documentation tools under the umbrella of nufu. Presentation at the Year of African Languages Symposium, April 2007, Gothenburg.

Chris Biemann. 2006. Unsupervised part-of-speech tagging employing efficient graph clustering. In *Proceedings of the COLING/ACL 06 Student Research Workshop*. The Association for Computer Linguistics.

Lars Borin and Anju Saxena. 2004. Grammar, incorporated. In Peter Juel Henriksson, editor, *CALL for the Nordic languages*, volume 30 of *Copenhagen Studies in Languages*, pages 125–146. Samfundslitteratur.

M. Brenzinger, editor. 2007. *Language Diversity Endangered*, volume 181 of *Trends in Linguistics: Studies and Monographs*. Mouton de Gruyter.

Leo Joachim Frachtenberg. 1920. *Alsea texts and myths*, volume 67 of *Bureau of American Ethnology Bulletin*. Smithsonian Institution, Washington, D.C.

Jost Gippert, Nikolaus P. Himmelmann, and Ulrike Mosel, editors. 2006. *Essentials of language documentation*, volume 178 of *Trends in linguistics: Studies and Monographs*. Mouton de Gruyter.

Raymond G. Gordon, Jr., editor. 2005. *Ethnologue: Languages of the World*. SIL International, Dallas, 15 edition.

Harald Hammarström. 2006a. A naive theory of morphology and an algorithm for extraction. In R. Wicentowski and G. Kondrak, editors, *SIGPHON 2006: Eighth Meeting of the Proceedings of the ACL Special Interest Group on Computational Phonology, 8 June 2006, New York City, USA*, pages 79–88. Association for Computational Linguistics. `http://www.cs.chalmers.se/~harald2/sigphon06.pdf`.

Harald Hammarström. 2006b. Poor man's stemming: Unsupervised recognition of same-stem words. In Hwee Tou Ng, Mun-Kew Leong, Min-Yen Kan, and Donghong Ji, editors, *Information Retrieval Technology: Proceedings of the Third Asia Information retrieval Symposium, AIRS 2006, Singapore, October 2006*, volume 4182 of *Lecture Notes in Computer Science*, pages 323–337. Springer-Verlag, Berlin.

Harald Hammarström. 2007. Unsupervised learning of morphology: Survey, model, algorithm and experiments. Thesis for the Degree of Licentiate of Engineering, Department of Computer Science and Engineering, Chalmers University, 91 pp.

Bonnie Henson. 2007. *The Phonology and Morphosyntax of Kol*. Ph.D. thesis, University of California at Berkeley.

Jouni Maho. 2003. A classification of the bantu languages: An update of guthrie's referential system. In Derek Nurse and Gérard Philippson, editors, *The Bantu Languages*, Routledge Language Family Series, pages 639–651. Routledge, London & New York.

Sebastian Nordhoff. 2007a. The grammar authoring system galoes. Presentation at the Wikifying Research Workshop, June 2007, Leipzig.

Sebastian Nordhoff. 2007b. Grammar writing in the electronic age. Presentation at the Conference of the Association of Linguistic Typology, September 2007, Paris.

Magnus Sahlgren. 2006. *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Ph.D. thesis, Stockholm University, Stockholm.

Anju Saxena and Lars Borin, editors. 2006. *Lesser-known languages of South Asia: status and policies, case studies and applications of information technology*, volume 175 of *Trends in linguistics: Studies and Monographs*. Mouton de Gruyter.

Tim Schlatter and Karen Schlatter. 2006. *[Tabo New Testament in Two Dialects]*. Bible Society of Papua New Guinea, Port Moresby.

Sudan Interior Mission. 1963. *Gwon this ki 'twam pa mo [Uduk New Testament]*. Sudan Interior Mission.

Christina Thornell and Yasuko Nagano-Madsen. 2004. Preliminaries to the phonetic structure of the bantu language mpiemo. *Africa & Asia: Göteborg working pa-*

pers on Asian and African languages and literatures, 4:163–180.

Christina Thornell. 2003. Data on the verb phrase in mpiemo. *Africa & Asia: Göteborg working papers on Asian and African languages and literatures*, 3:91–122.

Christina Thornell. 2004a. Minioritetsspråket mpiemos sociolingvistiska kontext. *Africa & Asia*, 5:167–191.

Christina Thornell. 2004b. Wild plant names in the mpiemo language. *Africa & Asia: Göteborg working papers on Asian and African languages and literatures*, 4:57–89.

Christina Thornell. 2008. *"Boulettes de graines de courge, pêche, hospitalit . . . ": Enregistrements transcrits et annotées pour une documentation du mpiemo (langue bantoue de la Réepublique Centrafricaine et du Cameroun)*, volume 25 of *Wortkunst und Dokumentartexte in afrikanischen Sprachen*. Rüdiger Köppe Verlag, Köln.

Hein van der Voort. 2007. Theoretical and social implications of language documentation and description on the eve of destruction in rondônia. Paper presented at Conference on Language Documentation and Linguistic Theory 7-8 December 2007, SOAS, London.