Det här verket har digitaliserats vid Göteborgs universitetsbibliotek och är fritt att använda. Alla tryckta texter är OCR-tolkade till maskinläsbar text. Det betyder att du kan söka och kopiera texten från dokumentet. Vissa äldre dokument med dåligt tryck kan vara svåra att OCR-tolka korrekt vilket medför att den OCR-tolkade texten kan innehålla fel och därför bör man visuellt jämföra med verkets bilder för att avgöra vad som är riktigt.

This work has been digitized at Gothenburg University Library and is free to use. All printed texts have been OCR-processed and converted to machine readable text. This means that you can search and copy text from the document. Some early printed books are hard to OCR-process correctly and the text may contain errors, so one should always visually compare it with the images to determine what is correct.

GÖTEBORGS UNIVERSITET
GÖTEBORGS UNIVERSITETSBIBLIOTEK

UNIVERSITY OF GOTHENBURG

Unsupervised Learning of Morphology and the Languages of the World

Harald Hammarström

Ph.D. thesis

Department of Computer Science and Engineering
Chalmers University of Technology & University of Gothenburg

Gothenburg, Sweden 2009

CHALMERS | UNIVERSITY OF GOTHENBURG

Unsupervised Learning of Morphology and the Languages of the World

HARALD HAMMARSTRÖM

Thesis for the Degree of Doctor of Engineering
to be defended in public at **10:15, December 11, 2009**
in room HB1, Hörsalsvägen 8,
Chalmers University of Technology (Campus Johanneberg), Gothenburg.

The defence will be held in English.
Faculty opponent: Professor Richard Sproat,
Oregon Health and Science University.

Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg
SE-412 96 Göteborg, Sweden
Telephone: +46 (0)31 772 1000

Unsupervised Learning of Morphology and the Languages of the World
HARALD HAMMARSTRÖM
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg
ISBN 978-91-628-7942-6

Abstract

This thesis presents work in two areas; Language Technology and Linguistic Typology.

In the field of Language Technology, a specific problem is addressed: *Can a computer extract a description of word conjugation in a natural language using only written text in the language?* The problem is often referred to as Unsupervised Learning of Morphology and has a variety of applications, including Machine Translation, Document Categorization and Information Retrieval. The problem is also relevant for linguistic theory. We give a comprehensive survey of work done so far on the problem and then describe a new approach to the problem as well as a number of applications. The idea is that concatenative affixation, i.e., how stems and affixes are stringed together to form words, can, with some success, be modelled simplistically. Essentially, words consist of high-frequency strings ("affixes") attached to low-frequency strings ("stems"), e.g., as in the English *play-ing*. Case studies show how this naive model can be used for stemming, language identification and bootstrapping language description.

There are around 7 000 languages in the world, exhibiting a bewildering structural diversity. Linguistic Typology is the subfield of linguistics that ams to understand this diversity. Many of the languages in the world today are spoken only by relatively small groups of people and are threatened by extinction and it is therefore a priority to record them. Language documentation, is and has been, an extremely decentralised activity, carried out not only by linguists, but also missionaries, travellers, anthropologists etc foremostly throughout the past 200 years. There is no central record of which and how many languages have been described. To meet the priority, we have attempted to list those languages which are the most poorly described which do not belong to a language family where some other languages is decently described – a task requiring both analysis and diligence. Next, the thesis includes typological work on one of the more tractable aspects of language structure, namely numeral systems, i.e., normed expressions used to denote exact quantities. In one of the first surveys to cover the whole world, we look at rare number bases among numeral systems. One major rarity is base-6-36 systems which are only attested in South/Southwest New Guinea and we make a special inquiry into its emergence.

Traditionally, linguists have had headaches over what counts as a language as opposed to a dialect, and have therefore been reluctant to give counts of the number of languages in a given area. One chapter of the present thesis shows that, contrary to popular belief, there is an intuitively sound way to count languages (as opposed to dialects). The only requirement is that, for each pair of varieties, we are told whether they are mutually intelligible or not.

Thesis for the Degree of Doctor of Engineering

Unsupervised Learning of Morphology and the Languages of the World

HARALD HAMMARSTRÖM

Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg
SE-412 96 Göteborg
Sweden

Gothenburg, December 2009

Unsupervised Learning of Morphology and the Languages of the World
HARALD HAMMARSTRÖM
ISBN 978-91-628-7942-6

© HARALD HAMMARSTRÖM, 2009

Technical report 64D
Department of Computer Science and Engineering
Language Technology Research Group

Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg
SE-412 96 Göteborg
Sweden
Telephone +46 (0)31 772 1000

Printed at Chalmers, Gothenburg, Sweden, 2009

Abstract

This thesis presents work in two areas; Language Technology and Linguistic Typology.

In the field of Language Technology, a specific problem is addressed: *Can a computer extract a description of word conjugation in a natural language using only written text in the language?* The problem is often referred to as Unsupervised Learning of Morphology and has a variety of applications, including Machine Translation, Document Categorization and Information Retrieval. The problem is also relevant for linguistic theory. We give a comprehensive survey of work done so far on the problem and then describe a new approach to the problem as well as a number of applications. The idea is that concatenative affixation, i.e., how stems and affixes are stringed together to form words, can, with some success, be modelled simplistically. Essentially, words consist of high-frequency strings ("affixes") attached to low-frequency strings ("stems"), e.g., as in the English *play-ing*. Case studies show how this naive model can be used for stemming, language identification and bootstrapping language description.

There are around 7 000 languages in the world, exhibiting a bewildering structural diversity. Linguistic Typology is the subfield of linguistics that aíms to understand this diversity. Many of the languages in the world today are spoken only by relatively small groups of people and are threatened by extinction and it is therefore a priority to record them. Language documentation, is and has been, an extremely decentralised activity, carried out not only by linguists, but also missionaries, travellers, anthropologists etc foremostly throughout the past 200 years. There is no central record of which and how many languages have been described. To meet the priority, we have attempted to list those languages which are the most poorly described which do not belong to a language family where some other languages is decently described – a task requiring both analysis and diligence. Next, the thesis includes typological work on one of the more tractable aspects of language structure, namely numeral systems, i.e., normed expressions used to denote exact quantities. In one of the first surveys to cover the whole world, we look at rare number bases among numeral systems. One major rarity is base-6-36 systems which are only attested in South/Southwest New Guinea and we make a special inquiry into its emergence.

Traditionally, linguists have had headaches over what counts as a language as opposed to a dialect, and have therefore been reluctant to give counts of the number of languages in a given area. One chapter of the present thesis shows that, contrary to popular belief, there is an intuitively sound way to count languages (as opposed to dialects). The only requirement is that, for each pair of varieties, we are told whether they are mutually intelligible or not.

Table of Contents

Acknowledgments

I would like to thank a number of people who, in various ways, were of importance for my writing this thesis. Of the people in the Computer Science Department I would especially like to thank my supervisor Bengt Nordström for his adamant support, erudite discussions and inspiring pathos for science. Likewise, I am much indebted to my second supervisor Aarne Ranta for his delicate advice and vibrant activity in Language Technology. However, thanks to Devdatt "Grälsjuk" Dubhashi I have some idea as to how the world *really* works. I hope to be able to forward his priceless gift.

Markus Forsberg, Björn Bringert, Håkan Burden, Alejandro Russo, David Wahlstedt, Jan-Willem Roorda, Krasimir Angelov, Libertad Tansini, Vilhelm Verendel, Wolfgang John, Merja Karjalainen and the other past and present PhD students at the CS(E) department have contributed to a great social and research environment.

Outside the department, I would also like to thank Jens Allwood (especially for bringing me to South Africa for some valuable experience), Lars Borin, Anju Saxena, John Löwenadler, Lilja Øvrelid, the Africanists Karsten Legère, Christina Thornell, Eva-Marie Ström, Malin Petzell, Helene Fatima Idris (especially for helping me and Therese while in Sudan), and a multitude of international colleagues with whom I have exchanged beers, laughs, ideas and materials over many conferences, visits and emails. Swintha Danielsen, Sophie Salffner, Roger Blench and Pushpak Bhattacharyya have even hosted me in times of travel. Similarly, I am lucky to be part of the GSLT network, crammed with too many intelligent and entertaining individuals to list.

Further back in time, I would like to thank my old classmates from Uppsala who were instrumental in getting me hooked on Computer Science in the first place: Tomas Fägerlind (who has also taught me everything I know about humans), Magnus Rattfeldt, Olof Dahlberg, Jim Wilenius, Lars "Ars" Göransson, Mattias Jakobsson, Per "upp över 100" Sahlin and the charming Jonas "Norris" Grönlund. Herman Geijer never cared much for Computer Science but has been a great friend over the years (in fact, I finished the thesis manuscript sitting in his couch). Of course, there would have been no thesis without Henrik Olofsson, my oldest friend, my mother and father, or without Therese Brolin, the world's dearest.

Harald Hammarström

Gothenburg
December 2009

Introduction

1 Language Technology

The work described in the first part of this thesis is in the area of Language Technology (LT), here defined as *the study of computer-aided processing of natural languages*. The ultimate goal of LT is to allow computers to deal with ("understand") natural language as humans do, which would make computers enormously more useful to humans. As of now, this goal is very far off, and we are happy if we can make progress on smaller subtasks, even if they do not achieve perfect accuracy. The problem studied in this thesis is one such subtask, and can be described as follows:

> Given a large collection of written text in a given natural language, can a computer, without any specific knowledge about the language, extract a description of how words are conjugated in that language?

The problem is often referred to as Unsupervised Learning of Morphology, but also (Automatic) Induction of Morphology, Morpheme Discovery, Word Segmentation, Algorithmic Morphology, quantitative Morphsegmentierung (in German) and other variants have been used. Of these, Unsupervised Learning of Morphology (ULM) is fairly common and faces the least risk of misunderstanding, so it will be used throughout the present work.

In the Computer Science tradition, the solution to task such as this amounts to a) providing a formal description of the problem (in terms of sets, strings, logical conditions and the like) into which real-world instances are approximated, b) providing a step-by-step description of a method, i.e., an algorithm, to compute the desired output from the input and c) a proof or argument for the correctness and (if known) the optimality of the algorithm. Remarkably, in the 1940s, long before the Computer Science had matured as a field, and long before computers became practical to use, so-called structural linguists were asking for a solution of the exactly the same kind to the ULM and related problems, but from a different perspective. The interest was not so much putting computers to work as to learn how linguistic analysis could be understood, which has particular implications for linguistic theory and possibly child language acquisition. As with most work in Language Technology, the present work will draw on experiences from both Computer Science and Linguistics, and hopefully contribute to all.

The ULM problem is stated above in rather abstract terms. One might ask for specifics in terms of which languages are targeted, what (implicit) knowledge is allowed, how high accuracy is the aim, if there are speed requirements, how

much text input is needed, what is meant by a description of conjugating words, is a black-box solution adequate or do we have to understand the inner workings, what is assumed about the written form of a language and so on. All these aspects with be elaborated on in the thesis. However, in essence, we target a much wider range of languages than English, but if the input language is the English New Testament[1] the desired output is any kind of description that tells us that forms like played and playing are conjugations of the same stem, and that see and sea aren't, perhaps reaching 90% accuracy on such pairs. No knowledge at all of forms is to be supplied but a small number of parameters and assumptions about suffix-length can be tolerated, whereas running time is not a priority.

Word-form analysis, or morphological analysis (see below), is generally the first step in computational analysis of natural language, and as such has a wide variety of LT applications, including Machine Translation, Document Categorization and Information Retrieval. ULM can also serve to boost investigations in Linguistics, especially the subfields Quantitative Linguistics and Linguistic Typology, and potentially contribute to linguistic theory.

A legitimate question is about the stipulation that distributional criteria alone should serve as the only source of knowledge for the computer. Why cannot a little or a lot of human knowledge about a language be hard-wired in order to describe how words are conjugated? This is indeed an option, and has been the way to handle the matter for virtually all languages committed to computational treatment, but it normally requires a lot of human effort. Roughly the amount of work of an MA thesis is needed to computationally implement conjugational patterns and an unspecified but huge amount of work to list legal lexical items.[2] Therefore, the ULM-problem as specified, has an important role to play. First, it would be a great benefit to rid us of the human effort of implementing conjugational patterns for the next range of languages to receive computational treatment. Second, even for languages which have this already, along with huge lists of lexical items, open domain texts will always contain a fair share of (inflected) previously unknown words, that are not in the lexicon (Forsberg et al. 2006, Lindén 2008, Mikheev 1997, Bharati et al. 2001). There has to be strategy for such out-of-dictionary words – a ULM-solving algorithm is one possibility. It could also turn out that the ULM-problem cannot, in some sense, be solved without explicit human-derived linguistic knowledge. If such a proof, or a convincing argument, is found this constitutes a resolution to the ULM-problem as good as one which proves the existence of an ULM-solving algorithm.

[1] 785066 tokens/running words versus 12999 unique words/types (King James 1977).

[2] Because of this, most such implementations have so far not been released to the public domain and have sometimes been kept in formats with poor portability, but there is in principle no reason why it should continue to be so, cf. Forsberg (2007).

2 Languages of the World

The work described in the second part of this thesis is in the area of Linguistics, here defined as *the study of natural languages.* More specifically, the work in this thesis falls in the subfield of Linguistic Typology, or *the systematic study of the unity and variation of the languages of the world.*

Among all the normed speech varieties occurring among the world's peoples, linguistics have long become accustomed to the concept of a language as a maximal set of mutually intelligible varieties. (As is well-known, the everyday usage of the word language, does not precisely correspond to this delineation, as other factors, such as attitudes or political power, play a role in forming the everyday status.) Empirically and theoretically, there are problems with the notion of mutual intelligibility and a strict yes/no property. However, if we assume for a moment that there is no problem with the notion of mutual intelligibility, that is, for each pair of varities, we can decide yes/no if they are intelligible. Then it is logically possible that A is mutually intelligible with B, B is mutually intelligible with C, but A is not mutually intelligible with C. The traditional manner in which linguists have approached this situation is to say that there is no way to assign languages over A, B, C, without somehow getting into contradictions, given the concept of language a maximal set of mutually intelligible varieties – A, B, C cannot all be the same language, as A and C are not mutually intelligible. If A, B is one language, then by the same token B, C should also be one language, but if A is the same as B and B is the same as C, then A and C must be the same, but they are not mutually intelligible! For this reason, linguistic have though the concept of language as being born with logical inconstiencies, and as a result, declared it impossible to count the number of languages in the world. This traditional view is too narrow, and to claim that there is no meaningful way to count the number of languages is wrong. In Chapter V, we give a novel intuitively sound interpretation to show that it is possible to count the number of languages without any inconsistencies in any arrangement of speech varieties, as long as we assume that each pair of varieties can be decided mutual intelligible or not.

In Linguistic Typology, cross-linguistic facts are noted and non-random discrepancies are sought to be explained. Many different kinds of explanations could a priori be invoked, psycholinguistic, historical, cultural etc. In Chapter VII we present a rigid definition and a thorough survey of facts on one aspect of human language, namely number bases in the numeral system. It is presumably the first such survey that is explicitly known to cover languages from every language family attested in the world and thereby we are able to set the record straight in a number of open cases. One major rarity is base-6-36 systems which are only attested in South/Southwest New Guinea. In Chapter VI we attempt to trace the emergence of the base-6-36 system in this area. Although the data is somewhat incomplete, there is evidence that the 6-36 system came from yams counting. A cultural explanation, as the neighbouring non-base-6 languages do not rely on tuber cultivation for subsistence.

Many of the languages in the world today are spoken only by relatively small

groups of people. Of these, many are on the path to extinction, in the sense that speakers, especially younger generations, are shifting to using another language, and consequently, as generations pass, no speakers at all will be left. Languages today die at a much faster rate than languages diverge to become new languages. Therefore the world's linguistic diversity is at risk of disappearing. For a scientific observer, the world's linguistic diversity is a unique gigantic experiment on human communication systems, which no laboratory can hope to achieve. For a small group of people, the language is part of their identity, and while a few are happy to shift, most groups would like to maintain their language, and, if anything, be bilingual in another, bigger, language. Languages documentation, i.e., to record languages (dictionary, grammar book, sound/video recordings), makes both scientists happy and helps the speaker community empower their language, and, if it dies anyway, allow descendants to see and hear their ancestral language.

Language documentation, is and has been, an extremely decentralised activity. It has been the outcome of linguists, missionaries, travellers, anthropologists, administrators etc stationed at missions, colonial establishments, universities in the first world and universities in the third world, over the past several centuries. There is no central record of which and how many languages have been described and to what level. From the perspective of science, the highest priority are languages otherwise poorly documented which are not genetically related to some other language which is not so poorly documented. In Chapter VIII, we list those languages. Making such a list involves considerable bookkeeping work and a vast amount of analysing unclear cases, judging extinctness, and gauging relatedness of partly described, dubiously attested language varieties.

3 Publications and Contributions

The chapters in this thesis are based on the following publications.

a. Hammarström, H. (2005). A New Algorithm for Unsupervised Induction of Concatenative Morphology In Yli-Jyrä, A., Karttunen, L., and Karhumäki, J., editors, *Finite State Methods in Natural Language Processing: 5th International Workshop, FSMNLP 2005, Helsinki, Finland, September 1-2, 2005. Revised Papers*, volume 4002 of *Lecture Notes in Computer Science*, pages 288–289. Springer-Verlag, Berlin.

b. Hammarström, H. (2006a). A naive theory of morphology and an algorithm for extraction. In Wicentowski, R. and Kondrak, G., 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.

c. Hammarström, H. (2006b). Poor man's stemming: Unsupervised recognition of same-stem words. In Ng, H. T., Leong, M.-K., Kan, M.-Y., and Ji, D., 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.

d. Hammarström, H. (2007a). A fine-grained model for language identification. In *Proceedings of iNEWS-07 Workshop at SIGIR 2007, 23-27 July 2007, Amsterdam*, pages 14–20. ACM.

e. Hammarström, H. (2007b). A survey and classification of methods for (mostly) unsupervised learning of morphology. In *NODALIDA 2007, the 16th Nordic Conference of Computational Linguistics, Tartu, Estonia, 25-26 May 2007*. NEALT.

f. Hammarström, H., Thornell, C., Petzell, M., and Westerlund, T. (2008). Bootstrapping language description: The case of Mpiemo (Bantu A, Central African Republic). In *Proceedings of LREC-2008*, pages 3350–3354. European Language Resources Association (ELRA).

g. Hammarström, H. (2009a). Poor man's word-segmentation: Unsupervised morphological analysis for indonesian. In *Proceedings of the Third International Workshop on Malay and Indonesian Language Engineering (MALINDO)*. Singapore: ACL.

h. Hammarström, H. (2009b). A Survey of Computational Morphological Resources for Low-Density Languages *Submitted.*

i. Forsberg, M., Hammarström, H., and Ranta, A. (2006). Lexicon extraction from raw text data. In Salakoski, T., Ginter, F., Pyysalo, S., and Pahikkala, T., editors, *Advances in Natural Language Processing: Proceedings of the 5th International Conference, FinTAL 2006 Turku, Finland, August 23-25, 2006*, volume 4139 of *Lecture Notes in Computer Science*, pages 488–499. Springer-Verlag, Berlin.

j. Hammarström, H. (2008a). Automatic annotation of bibliographical references with target language. In *Proceedings of MMIES-2: Wokshop on Multi-source, Multilingual Information Extraction and Summarization*, pages 57–64. ACL.

k. Hammarström, H. (2008b). Counting languages in dialect continua using the criterion of mutual intelligibility. *Journal of Quantitative Linguistics*, 15(1):34–45.

l. Hammarström, H. (2009c). Whence the Kanum base-6 numeral system? *Linguistic Typology*, 13(2):305–319.

m. Hammarström, H. (2009d [to appear]). Rarities in numeral systems. In Wohlgemuth, J. and Cysouw, M., editors, *Rara & Rarissima: Collecting and interpreting unusual characteristics of human languages*, Empirical Approaches to Language Typology, pages 7–55. Mouton de Gruyter.

n. Hammarström, H. (2009e). The Status of the Least Documented Language Families in the World *Submitted.*

All the work in the present thesis is the sole and original work of the author, except Chapter III and the last section of Chapter 8. In Chapter III, the present author conducted the experiment, took part in discussions, wrote the related work section and did the proof of NP-completeness, whereas the design, description and implementation of the extraction-tool was the work of Markus Forsberg and Aarne Ranta. In section III, the present author did the design, implementation and write-up of the experiment, whereas Christina Thornell collected the text data in the field in the Central African Republic and Torbjörn Westerlund as well as Malin Petzell offered feedback and took part in discussions.

References

Bharati, A., Rajeev Sangal, S. B., Kumar, P., and Aishwarya (2001). Unsupervised improvement of morphological analyzer for inflectionally rich languages. In *Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium (NLPRS-2001), November 27-30, 2001, Hitotsubashi Memorial Hall, National Center of Sciences, Tokyo, Japan*, pages 685–692. Tokyo, Japan.

Forsberg, M. (2007). *Three Tools for Language Processing: BNF Converter, Functional Morphology, and Extract.* PhD thesis, Chalmers University of Technology, Gothenburg.

Forsberg, M., Hammarström, H., and Ranta, A. (2006). Lexicon extraction from raw text data. In Salakoski, T., Ginter, F., Pyysalo, S., and Pahikkala, T., editors, *Advances in Natural Language Processing: Proceedings of the 5th International Conference, FinTAL 2006 Turku, Finland, August 23-25, 2006*, volume 4139 of *Lecture Notes in Computer Science*, pages 488–499. Springer-Verlag, Berlin.

King James (1977). *The Holy Bible, containing the Old and New Testaments and the Apocrypha in the authorized King James version.* Nashville, New York: Thomas Nelson.

Lindén, K. (2008). A probabilistic model for guessing base forms of new words by analogy. In Gelbukh, A. F., editor, *Proceedings of CICLing-2008: 9th International Conference on Intelligent Text Processing and Computational Linguistics*, volume 4919 of *Lecture Notes in Computer Science*, pages 106–116. Springer.

Mikheev, A. (1997). Automatic rule induction for unknown-word guessing. *Computational Linguistics*, 23(3):405–423.

Part One: Computational Linguistics

Chapter I Unsupervised Learning of Morphology: A Naive Model and Applications

Edited synthesis of the following papers, where Hammarström (2007b) has been substantially updated:

a. Hammarström, H. (2005). A New Algorithm for Unsupervised Induction of Concatenative Morphology In Yli-Jyrä, A., Karttunen, L., and Karhumäki, J., editors, *Finite State Methods in Natural Language Processing: 5th International Workshop, FSMNLP 2005, Helsinki, Finland, September 1-2, 2005. Revised Papers*, volume 4002 of *Lecture Notes in Computer Science*, pages 288-289. Springer-Verlag, Berlin.

b. Hammarström, H. (2006a). A naive theory of morphology and an algorithm for extraction. In Wicentowski, R. and Kondrak, G., 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.

c. Hammarström, H. (2006b). Poor man's stemming: Unsupervised recognition of same-stem words. In Ng, H. T., Leong, M.-K., Kan, M.-Y., and Ji, D., 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.

d. Hammarström, H. (2007a). A fine-grained model for language identification. In *Proceedings of iNEWS-07 Workshop at SIGIR 2007, 23-27 July 2007, Amsterdam*, pages 14–20. ACM.

e. Hammarström, H. (2007b). A survey and classification of methods for (mostly) unsupervised learning of morphology. In *NODALIDA 2007, the 16th Nordic Conference of Computational Linguistics, Tartu, Estonia, 25-26 May 2007*. NEALT.

f. Hammarström, H., Thornell, C., Petzell, M., and Westerlund, T. (2008). Bootstrapping language description: The case of Mpiemo (Bantu A, Central African Republic). In *Proceedings of LREC-2008*, pages 3350-3354. European Language Resources Association (ELRA).

g. Hammarström, H. (2009a). Poor man's word-segmentation: Unsupervised morphological analysis for indonesian. In *Proceedings of the Third International Workshop on Malay and Indonesian Language Engineering (MALINDO)*. Singapore: ACL.

Unsupervised Learning of Morphology: A Naive Model and Applications

Harald Hammarström
Department of Computer Science and Engineering
Chalmers University of Technology
and University of Gothenburg
SE-412 96 Göteborg, Sweden
harald2@chalmers.se

1 Introduction

The problem addressed in the present chapter can be described as follows:

Input: An unlabeled corpus of an arbitrary natural language

Output: A (possibly ranked) set of prefixes and suffixes corresponding to true prefixes and suffixes in the linguistic sense, i.e., well-segmented and with grammatical meaning, for the language in question.

Restrictions: We consider only concatenative morphology and assume that the corpus comes already segmented on the word level.

The problem, in practice and in theory, is relevant for information retrieval, child language acquisition, and the many facets of use of computational morphology in general.

The reasons for attacking this problem in an unsupervised manner include advantages in elegance, economy of time and money (no annotated resources required), and the fact that the same technology may be used on new languages.

We begin with a survey on ULM in general, i.e., the problem as above, but without the restrictions.

Next, we describe two components in the broader line of attack on the ULM-problem. The first component extracts a list o salient prefixes and suffixes from an unlabeled corpus of a language. The underlying theory makes no assumptions on whether the language uses a lot of morphology or not, whether it is prefixing or suffixing, or whether affixes are long or short. It does however make the assumption that 1. salient affixes have to be frequent, i.e., occur much more often that random segments of the same length, and that 2. words essentially are variable length sequences of random characters, e.g., a character should not

occur in far too many words than random without a reason, such as being part of a very frequent affix. The second component, extract paradigms, i.e., sets of affixes, that tend to occur on the same stems. The underlying idea is that the members of a paradigmatic set of affixes alternate on a stem set in higher combined proportions than non-members. It is not necessary that the members pairwise occur with high absolute frequency on the same stems.

The two components are then used, with various additional measures, in four applications, which are given a separate section, and are empirically evaluated individually.

2 A Survey of Work on Unsupervised Learning of Morphology

For the purposes of the present survey, we use the following definition of Unsupervised Learning of Morphology (ULM).

Input: Raw (unannotated, non-selective) natural language text data

Output: A description of the morphological structure (there are various levels to be distinguished; see below) of the language of the input text

With: As little supervision, i.e., parameters, threholds, human intervention, model selection during development etc., as possible

Some approaches have explicit or implicit biases towards certain kinds of languages; they are nevertheless considered to be ULM for this survey.

Morphology may be narrowly taken as to include only derivational and grammatical affixation, where the number of affixations a root may take is finite and the order of affixation may not be permuted. This survey also subsumes attempts that take a broader view including clitics and compounding (and there seems to be no reasons in principle to exclude incorporation and lexical affixes). A lot of, but not all, approaches focus on concatenative morphology/compounding only.

All works considered in this survey are designed to function on orthographic words, i.e., raw text data in an orthography that segments on the word-level. Crucially, this excludes work the rather large body of work that only targets word-segmentation, i.e., segmenting a sentence or a full utterance into words. However, work that explicitly aims to treat both word-segmentation and morpheme-segmentation in one algorithm are included. Hence, subsequent uses of the term segmentation in the present survey is to be understood as morpheme-segmentation rather than word-segmentation. We prefer the term segmentation to analysis since, in general in ULM, the algorithm does not label the segments.

Work that requires selective input, such as 'singular-plural pairs', or 'all members of a paradigm' are excluded, unless such pairs/sets are extracted from raw text in an unsupervised manner as well. Similarly, we exclude work where some (small) amount of annotated data, some (small) amount of existing rule sets, or resources such as a parallel corpus, are mandatory.

One of the matters that varies the most between different authors is the desired outcome. It is useful to set up the implicational hierarchy shown in Table 1 (which need of course not correspond to steps taken in an actual algorithm). The division is implicational in the sense that if one can do the morphological analysis of a lower level in the table, one can also easily produce the analysis of any of the above levels. For example, if one can perform segmentation into stem and affixes, one can decide if two word are of the same stem. The converse need not hold, it is perfectly possible to answer the question of whether two words

Affix list	A list of the affixes.
↑	
Same-stem decision	Given two words, decide if they are affixations of the same stem.
↑	
Segmentation	Given a word, segment it into stem and affix(es).
↑	
Paradigm list	A list of the paradigms.
↑	
Lexicon+Paradigm	A list of the paradigms and a list of all stems with information of which paradigm each stem belongs to.

Table 1. Levels of power of morphological analysis. No distinction is made between probabilistic and non-probabilistic versions.

are of the same stem with high accuracy, without having to commit to what the actual stem should be.

Many recent articles fail to deal properly with previous and related work, some reinvent heuristics that have been sighted earlier, and there little modularization taking place. Previous surveys and overviews are Kurimo et al. 2007a, McNamee 2008, Kurimo and Varjokallio 2008, Kurimo et al. 2007c,b, Hammarström 2007a, Kurimo et al. 2008, Kurimo and Turunen 2008, Powers 1998, Borin 1991, Clark 2001, Roark and Sproat 2007, Goldsmith pear, Borin 2009, Batchelder 1997:66-68 and the related-work sections of research papers. Nevertheless, there is no survey to date which is comprehensive and which discusses the ideas in the field critically.

We will not attempt a comparison in terms of accuracy figures as this is wholly impossible, not only because of the great variation in goals but also because most descriptions do not specify their algorithm(s) in enough detail. Furtunately, this aspect is better handled in controlled competitions, such as the Unsupervised Morpheme Analysis – Morpho Challenge[1] which offers tasks of segmentation of Finnish, English, German, Arabic and Turkish.

2.1 Roadmap and Synopsis of Earlier Studies

A chronological listing of earlier work (with very short characterizations) is given in Table 2-4. Several papers are co-indexed if they represesent essetially the same line of work bu essentially the same author(s).

Given the number of algorithms proposed, it is impossible to go through the methods and ideas individually. However, the main trends are as follows.

[1] Website `http://www.cis.hut.fi/morphochallenge2009/` accessed 10 September 2009.

	Model	Superv.	Experimentation	Learns what?
Harris 1955, 1968, 1970	C	T	English	Segmentation
Andreev 1965, 1967:Chapter 2	C	T	Hungarian/Russian (I)	Unclear
Gammon 1969	C	T	English	Segmentation
Lehmann 1973:71-93	C	T	German (I)	Segmentation
de Kock and Bossaert 1969, 1974, 1978	C	T	French/Spanish	Lexicon+Paradigms
Hafer and Weiss 1974	C	T	English (IR)	Segmentation
Faulk and Gustavson 1990	C	T	English (I)	Segmentation
Klenk and Langer 1989	C	T	German	Segmentation
Langer 1991	C	T	German	Segmentation
Redlich 1993	C	T	English (I)	Segmentation
Klenk 1992, 1991	C	T	Spanish	Segmentation
Flenner 1992, 1994, 1995	C	T	Spanish	Segmentation
Janßen 1992	C	T	French	Segmentation
Juola et al. 1994	C	T	English	Segmentation
Brent 1993, 1999, Brent et al. 1995, Snover 2002, Snover et al. 2002, Snover and Brent 2001, 2003	C	T	English/Child-English/Polish/ French	Segmentation
Deligne and Bimbot 1997, Deligne 1996	C	T	English/French (I)	Segmentation
Yvon 1996	C	T	French (I)	Segmentation
Kazakov 1997, Kazakov and Manandhar 1998, 2001	C	T	French/English	Segmentation
Jacquemin 1997	C	T	English	Segmentation
Cromm 1997	C	T	German	Unclear
Gaussier 1999	C	T	French/ English (I)	Lexicon+Paradigms
Déjean 1998a,b	C	T	Turkish/English/ Korean/French/ Swahili/ Vietnamese (I)	Affix Lists
Medina Urrea 2000, 2003, 2006	C	T	Spanish	Affix List
Schone and Jurafsky 2000, 2001a, Schone 2001	C	T	English	Segmentation
Goldsmith 2000, 2001, 2006, Belkin and Goldsmith 2002, Goldsmith et al. 2001, Hu et al. 2005b, Xanthos et al. 2006	C	T	English (I)	Lexicon+Paradigms
Baroni 2000, 2003	C	T	Child-English/ English	Affix List
Cho and Han 2002	C	T	Korean	Segmentation
Sharma et al. 2002, 2003, Sharma and Das 2002	C	T	Assamese	Lexicon+Paradigms
Baroni et al. 2002	C/NC	T	English/German (I)	Related word pairs
Bati 2002	C/NC	T	Amharic	Lexicon+Paradigms

Table 2. Very brief roadmap of earlier studies [Page 1(3)]. Abbrevations in the Table: C = Concatenative, NC = Non-concatenative, T = Threshold(s) and Parameter(s) to be set by a human, I = Impressionistic evaluation, IR = Evaluation only in terms of Information Retrieval Performance. RR = Hand-written rewrite rules.

	Model	Superv.	Experimentation	Learns what?
Creutz 2003, 2006, Creutz and Lagus 2002, 2005c, 2004, 2005a,b, 2007, Creutz et al. 2005b, Hirsimäki et al. 2003, Creutz et al. 2005a	C	T	Finnish/Turkish/ English	Segmentation
Kontorovich et al. 2003	C	T	English	Segmentation
Medina Urrea and Díaz 2003, Medina-Urrea 2006, 2008	C	T	Chuj/Ralámuri/Czech	Affix List
Mayfield and McNamee 2003, McNamee and Mayfield 2007	-	-	8 West European languages (IR)	Same-stem
Zweigenbaum et al. 2003, Hadouche 2002	C	T	Medical French	Segmentation
Pirrelli et al. 2004, Pirrelli and Herreros 2007	C	T	Italian/English/Arabic	Unclear
Johnson and Martin 2003	C	T	Inuktitut	Unclear
Katrenko 2004	C	T	Ukrainian	Lexicon+Paradigms
Ćavar et al. 2004a,b, Ćavar et al. 2005, 2006	C	T	Child-English	Unclear
Rodrigues and Ćavar 2005, 2007	NC	T	Arabic	Segmentation
Monson 2004, 2009, Monson et al. 2007b, 2004, 2007a, 2008a,b,c	C	T	English/Spanish/ Mapudungun (I)	Segmentation
Yarowsky and Wicentowski 2000, Wicentowski 2002, 2004	C/NC	AP	30-ish mostly European type languages	Segmentation
Gelbukh et al. 2004	C	-	English	Segmentation
Argamon et al. 2004	C	T	English	Segmentation
Goldsmith et al. 2005, Hu et al. 2005a	C/NC	T	Unclear	Unclear
Bacchin et al. 2005, 2002b,a, Nunzio et al. 2004	C	T	Italian/English	Segmentation
Oliver 2004:Chapter 4-5	C	T	Catalan	Paradigms
Bordag 2005b,a, 2007b,a,c	C	T	English/German	Segmentation
Hammarström 2006a, 2005, 2006a,b, 2007b, 2009a	C	-	Maori to Warlpiri	Same-stem
Bernhard 2005a,b, 2006, 2007a,b	C	T	Finnish/Turkish/English	Segmentation+Relate sets of words
Keshava and Pitler 2005	C	T	Finnish/Turkish/English	Segmentation
Johnsen 2005	C	T	Finnish/Turkish/English	Segmentation
Atwell and Roberts 2005	C	T	Finnish/Turkish/English	Segmentation
Dang and Choudri 2005	C	T	Finnish/Turkish/English	Segmentation
ur Rehman and Hussain 2005	C	T	Finnish/Turkish/English	Segmentation
Jordan et al. 2006, 2005	C	T	Finnish/Turkish/English	Segmentation
Goldwater et al. 2005, Goldwater 2007, Naradowsky and Goldwater 2009	C	T	English/Child-English	Segmentation
Freitag 2005	C	T	English	Segmentation
Golcher 2006	C	-	English/German	Lexicon+Paradigms
Arabsorkhi and Shamsfard 2006	C	T	Persian	Segmentation
Chan 2006	C/NC	T	English	Paradigms
Demberg 2007	C/NC	T	English/German/Finnish/ Turkish	Segmentation
Dasgupta and Ng 2006, 2007, Dasgupta and Ng. 2007, Dasgupta 2007	C	T	Bengali	Segmentation
De Pauw and Wagacha 2007	C/NC	T	Gikuyu	Segmentation
Tepper 2007, Tepper and Xia 2008	C/NC	T+RR	English/Turkish	Analysis

Table 3. Very brief roadmap of earlier studies [Page 2(3)]. Abbrevations in the Table: C = Concatenative, NC = Non-concatenative, T = Threshold(s) and Parameter(s) to be set by a human, I = Impressionistic evaluation, IR = Evaluation only in terms of Information Retrieval Performance. RR = Hand-written rewrite rules.

	Model	Superv.	Experimentation	Learns what?
Xanthos 2007	NC	T	Arabic	Lexicon+Paradigms
Majumder et al. 2007, 2008	C	T	French/Bengali/French/ Bulgarian/Hungarian	Analysis
Zeman 2007, 2008a,b	C	-	Czech/English/German/ Finnish	Segmentation+ Paradigms
Kohonen et al. 2008	C	T	Finnish/Turkish/English	Segmentation
Goodman 2008	C	T	Finnish/Turkish/English	Segmentation
Pandey and Siddiqui 2008	C	T	Hindi	Segmentation+ Paradigms
Johnson 2008	C	T	Sesotho	Segmentation
Snyder and Barzilay 2008	C/NC	T	Hebrew/Arabic/Aramaic/ English	Segmentation
Spiegler et al. 2008	C	T	Zulu	Segmentation
Moon et al. 2009	C	T	English/Uspanteko	Segmentation
Poon et al. 2009	C	T	Arabic/Hebrew	Segmentation

Table 4. Very brief roadmap of earlier studies [Page 3(3)]. Abbrevations in the Table: C = Concatenative, NC = Non-concatenative, T = Thresholds and Parameters to be set by a human, I = Impressionistic evaluation, IR = Evaluation only in terms of Information Retrieval Performance. RR = Hand-written rewrite rules.

There are basically three approaches to the problem:

a. Group and Abstract: In this family of methods, words are first grouped (clustered into sets, paired, shortlisted etc) according to some metric, which is typically string edit distance, but may include semantic features (Schone 2001), distributional similarity (Freitag 2005) or frequency signatures (Wicentowski 2002). The next step, is to abstract some morphological pattern that recurs among the groups. Such emergent patterns provide enough clues for segmentation and can sometimes be formulated as rules or morphological paradigms.

b. Frequency and Border: In this family of methods, frequent segments have a direct interpretation as candidates for segmentation. In addition, if a segment occurs with a variety of segments immediately adjacent to it, this is interpreted as evidence for a segmentation border. A typical implementation is to subject the data to a compression formula of some kind, where frequent long segments with clear borders offer the optimal compression gain. The outcome of such a compression scheme gives the segmentation and occasionally paradigm information can be gleaned from co-occurrence and border properties.

c. Features and Classes: In this family of methods, a word is seen as made up of features – n-grams in Mayfield and McNamee (2003), McNamee and Mayfield (2007), and initial/terminal/mid-segment in De Pauw and Wagacha (2007). Features which occur on many words have little selective power across

the words, whereas features which occur seldom, pinpoint a specific word or stem. To formalize this intuition, Mayfield and McNamee (2003) and McNamee and Mayfield (2007) use TF-IDF and De Pauw and Wagacha (2007) use entropy. Classifying an unseen word reduces to using its features to select which word(s) it may be morphologically related to. This decides whether the unseen word is a morphological variant of some other word, and allows extracting the "variation" by which they are related, such as an affix.

The first two, a. and b., enjoy a fair amount of popularity in the present collection of work, though b. is more common and was the only kind used up to about 1997. The last, c., is used only by two sets of authors (cited above). Xanthos (2007) falls outside either category as it attempts to first learn phonological categories and then uses these to infer intercalated morphology (with the observation that, empirically, intercalated morphology does seem to depend on vowel/consonant considerations). The work by de Kock and Bossaert 1969, 1974, 1978, Yvon 1996, Medina Urrea 2003 can favourably be seen as a midway between a. and b. as they rely on sets of four members with a particular affixation arrangement ("squares"), whose existence is governed much by the frequency of the affixes in question. There are, of course, many other lines of work that draw from both a. and b., but in a less cross-cut way.

An obvious advantage of the a. (and to some extent c.) family of methods is that they are capable of handling non-concatenative morphology.

2.2 Discussion

Within the a. family of methods, the main challenge is to avoid the use of thresholds to filter out spurious groupings that come with all of the so far employed grouping criteria.

In the b. family of methods, there are several open questions of interest.

Most (if not all) authors trace the inspiration for their border/frequency heuristics back to Harris (1955). Although Harris was far ahead in conceiving of an algorithm using such counts for segmentation, his description is vague on the role/need for thresholds[2], and the exact formulation of his criterion, namely the size of a segment's successor character set, was shown (in various interpretations) as early as Hafer and Weiss (1974) not to be quite sound – even for English. (Kazakov and Manandhar (2001) identify further theoretical shortcomings). More modern versions have considered the branching signature of a segment's character trie, with better empirical results, but we still do not have a theoretical understanding of the signs of segment combination and alternation.

Another way to use character sequence counts is that associated with Ursula Klenk and various colleagues (see, e.g., Klenk and Langer (1989) for a good explanation). For each character bigram c_1c_2, they record at what percentage there is a morpheme boundary before $|c_1c_2$, between $c_1|c_2$, after $c_1c_2|$, or none.

[2] Though, this is still far superior to the cascade of thresholds advised by the other early pioneer, Andreev (1965).

A new word can then be segmented by sliding a bigram window and taking the split which satisfies the corresponding bigrams the best. For example, given a word *singing*, if the window happens to be positioned at *-gi-* in the middle, the bigram splits $ng|$, $g|i$ and $|in$ are relevant to deciding whether sing$|$ing is a good segmentation. Exactly how to do the split by sliding the window and combining such bigram split statistics is subject to a fair amount of discussion. However, it became apparent that, bigram-splithood is dependent on, e.g., the position in a word – *-ed* is likely at the end of a word, but hardly in any other position – and exception lists and cover-up rules had to be introduced, before the approach was abandoned altogether.

Several different authors in the b. paradigm have hailed Minimum Description Length (MDL) as the motivation for a given formula to compress input data into a morphologically analysed representation. The Minimum Description Length (MDL) principle is a general purpose method of statistical inference. It views the learning/inference process as data compression: for a given set of hypotheses H and data set D, we should try to find the hypothesis in H that compresses D most (Grünwald 2007:3-40). Concretely, such a calculation can take the the following form. If $L(H)$ is the length, in bits, of the description of the hypothesis; and $L(D|H)$ is the length, in bits, of the description of the data when encoded with the help of the hypothesis, then MDL aims to minimize $L(H) + L(D|H)$. In principle, all of the works that have invoked MDL in the ULM-method act as follows. A fix way Q of describing morphological regularities is conceived, which has two components which we may call patterns H and data D. A coding scheme is devised to describe any H and to describe any set of actual words with some specific H and D. A greedy search is done for a local minimum of the sum $L(H) + L(D|H)$ to describe the set of words W (in some approaches) or the bag of tokens C (in other approaches) of the input text data[3]. In these cases, the label MDL, in at least the terminology of Grünwald (2007:37-38), seems to be ill-founded since, crucially, the Q, H, D-search is not among different description *languages*, but among parameters in a fix language. In this respect it is important to note that, compared to the schemes devised so far, Lempel-Ziv compression should yield a superior compression (as, in fact, conceded by Baroni 2000:146-147). However, MDL-inspired optimization schemes have achieved very competetive results in practice.

Lastly, several pieces of work in the b. tradition have attempted to address morphophonological changes in a principled way, though so far these have been developed in close connection with a particular segmentation method and target language.

A perhaps worrying tendency is that, despite extensive cross-citation, there is little transfer between different groups of authors and there is a fair amount of duplication of work. The lack of a broadly accepted theoretical understanding is possibly related to this fact. Few approaches have an abstract model of how words are formed, and thus cannot explain why (or why not) the heuristics

[3] As most approaches define their task as capturing the set of legal morphological forms, their goal should be to compress W, but see Goldwater (2007:53-59) for arguments for compressing C.

employed fail, what kind of errors are to be expected and how the heuristics can be improved. Nevertheless, a model for the simplest kind of concatenative morphology is emerging. Namely, that two sets of random strings, B and S, combine in some way to form a set of words W. For Gelbukh et al. (2004), the segmentation task is to find minimal size $|X| + |Y|$ such that $W \subset \{xy | x \in X, y \in Y\}$. For Bacchin et al. (2005) as well as in the word-segmentation version of Deligne (1996), the segmentation task is to find a configuration of splits for each $w = xy \in W$ such that each x and y occur in as many splits as possible (more precisely, the product, over all words, of the number number of splits for the parts x and y should be maximized). Hammarström (2006a) adds that the formation of W from B and S should be such that each $s \in S$ should occur frequently, which has implications for the segmentation strategy. Brent (1999) devises a precise, but more elaborate, way of constructing W from B and S, but at the price of a large search space, and whose global maximum is hard to characterize intuitively. Kontorovich et al. (2003), Snyder and Barzilay (2008), Goldwater (2007) and Poon et al. (2009) should also be noted for containing generative models.

3 A Naive Theory of Affixation and an Algorithm for Extraction

In this section we present a naive theory on how the simplest kind of affixation in natural languages may behave. The theory allows us to devise an extraction algorithm, i.e., an algorithm that partially undoes the affixation. We discuss the assumptions and thinking behind the theory and algorithm, which actually requires only a few lines to define mathematically. Next, we present and discuss some experimental results on typologically different languages. Finally, we state some conclusions and ideas on future components of unsupervised morphological analysis.

3.1 A Naive Theory of Affixation

Notation and definitions:

- $w, s, b, x, y, \ldots \in \Sigma^*$: lowercase-letter variables range over strings of some alphabet Σ and are variously called words, segments, strings, etc.
- $s \triangleleft w$: s is a terminal segment of the word w, i.e., there exists a (possibly empty) string x such that $w = xs$
- $b \triangleright w$: b is an initial segment of the word w, i.e., there exists a (possibly empty) string x such that $w = bx$
- $W, S, \ldots \subseteq \Sigma^*$: capital-letter variables range over sets of words/strings/segments
- $f_W(s) = |\{w \in W | s \triangleleft w\}|$: the (suffix) frequency, i.e., the number of words in W with terminal segment s
- $S_W = \{s | s \triangleleft w \in W\}$: all terminal segments of the words in W
- $B_W = \{b | b \triangleright w \in W\}$: all initial segments of the words in W
- $uf_W(u) = |\{(x, y) | xuy = w \in W\}|$: the substring frequency of u, i.e., the number times u occurs as a substring in the set of words W (x and y may be empty).
- $nf_W(u) = uf_W(u) - f_W(u)$: the non-final frequency of u, i.e. the substring frequency minus those in which it occurs as a suffix.
- $|\cdot|$: is overloaded to denote both the length of a string and the cardinality of a set
- $''$: denotes the empty string

Assume we have two sets of random strings over some alphabet Σ:

- Bases $B = \{b_1, b_2, \ldots, b_m\}$
- Suffixes $S = \{s_1, s_2, \ldots, s_n\}$

Such that:

Arbitrary Character Assumption (ACA): The probability of each character c in a word $w = xcy \in B, S$ does not depend on the strings x, y around it

Note that B and S need not be of the same cardinality and that any string, including the empty string, could end up belonging to both B and S. They need neither to be sampled from the same distribution; pace the requirement, the distributions from which B and S are drawn may differ in how much probability mass is given to strings of different lengths. For instance, it would not be violation if B were drawn from a a distribution favouring strings of length, say, 42 and S from a distribution with a strong bias for short strings.

Next, build a set of affixed words $W \subseteq \{bs | b \in B, s \in S\}$, that is, a large set whose members are concatenations of the form bs for $b \in B, s \in S$, such that:

Frequent Flyer Assumption (FFA): The members of S are frequent. Formally: Given any $s \in S$: $f_W(s) >> f_W(x)$ for all x such that 1. $|x| = |s|$; and 2. not $x \lhd s'$ for all $s' \in S$).

In other words, if we call $s \in S$ a *true suffix* and we call x an *arbitrary segment* if it neither a true suffix nor the terminal segment of a true suffix, then any true suffix should have much higher frequency than an arbitrary segment of the same length.

One may legimately ask to what extent words of real natural languages fit the construction model of W, with the strong ACA and FFA assumptions, stated above. For instance, even though natural languages often aren't written phonemically, it is not difficult to find examples of languages that have phonotactic constraints on what may appear at the beginning or end of a word, e.g., Spanish **st-* may not begin a word and yields *est-* instead. This is a violation of ACA because the probability of observing *s* is much lower and that of *e* significantly higher, depending on whether or not the empty string is on the left of it, i.e., when initial. Another violation of ACA is that (presumably all (Ladefoged 2005)) languages disallow or disprefer a consonant vs. a vowel conditioned by the vowel/consonant status of its predecessor. However, for the present extraction algorithm, if a certain element occurs with *less* frequency than uniform random (the best example would be click consonants which, in some languages, e.g., Eastern !Xóõ (Traill 1994), occur only initially), this is less of a problem in practice.

As for FFA, we may have breaches such as Biblical Aramaic (Rosenthal 1995) where an old -$\bar{a}$ element appears on virtually everywhere on nouns, making it very frequent, but no longer has any meaning synchronically. Also, one can doubt the requirement that an affix should need to be frequent; for instance, the Classical Greek inflectional (lacking synchronic internal segmentation) alternative medial 3p. pl. aorist imperative ending -$\sigma\theta\omega\nu$ (Blomqvist and Jastrup 1998), is not common at all.

Just how realistic the assumptions are is an empirical question, whose answer must be judged by experiments on the relevant languages. In the absense of fully

Positions	Distance
$\|\|p_1 - p_2\|\|$	0.47
$\|\|p_1 - p_3\|\|$	0.36
$\|\|p_1 - p_4\|\|$	0.37
$\|\|p_2 - p_3\|\|$	0.34
$\|\|p_2 - p_4\|\|$	0.23
$\|\|p_3 - p_4\|\|$	0.18

Table 5. Difference between character distributions according to word position.

annotated annotated test sets for diverse languages, and since the author does not have access to the Hutmegs/CELEX gold standard sets for Finnish and English (Creutz and Lindén 2004), we can only give some illustrative experimental data.

ACA: If the probability of a character does not depend on the segment preceding it, it follows that it should not depend on the length of the segment preceding it either. On a New Testament corpus of Basque (Leizarraga 1571) we computed the probability of a character appearing in the initial, second, third or fourth position of the word. Since Basque is entirely suffixing, if it complied to ACA, we'd expect those distributions to be similar. However, when we look at the difference of the distributions in terms of variation distance between two probability distributions ($||p - q|| = \frac{1}{2}\sum_x |p(x) - q(x)|$), it shows that they differ considerably – especially the initial position proves more special – as shown in Table 5.

FFA: As for the FFA, we checked a corpus of bible portions of Warlpiri (Summer Institute of Linguistics 2001). This was chosen because it is one of the few languages known to the author where data was available and which has a decent amount of frequent suffixes which are also long, e.g., case affixes are typically bisyllabic phonologically and five-ish characters long orthographically. Since the orthography employed marks segmentation, it is easy to compute FFA statistics on the words by removing the segmentation marking artificially. Comparing with the lists in Nash (1980:Chapter 2) it turns out that FFA is remarkably stable for all grammatical suffixes occuring in the outermost layer. There are, however, the expected kind of breaches; e.g., a tense suffix *-ku* combined with a final vowel *-u* which is frequent in some frequent preceding affixes making the terminal segment *-uku* more frequent than some genuine three-letter suffixes.

The language known to the author which has shown the most systematic disconcord with the FFA is Haitian Creole (also in bible corpus experiments (American Bible Society 1999)). Haitian creole has very little morphology of its own but owes the lion's share of its words to French. French derivational morphemes abound in these words, e.g., *-syon*, which have been carefully shown by Lefebvre (2004) not to be productive in

Haitian Creole. Thus, the little morphology there is in Haitian creole is very difficult to get at without also getting the French relics.

3.2 An Algorithm for Affix Extraction

The key question is, if words in natural languages are constructed as W explained above, can we recover the segmentation? That is, can we find B and S, given only W? The answer is yes, we can partially decide this. To be more specific, we can compute a score Z_W such that $Z_W(x) > Z_W(y)$ if $x \in S$ and $y \notin S$. In general, the converse need not hold, i.e., if both $x, y \in S$, or both $x, y \notin S$, then it may still be that $Z_W(x) > Z_W(y)$. This is equivalent to constructing a ranked list of all possible segments, where the true members of S appear at the top, and somewhere down the list the junk, i.e., non-members of S, start appearing and fill up the rest of the list. Thus, it is not said *where* on the list the true-affixes/junk border begins, just that there is a consistent such border.

Now, how should this list be computed? All terminal segments are contained in the set S_W, the question is just to order them. We shall now define three properties that we argue will be enough to put the S-belonging affixes at the top. For a terminal segment s, define:

Frequency The frequency $f_W(s)$ of s (as a terminal segment).

Curve Drop First, for s, define its curve $C_s(c)$ which is a probability distribution on Σ:

$$C_s(c) = \frac{f_W(cs)}{f_W(s)}$$

Next, more importantly, define its *curve drop* $\overline{C}(s)$ which is a value in $[0, 1]$:

$$\overline{C}(s) = \frac{1 - max_c(C_s(c))}{1 - \frac{1}{|\Sigma|}}$$

Random Adjustment First, for s, define its probability as:

$$P_W(s) = \frac{f_W(s)}{\sum_{s'} f_W(s')}$$

Second, equally straighfowardly, for an arbitrary segment u, define its non-final probability as:

$$nP_W(u) = \frac{nf_W(u)}{\sum_{u'} nf_W(u')}$$

Finally, for a terminal segment s, define its *random adjustment* $RA(s)$ which a value in Q^+:

$$RA(s) = \begin{cases} \frac{P_W(s)}{nP_W(s)} & \text{if } nP_W(s) > 0 \\ 1.0 & \text{otherwise} \end{cases}$$

It is appropriate now to show the intuition behind the definitions. There isn't much to comment on frequency, so we'll go to curve drop and random adjustment. All examples in this section come from the Brown corpus (Francis and Kucera 1964) of one million tokens ($|W| = 47178$ and $|S_W| = 154407$).

The curve drop measure is meant to predict when a suffix is well-segmented to the left. Consider a suffix s, in all the words on which it appears, there is a preceding character c. Figure 1 shows examples of the frequency distribution on preceding character for example suffixes *-ing* and *-ng*. The reasoning is as follows. If s is a true suffix and is well-segmented to the left, then its curve-drop value should be high. Frequent true suffixes that attach to bases whose last character is random should have a close to uniform curve. On the other hand, if the curve drop value is low it means there is a character that suspiciously often precedes s. However, if s weren't a true suffix to begin with, perhaps just a frequent but random character, then we expect its curve drop value to be high too! To exemplify this, we have $\overline{C}(ing) \approx 0.833$, $\overline{C}(ng) \approx 0.029$ and $\overline{C}(a) \approx 0.851$.

The random adjustment measure it precisely to distinguish what a "frequent but random segment" is, that is, discriminate, e.g., *-a* versus *-ing* as well as *-a* versus *-ng*. Now, how does one know whether something is random or not? One approach would be to say the shorter the segment the more random. Although it is possible to get this to work reasonably well in practice, it has some drawbacks. First, it treats all segments of the same length the same, which may be too brutal, e.g., should *-s* be penalized as much as *-a*? Second, it might be considered too vulnerable to orthography. For example if a language has an odd trigraph for some phoneme, we are clearly going to introduce an error source. Instead we propose that a segment is random iff it has similar probability in any position of the word. This avoids the "flat length"-problems but has others, which we think are less harmful. First, we might get sparse data which can either be back-off smoothed or, like here, effectively ignored (where we lack occurence we set the RA to 1.0). Second, phonotactic or orthographic constraints may cause curiousities, e.g., English *y* is often spelled *i* when medial as in *fly* vs. *flies*.

To put it all together, we propose the characterization of suffixes in terms of the three properties as shown in Table 6. The terms high and low are of course idealized, as they are really gradient properties.

As seen from the table, we hold that true suffixes (and only true suffixes) are those which have a high value for all three properties. Therefore, we define our final ranking score, the $Z_W : S_W \rightarrow \mathbf{Q}$:

$$Z_W(s) = \overline{C}(s) \cdot RA(s) \cdot f_W(s) \tag{1}$$

Thus we are deliberate saying that if you have a not-so-high relative value for one of the properties, you can compensate to some extent by having very

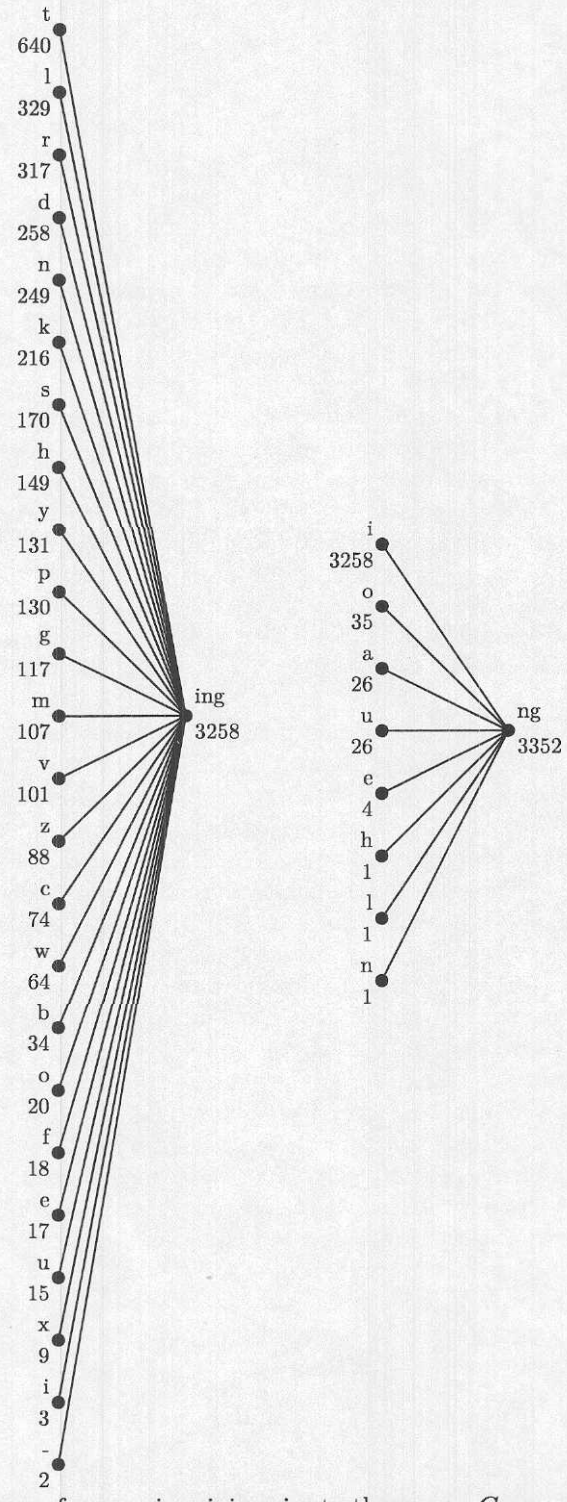


Figure 1. The curve frequencies giving rise to the curves C_{ing} and C_{ng} respectively.

f_W	$\overline{C}$	RA	Example	Label
high	high	high	*-ing*	True suffix
high	high	low	*-a*	Frequent random segment
high	low	high	*-ng*	Tail of true suffix
high	low	low	N/A	Second part of a digraph
low	high	high	*-oholic*	Infrequent true suffix
low	high	low	*-we*	Happenstance low RA-segment?
low	low	high	*-icz*	Tail of foreign personal name ending
low	low	low	*-ebukadnessar*	Infrequent segment

Table 6. The logically possible condigurations of the three suffix properties, accompanied by an appropriate linguistically inspired label and an example from English.

high relative values for the other properties (relative here means relative to the corresponding values of other suffixes). It is instructive to look at what happens in a few interesting cases:

1. We have two suffixes such that one is an enlargement of the other by a random segment, e.g., *-ting* versus *-ing*, where the true suffix is the shorter one.

   In this case, we expect both to have similar high $\overline{C}$, the longer one should have higher RA and, by necessity, the shorter one should have signficantly higher frequency. Example values for *-ing* versus *-ting* are shown in Table 7.

   Here, we see that the shorter wins out and we can use that fact to weed out the longer one (cf. purging below). (One might think that in a "perfect" situation, the f_W and RA would cancel out, leaving the situation a tie. However, RA will not cancel f_W in a language which, like all languages we know of, has more non-final than final "positions of segments in words", and also, ceteris paribus, we expect a higher frequency to yield a more reliable curve drop value.)

2. We have two suffixes such that one is a tail of the other, but both are true suffixes, and they just happen to share a segment, e.g., *-ly* versus *-y*.

   In this case, we succeed in keeping both if the longer wins out on a better curve-drop and random adjustment. In fact, as shown in Table 7 this is exactly what happens with *-ly* versus *-y*.

3. We have two true suffixes which incidentally share an ending which is not a true suffix. Although easy to find in other languages, I failed to find an example of this in English without confounding factors, but we can imagine one, for example *-xz* versus *-yz*. Given the assumption that *-z* itself is not a true suffix, $f_W(z)$ should not be many times higher than $f_W(xz) + f_W(yz)$, thus its curve-drop not many percent, if at all, higher

s	$f_W(s)$	$\overline{C}(s)$	$RA(s)$	$Z_W(s)$
-ing	3258	0.83	19.6	53309.3
-ting	640	0.69	31.5	13929.5
-y	3931	0.63	5.8	14402.7
-ly	1532	0.76	23.4	27282.2
-t	2796	0.74	0.50	1040.6
-st	561	0.64	0.68	246.3
-ist	202	0.81	1.29	213.9
-est	213	0.88	1.82	341.4
-s	11220	0.80	2.49	22514.8
-ings	205	0.89	60.5	11034.2
-ations	215	0.86	110.9	20482.1

Table 7. Values for some borderline cases.

than 0.5, and of course, $RA(z) \approx 1$. On the other hand, by assumption of being true suffixes, *-xz* and *-yz* should have high curve-drop values, and outperform *-z* on RA.

Empirically, the prediction is wrong in the case *-est* versus *-est*, as shown in 7. But, on the other hand, *-ist* and *-est* can hardly be said to satisfy FFA.

4. We have two true stacked suffixes which share an ending and this ending is also true suffix, e.g., *-ations* versus *-ings*.

   As opposed to the above case, *-s* will appear in a lot of other places than after *-ing* and *-ation*, and is consequently given a higher score as shown in Table 7.

As these considerations exemplify, the formal criterion mostly conforms to linguistic analysis, but as noted as noted in the third example, the outcomes occasionally disconcords with linguistic analysis.

A theoretical weakness with the RA-value as computed at present is when applied to languages which stack suffixes after each other. English does this to a small extent, as in *-ing* vs. *-ings*. In such cases, when calculating the non-final frequency of *-ing* one would like to count an occurence of *-ing-* in *-ings* as a final occurence. But this would require knowing beforehand that *-s* is a true suffix as opposed to *-ings*. Fortunately, the impact of this drawback, for languages such as English where stacking is rather uncommon, appears not to be crucial. Even if suffixes occur when they are "almost" final, they still don't occur in the when initial or in the mid-span of the word.

As a last discussion note, it is tempting to leave out the f_W-component in the calculation of the ranking. The frequency is really only needed when deciding between suffixes which are tails of each other – it plays no crucial role in ranking between suffixes which don't share a tail. If frequencies are used only to purge out losers in tail-indexed sets of suffixes, the resulting list will also

Input: A text corpus C

Step 1. Extract the set of words W from C (thus all contextual and word-frequency information is discarded)

Step 2. Calculate $f_W(s)$, $\overline{C}(s)$ and $RA(s)$ for each $s \in S_W$

Step 3. Combine $Z_W(s) = \overline{C}(s) \cdot RA(s) \cdot f_W(s)$

Table 8. Summary of affix-extraction algorithm.

contain some non-FFA true suffixes but also too many spurious things, such as foreign personal name endings.

To sum up, the final Z_W-score in equation 1 is the one that purports to have the property that $Z_W(x) > Z_W(y)$ if $x \in S_W$ and $y \notin S_W$ – at least if purged (see below). We cannot give a formal proof that languages satisfying ACA and FFA should get a faultless ranking list because this is true only in a heuristic sense. To set bounds on the probability for it to hold is also depends on a lot of factors that are hard, or at least inelegant, to characterize. We hope, however, to have sketched the how the ACA and FFA assumptions are used.

A summary of the algorithm described in this section is displayed in Table 8.

The time-complexity bounding factor is the number of (final and non-final) segments, which is linear (in the size of the input) if words are bounded in length by a constant and quadratic in the (really) worst case if not.

3.3 Experimental Results

For an English bible corpus (King James 1977) we get the top 30 plus bottom 3 suffixes as shown in Table 9.

English has little affixation compared to, e.g., Turkish which is at the opposite end of the typological scale (Dryer 2005). The corresponding results for Turkish on a bible corpus (American Bible Society 1988) is shown in Table 10.

The results largely speak for themselves but some comments are in order. As is easily seen from the lists, some suffixes are suffixes of each other so one could *purge* the list in some way to get only the most "competitive" suffixes. One purging strategy would be to remove x from the list if there is a z such that $x = yx$ and $Z_W(z) > Z_W(x)$ (this would remove, e.g., *-ting* if *-ing* is above it on the list). A more sophisticated purging method is the following, which does slightly more. First, for a word $w \in W$ define its best segmentation as: $Segment(w) = argmax_{s \lhd w} Z_W(s)$. Then purge by keeping only those suffixes which are the best parse for at least one word: $S'_W = \{s \in S_W | \exists w Segment(w) = s\}$.

Such purging kicks out the bulk of "junk" suffixes. Table 11 shows the numbers for English, Turkish and the virtually affixless Maori (Bauer et al. 1993). It should noted that "junk" suffixes still remain after purging – typically

-ed	15448.4	*-s*	3407.3
-eth	12797.1	*-ions*	2684.5
-ted	11899.4	*-est*	2452.6
-iah	11587.5	*-sed*	2313.7
-ly	10571.2	*-y*	2239.2
-ings	8038.9	*-leth*	2166.3
-ing	7292.8	*-nts*	2122.6
-ity	6917.6	*-ied*	1941.7
-edst	6844.7	*-ened*	1834.9
-ites	5370.2	*-ers*	1819.5
-seth	5081.6	*-ered*	1796.7
-ned	4826.7	*-ded*	1582.2
-s'	4305.2	*-neth*	1540.0
-nded	3833.8	...	...
-ts	3783.1	*-ig*	0.0
-ah	3766.9	*-io*	0.0
-ness	3679.3	*-ti*	0.0

Table 9. Top 30 and bottom 3 extracted suffixes for an English bible corpus. The high placement of English *-eth* and *-iah* are due to the fact that the bible version used has drinketh, sitteth etc and a lot of personal names in *-iah*.

common stem-final characters – and that there is no simple relation between the number of suffixes left after purging and the amount of morphology of the language in question. Otherwise we would have expected the morphology-less Maori to be left with no, or 28-ish, suffixes or at least less than English.

A good sign is that the purged list and its order seems to be largely independent of corpus size (as long as the corpus is not very small) but we do get some significant differences between bible English and newspaper English.

We have chosen to illustrate using affixes but the method readily generalizes to prefixes as well and even prefixes and suffixes at the same time. As an example of this, we show top-10 purged prefix-suffix scores in the same table also for some typologically differing languages in Table 12. Again, we use bible corpora for cross-language comparability (Swedish (Svenska Bibelsällskapet 1917) and Swahili (British & Foreign Bible Society 1953)). The scores have been normalized in each language to allow cross-language comparison – which, judging from the table, seems meaningful. Swahili is an exclusively prefixing language but verbs tend to end in *-a* (whose status as a morpheme is the linguistic sense can be doubted), whereas Swedish is suffixing, although some prefixes are or were productive in word-formation.

A full discussion of further aspects such as a more informed segmentation of words, peeling of multiple suffix layers and purging of unwanted affixes requires, is beyond the scope of this paper.

-larına	71645.4	*-adılar*	16587.9
-larından	47941.9	*-lerinden*	15201.1
-lerinin	43917.3	*-nden*	14082.2
-lerden	36294.0	*-sinin*	13493.9
-inden	35258.2	*-nin*	12340.9
-iyorlardı	28716.2	*-yorsunuz*	12135.0
-arak	27774.1	*-larla*	12069.7
-iyorsunuz	25403.1	*-en*	11513.5
-inin	25045.5	*-ten*	11424.0
-dılar	20718.7	*-siniz*	11043.0
-lere	20718.2	*-madılar*	10958.9
-ip	20431.2	*-lardan*	10428.1
-dan	19468.4	*-siniz*	10391.1
-ndan	18556.3	*-...*	...
-ından	18226.3	*-ist*	0.0
-yorlardı	18097.1	*-iy*	0.0
-acaksınız	16751.1	*-yo*	0.0

Table 10. Top 30 and bottom 3 extracted suffixes for Turkish. Most of these are really compounds of two suffixes, showing that some adaptation to multi-layer suffixing languges is appropriate.

	Corpus	Tokens	$\|W\|$	$\|S_W\|$	$\|S'_W\|$
Maori	British & Foreign Bible Society 1996	1 101 665	8 354	23 007	78
English	King James 1977	917 634	12 999	39 845	63
Turkish	American Bible Society 1988	574 592	56 881	175 937	122

Table 11. Figures for different languages on the effects on the size of the suffix list after purging.

3.4 Conclusion

We have presented a new theory of affixation and a parameter-less efficient algorithm for collecting affixes from raw corpus data of an arbitrary language. Depending on one's purposes with it, a cut-off point for the collected list is still missing, or at least, we do not consider that matter here. The results are promising and competitive but at present we lack formal evaluation in this respect. Future directions also include a more specialized look into the relation between affix-segmentation and paradigmatic variation and further exploits into layered morphology.

Acknowledgements

The author has benefited much from discussions with Bengt Nordström. The author is also grateful to Bob Carpenter for pointing out a grave technical error

Swedish		English		Swahili	
för-	0.097	*-ed*	0.132	*-a*	0.100
-en	0.086	*-eth*	0.109	*wa-*	0.095
-na	0.036	*-iah*	0.099	*ali-*	0.065
-ade	0.035	*-ly*	0.090	*nita-*	0.059
-a	0.034	*-ings*	0.068	*aka-*	0.049
-ar	0.033	*-ing*	0.062	*ni-*	0.046
-er	0.033	*-ity*	0.059	*ku-*	0.044
-as	0.032	*-edst*	0.058	*ata-*	0.042
-s	0.031	*-ites*	0.046	*ha-*	0.032
-de	0.031	*-s'*	0.036	*a-*	0.031
...	...	...	...	...	...

Table 12. Comparative figures for prefix vs. suffix detection.

in an earlier version of this paper. We also wish to extend special thanks to ASEDA for granting access to electronic versions of the Warlpiri bible texts.

4 Affix Alternation

An important generalization concerns that of how affixes and stems may and may not combine. In many languages, not all affixes can occur on all stems, rather, one finds that affixes form groups, where each group is associated with a subset of stems on which any affix from the group may occur. We may call such a group of affixes a *paradigm*. For example, according to linguistic analysis of English morphology, the suffix set $\{-ing, -ed, -s, -'', -er, -ers\}$ can be argued to form a valid paradigm which occurs on stems such as 'play' and 'ask' but not 'bus', whereas $\{-xt, -ing, -blurb, -ation\}$ is not a valid paradigm.

In the present section we develop a method to find morphological paradigms given raw text data. First we go through some key properties of paradigms which have immediate bearing on how to formally characterize paradigms and explain why baseline approaches such as k-means clustering on Hamming distance are less attractive. Next we introduce a metric that takes a candidate paradigm and gives a score between 0 and 1 indicating "how much" the affixes in the paradigm occur on the same stems, i.e., how "good" the paradigm is. This metric can then be used for greedy searching through the paradigm space to find locally optimal paradigms. The resulting paradigms, with their scores, can be used in various applications (as in the coming sections of this chapter) and are evaluated indirectly this way. We have not yet conceived a final "extraction"-phase to extract all the salient paradigms (or the like), wherefore there is no evaluation simpliciter against some gold standard of paradigms.

4.1 Paradigms

The Nature of Paradigms

For our purposes, the definition of a paradigm is a maximally large set of affixes whose members systematically occur on an open class of stems. There are a few things to note:

- The number of theoretically possible paradigms is exponential in the number of affixes (as paradigms are sets of affixes).
- Empirically, languages tend to have a small number of paradigms.
- Empirically, languages tend to have only small paradigms, i.e., the number of affixes in each paradigm is small. Languages which have several layers of affixes can be said to obey generalization in the sense that each layer has few members, whereas, conversely, the full paradigm achieves considerable size combinatorially.
- Paradigms do not need to be disjoint (in real languages they are typically not)

The Evidence for Paradigms

Without any language specific knowledge, basically the only evidence at hand is co-occurrence of stems and affixes (i.e., when a word occurs in the corpus it evidences the co-occurrence of a (hypothetical) stem and suffix making up that word). Seen like this, the problem at hand bears obvious similarities to other areas of computational linguistics, e.g., word/document co-occurrence models in Information Retrieval (Baeza-Yates and Ribeiro-Neto 1997) or word/word co-occurrence in Part-of-Speech clustering (Redington et al. 1998, Pereira et al. 1993, Schütze 1993), and belongs more generally within a vector space pattern classification framework (Duda et al. 2001, Gordon 1999).

Paradigm induction would be an easy problem if all affixes that *could* legally appear on a word *did* appear on each such word in a raw text corpus. This is, as is well-known, far from the case. A typical corpus distribution is that a few lexemes appear very frequently but by far most lexemes appear once or only a few times (Baayen 2001). What this means for morphology is that most lexemes will appear with only one or a minority of their possible affixes, even in languages with relatively little morphology.

More formally, consider a morphological paradigm (set of suffixes) P that is a true paradigm according to linguistic analysis. If k lexemes that are conjugated according to P occur in a corpus, each of the k lexemes will occur in $1 \leq i \leq |P|$ forms. The number of forms i that a lexeme occurs in will not be normally distributed. Most lexemes will occur in only one form, and only very few, if any, lexemes will occur in all $|P|$ forms. It appears that for most languages and most paradigms, the number of lexemes that occur in i forms tends to decrease logarithmically in i.

The second challenge is segmentation noise. In general, an unsupervised learner has no lexical resource to distinguish between lookalikes, i.e., words that happen to end in one of the paradigm's members, and lexemes that are truly P-conjugated. For example, a word such as 'bilen' looks just as much *'bile-n' as the correct 'bil-en'. Segmentation ignorance will obviously introduce a lot of spurious stem-affix pairs into consideration.

4.2 Paradigm Induction Techniques

Binary Vector Similarity Measures

It is instructive to look at the most straightfoward approach to the problem at hand. Consider the co-occurrence matrix $M : S_W \times B_W$ engendered by the set of words W in some input corpus.

$$M_{xy} = 1 \text{ iff } xy \in W \text{ (and } M_{xy} = 0 \text{ otherwise)}$$

For example, such a matrix could begin as in Table 13.

The rows in the matrix make up binary vectors in the $|B_W|$-dimensional stem space. Thus one potential line of attack is to characterize each suffix by its stem-occurrence binary vector. This is especially attractive given the long history and

	play-	drink-	s-	...
-ing	1	1	1	
-ed	1	0	0	
-s	0	1	0	
-axophone	0	0	1	
...				

Table 13. Contrived example fragment of a plausible co-occurrence matrix.

(x, y)	$Sim_1(x, y)$	$Sim_2(x, y)$
(s, t)	0.9052	0.0005
(ing, ng)	0.9608	0.0000
(ation, ed)	0.9678	0.0004
(ed, ing)	0.9708	0.0107
(ed, nrighteousness)	0.9803	0.0000
(ebuchadnezzar, ing)	0.9807	0.0000
(ebuchadnezzar, nrighteousness)	0.9999	0.0000

Table 14. Sample similarity values using Hamming distance, computed from an English bible corpus (King James 1977). $Sim_1(x, y)$ counts the number of matches between x and y:s vectors and $Sim_2(x, y)$ counts only the number of positive matches only. Both are normalized to $|B_W|$.

success of binary vector similarity approaches for other pattern classification problems (see Hyuk-Cha et al. (2005) for one of many surveys). One obvious approach to the problem at hand would then be to do k-means clustering on the set of suffixes using a vector distance such as Hamming distance.

Let us first consider Hamming distances. As explained in Hyuk-Cha et al. (2005), for other problems, the outstanding issue has been whether to count negative matches or not (or, alternatively to use supervised learning to train corresponding weights). Table 14 exemplifies the Hamming distance between a selection of suffixes from an English bible corpus (King James 1977). $Sim_1(x, y)$ counts the number of matches between x and y:s vectors. $Sim_2(x, y)$ counts only the number of positive matches, i.e., matching occurrences of 1. Both are normalized to $|B_W|$.

Clearly, in our situation, counting negative matches will favour matches between suffixes that are very uncommon (especially since we do not here consider the matter of well-segmentedness to counteract). On the other hand, counting only positive matches will favour matches between suffixes that are common, e.g., two one-character suffixes, even if they do not actually contrast in the language. The only solution seems to be to use proportions, but this does not without effort result in a distance metric. Thus in the next section we introduce a metric which has a lot in common in Mutual Information. Various tf-idf-related (Baeza-Yates and Ribeiro-Neto 1997) measures could also be considered.

Crucially, k-means clustering on most simple vector distances yields very

poor results. What tends to happen is that two suffixes of some true P form a tight pair and won't let the other members of P in, which are instead sucked into large clusters of junk. Therefore, except for languages where all paradigms are of cardinality 2 (if such languages exist), we have less faith in developing an approach based on strictly binary distances.

4.3 Formalizing Same Stem Co-Occurrence

As explained above, from the word distributions characteristic of natural language corpora, it is not straightforward to come up with a measure of how much a set of suffixes show up on the "same stems" that is not such that it favours the inclusion of any simply frequent, rather than truly contrasting, terminal segment. The measure we present here is valid for an arbritrary set of suffixes, including pairs (size two sets) but not restricted to pairs.

First, for each suffix x, define its quotient function $H_x(y) : S_W \rightarrow [0, 1]$ as:

$$H_x(y) = \frac{|Stems(x) \cap Stems(y)\}|}{|Stems(x)|} \tag{2}$$

where $Stems(x) = \{z|zx \in W\}$. The formula is conveying the following. We are given a suffix x, and we want to construct a quotient function which is a function from any other suffix to a score between 0 and 1. The score is calculated as: look at all the stems of x, other suffixes y will undoubtedly also occur on some of these stems. For each other suffix y, find the proportion of x:s stems on which y also appears. This proportion will be the quotient associated with y. Two examples of quotient functions (sorted on highest value) are given in Table 15.

Now, given a set of affixes P, construct a rank by summing the quotient functions of the members of P:

$$V_P(y) = \sum_{x \neq y \in P} H_x(y) \tag{3}$$

The $x \neq y$ is just there so that the y:s that are also in P do not get an "extra" 1.0, since $H_x(x) = 1.0$ for any data. The rank is just y sorted on highest $V_P(y)$.

As an example, take W from the Swedish PAROLE-Corpus (Borin 1997). We can compare in Table 16 the very common paradigm $\{a, an, as, ans, or, orna, ors, ornas\}$ with the nonsense paradigm $\{ungen, ig, ar, ts, s, de, ende, er\}$ consisting only of individually frequent suffixes. In Table 16, the ranks of the members of P to the left are [0, 1, 2, 4, 6, 8, 22, 31], and for P to the right, the ranks are [115044, 127, 17, 28, 4, 10, 100236, 14].

Now, if we can generalize from these cases it seems that we can rank different hypotheses of paradigms (of the same size) by looking at their quotient ranks. If the members of P "turn up high in" the quotient rank then the members of P tend to turn up on the same stems. There are several issues in formalizing the notion of "turn up high in". The places in the ranked list alone? Also incorporate the scores? Average place or total sum of places? For now we will

y	$H_{ing}(y)$	y	$H_{ed}(y)$
ing	1.00	*ed*	1.00
ed	0.59	*ing*	0.42
"	0.41	*"*	0.33
s	0.25	*e*	0.21
e	0.24	*s*	0.20
es	0.19	*es*	0.17
er	0.12	*er*	0.08
ers	0.10	*ion*	0.07
ion	0.07	*ers*	0.05
y	0.05	*y*	0.04
ings	0.05	*ions*	0.03
ions	0.03	*ation*	0.03
in	0.03	*able*	0.02
ation	0.03	*ings*	0.02
's	0.03	*'s*	0.02
ingly	0.03	*or*	0.02
or	0.02	*in*	0.01
able	0.02	*ly*	0.01
ive	0.02	*ive*	0.01
ors	0.02	*ingly*	0.01
ations	0.01	*al*	0.01
er's	0.01	*ment*	0.01
ment	0.01	*ors*	0.01
ly	0.01	*ations*	0.01
...	...	...	...

Table 15. Sample quotient functions/lists for *ing* and *ed* on the Brown Corpus. H_{ing} and H_{ed} have 68337 and 75853 nonzero values respectively.

y	$V_P(y)$	y	$V_P(y)$
a	3.93	"	3.32
an	2.82	t	1.48
or	2.71	a	1.19
"	1.91	r	1.18
orna	1.76	**s**	1.15
ar	1.13	en	1.14
as	1.06	iga	0.86
ade	1.05	d	0.80
ans	0.94	igt	0.73
at	0.89	as	0.66
en	0.82	**de**	0.59
s	0.76	des	0.57
t	0.73	ade	0.55
e	0.71	ung	0.49
er	0.66	**er**	0.49
ad	0.61	at	0.48
ande	0.52	n	0.46
ades	0.47	**ar**	0.45
ats	0.40	an	0.44
i	0.36	e	0.42
...	...	...	...
ors	0.35		
...	...		
ornas	0.27		
...	...		

Table 16. Example ranks for $P = \{a, an, as, ans, or, orna, ors, ornas\}$ (left) and $P = \{ungen, ig, ar, ts, s, de, ende, er\}$ (right).

P	$VI(P)$	P	$VI(P)$
{'ation'}	0.00	{'xt'}	0.00
{'ated', 'ation'}	0.14	{'xt', 'n'}	0.04
{'ate', 'ated', 'ation'}	0.40	{'xt', 'n', 'ns'}	0.12
{'ate', 'ated', 'ating', 'ation'}	0.75	{'n', 'ns'}	0.55
{'ate', 'ated', 'ating', 'ation', 'ations'}	1.00	...	...

Table 17. Example iterations of $G^*('ation')$ and $G^*('xt')$.

just do a simple sum of places in the ranked list, divide by the optimum sum (which depends on $|P|$ and is $0+\ldots+|P|-1$), and take the inverse. This gives a score between 0 and 1 where a high score means the members of P tend to appear on the same stems:

$$VI(P) = \frac{|P|(|P|-1)}{2\sum_{x\in P} place(x, V_P)} \tag{4}$$

The VI-score from the last section may be used for a greedy hill-climbing search through the affix set space. For example, we may start with an affix, a one member set, and see whether we can improve the affix score by including another member, and perhaps another after that until we cannot improve the score anymore. One such ever-expanding search through the affix set space is bounded by the total number of affixes and is thus polynominal in the number of affixes. A bolder alternative is to also entertain the possibility of kicking some member out if that improves the score. Then, in the worst case, we may have to step through the whole affix set space before convergence. In practice, however, searches are not at all close to exploring large parts of the affix set space before converging, so allow expulsions is not prohibitive. Formally, define the growing function of a set P of affixes as:

$$G(P) = argmax_{p\in\{P\}\cup\{P\cup\{s\}|s\in S_W\}\cup\{P\setminus\{s\}|s\in S_W\}} VI(p) \tag{5}$$

$$G^*(P) = \begin{cases} P & \text{if } G(P) = P \\ G^*(G(P)) & \text{if } G(P) \neq P \end{cases} \tag{6}$$

Two growth-examples are shown in Table 17, one which attains a perfect 1.0 score and one in which the original member is expelled in a later iteration.

4.4 Discussion

Impressionistically, the growing algorithm makes sense. Where one as a linguist has sure intentions, the score tends to agree. For example, $G^*(ing) = \{'', e, ed, es, ing, s\}$, and $G^*(ied) = \{ied, ies, y, ying\}$. However, the $VI(P)$-score knows nothing about well-segmentatedness, so, e.g., $VI(\{'xcellent', 'xcellently'\}) = 1.0$. It is clear that scores for well-segmentedness could come from a different component, such as the affix extraction algorithm in Section 3, and, indeed, such combinations are variously explored in the applications sections of the present

s_1-	s_2-	s_3-	s_4-	s_5-	s_6-	s_7-	s_8-	s_9-
s_1a	s_2a	s_3a	s_4b	s_5b	s_6b	s_7c	s_8c	s_9c
s_1x	s_2x	s_3x	s_4y	s_5y	s_6y	s_7z	s_8z	s_9z

s_{10}-	s_{11}-	s_{12}-
$s_{10}a$	$s_{11}a$	$s_{12}d$
$s_{10}b$	$s_{11}b$	$s_{12}b$
$s_{10}c$	$s_{11}c$	$s_{12}c$

Table 18. Definition of a set W ($|W| = 27$) of strings arranged according to prefix for readability.

chapter. However, we have not yet conceived a final "extraction"-phase to extract all the salient paradigms (or the like), wherefore there is no evaluation simpliciter against some gold standard of paradigms.

Since paradigms tend to be small, growing from a single paradigm is a better heuristic than shrinking from the whole affix set space. However, it still happens frequently for large $|P|$ (10-20) that the growing sticks in a local maximum. And for languages, e.g., Turkish, with paradigms that number hundreds of members (because they are really combinatorial in nature) the approach is not sufficient alone.

The paradigm-growing approach, as opposed to many fixed-cluster approaches, has the positive side-effect that one affix can readily be the member of several locally maximal paradigms. Though for each paradigm, there needs to be at least one starting-affix that grows towards it.

In practice, growing paradigms take a long time because of sorting and summing with very long lists (typically 100000-ish items).

It is important to note the set-oriented (rather than pair-oriented) nature of this measure. Let us say x is at the top of y:s quotient list H_y (except for y itself) and y is at the top of H_x (except for x self). Even so, $VI(\{x, y\})$ might get a smaller score than $VI(\{x, y, z\})$ with the inclusion of another member z. This depends, of course, but crucially z may get a lot of "points" both from $H_x(z)$ and $H_y(z)$ whereas x doesn't get points from H_x nor y from H_y.

The $VI(P)$-measure cannot be reduced to a monotone function of the set of $VI(P)$-values for the set of suffix-pairs in P as the following counterexample shows. (Unfortunately I do not know of a smaller counterexample.). Let W be the strings in the cells of the tables in 18, arranged according to stems s_i for readability.

The set of W so defined yields the quotient lists for the suffixes a, b, c and d as shown in Table 19.

Now $VI(\{a, b, c\}) = 1.00$ since a, b, c get [0.66, 0.90, 0.90] respectively in the summed list V_P, which is more than any other suffix. Further, the $VI(\{x, y\})$-scores for each pair (x, y) in set $\{a, b, c\}$ are:
$VI(\{a, b\}) = 0.14$
$VI(\{a, c\}) = 0.14$

H_a		H_b		H_c		H_d	
a	1.00	b	1.00	c	1.00	d	1.00
x	0.60	c	0.50	b	0.50	b	1.00
b	0.40	y	0.50	z	0.50	c	1.00
c	0.40	a	0.33	a	0.33	...	0.00
...	0.00	...	0.00	...	0.00		

Table 19. Given W as in Table 18, quotient lists for a, b, c and d.

$VI(\{b,c\}) = 0.14$

On the other hand, $VI(\{b,c,d\}) = 0.5$ since the places for b, c, d are [0, 1, 5] and the optimal sum is 3 (= [0, 1, 2]). But the $VI(\{x,y\})$-scores for each pair (x,y) in set $\{b,c,d\}$ are higher than in the previous case:
$VI(\{b,c\}) = 0.14$
$VI(\{b,d\}) = 0.20$
$VI(\{c,d\}) = 0.20$

This shows that the pair VI-scores for $\{b,c,d\}$ are higher than for $\{a,b,c\}$, yet $VI(\{a,b,c\}) > VI(\{b,c,d\})$.

4.5 Conclusion

For three reasons, finding morphological paradigms from raw text data is an exceedingly difficult problem. First, paradigms – being sets of affixes – are exponentially many in the number of affixes. Second, almost all stems occur in one or only a few of the forms of their paradigm, making the evidence slight. Third, since raw text data does not mark the segmentation boundary between the true stem and its affix, all possibilities must be consídered the same, introducing noise in the stem-affix co-occurrence analysis. We present a metric that takes a paradigm and gives a score of "how good" the paradigm is, i.e., how much the members tend to occur on the same stem and how much non-members do not occur on those stems. The metric has the novelty that it is not the aggregation of a pair-wise defined metric. Using the metric there is a natural greedy search algorithm that finds locally optimal paradigms, which also takes advantages of the domain specific actuality that paradigms tend to be small. It is also well-behaved with respect to event that one affix may belong to several different paradigms. Impressionistically, the metric produces sensible results, at least on languages without combinatorial paradigms, but there remain many issues to discuss. It also remains to work out how to make use of the metric to extract a fix list of paradigms and thereafter to evaluate.

5 Application 1: A Fine-Grained Model for Language Identification

Abstract

Existing state-of-the-art techniques to identify the language of a written text most often use a 3-gram frequency table as basis for 'fingerprinting' a language. While this approach performs very well in practice (99%-ish accuracy) if the text to be classified is of size, say, 100 characters or more, it cannot be used reliably to classify even shorter input, nor can it detect if the input is a concatenation of text from several languages. The present paper describes a more fine-grained model which aims at reliable classification of input as short as one word. It is heavier than the classic classifiers in that it stores a large frequency dictionary as well as an affix table, but with significant gains in elegance since the classifier is entirely unsupervised. Classifying a short input query in multilingual information retrieval is the target application for which the method was developed, but also tools such as spell-checkers will benefit from recognising occasional interspersed foreign words. It is also acknowledged that a lot of practical applications do not need this fine level of granularity, and thus remain largely unbenefited by the new model. Not having access to real-world multi-lingual query data, we evaluate rigorously, using a 32-language parallel bible corpus, that accuracy is competitive on short input as well as multi-lingual input, and not only for a set of European languages with similar morphological typology.

5.1 Introduction

The language identification problem is to decide for a natural language text which language it is written in. The usual setting is to assume that one has access to training corpora beforehand for the languages to be considered. Some language fingerprint model is built from the training corpora and then classification of unseen text (belonging to one of the languages at hand) is performed through this model.

Existing state-of-the-art techniques rely on a surprisingly simple model, namely, a frequency table of character 3-grams for each language, read off directly from the training corpora. The corresponding 3-gram frequency table for the text to be classified is then compared to each stored language by some rank-frequency metric. In practice, this approach performs very well (99%-ish accuracy) if the text to be classified is of size, say, 100 characters or more (Juola 2006). Thus the language identification problem is a solved problem for most practical applications.

However, the crude 3-character gram method has a certain drawback (which may or may not be practical problem), in that it is not monotone. That is, if two texts s_1, s_2 are classified as l_1, l_2 respectively, then it is not certain that the concatenation of s_1 and s_2 is classified as either l_1 or l_2.

We will present an alternative model which aims at reliable classification of new text as short as one word. This model combines a frequency dictionary from

each training corpus and a component that tries to recognize completely unseen words by looking at affixes (which would, e.g., identify a word like *jihading* 'fighting the jihad' correctly as English). This latter component is crucial, not only for languages which make more use of affixes than English, but because there will always pop up completely novel words for any natural language no matter what size the training data. The affix detection technique implemented also builds from the same training corpora and requires no extra supervision or work by a human.

There are certainly practical applications which do require reliable classification of small segments and autodetection of language switches. These include spell checkers that wish to disregard interspersed foreign words, text-to-speech systems that make intermediate use of grapheme-to-phoneme conversion likewise wish to indentify interspersed foreign words, and multilingual information retrieval systems would benefit from knowing the language(s) of the words of a short query. For a lot of other practical applications, the granularity of the proposed new model is superfluous. For these applications, the only advantage of the proposed model is elegance and absolute lack of training supervision.

The resultant language identifier is evaluated using bible corpora for 32 languages, spanning the full range of morphological typology of languages of the world (Dryer 2005). Both its ability to classify short segments into one language and to autodetect short segments that may be composed of several languages, are evaluated. However, we do not compare these figures to existing systems, because they were not designed for classifying short segments accurately (and thus perform very poorly)[4]. On longer segments, i.e., 100 characters, performance is near perfect, and it is presumed that the state-of-the-art systems would also perform near perfect if tested on the same set.

With the improved accuracy on short segments and wide typological testing range, we hope to have met the challenges for written language identification set out in a recent survey article by Hughes et al. (2006).

All the training corpora used in this paper are bible corpora, since they are the only sufficiently large corpora available for a reasonably varied set of languages.

5.2 Previous Work

Our full bibliography of works dealing narrowly with written language identification spans over 100 articles, a handful of technical reports and one PhD thesis (Ziegler 1991) – it is therefore not possible to review them all here. Many pointers to older work and language identification of speech signals are given in Muthusamy and Spitz (1997) and Caseiro (1999). Sibun and Reynar (1996) is an excellent review and comparison of techniques used in early work.

[4] There would also have been practical problems in doing justice as many descriptions of existing systems hide information on parameter tweaking. Online systems we have found do not allow uploading the training/test set we use, which is crucial in order to assess language-dependentness.

For the language identification problem in the setting as in this paper, namely, written language identification trained on reference language data, two different feature models have been prevalent. One that looks at common words and one based on character n-grams (Grefenstette 1995, Cavnar and Trenkle 1994, Damashek 1995, Dunning 1994) and see Martin et al. (2006), Kruengkrai et al. (2005) for refinements of the n. The classification can then be done by comparing input text features to reference language features using rank-order statistics. More recent work in this direction has aimed at trimming overweight feature models (Poutsma 2002, Takci and Sogukpinar 2004) or at combining n-gram and whole word features (Prager 2000). See, however Biemann and Teresniak (2005) for a novel, completely different approach based on words clustered on sentence-co-occurrence. (The accuracy of this identifier is comparable to the older approaches, but it is not, as claimed therein, unsupervised, because there is a very large number of manually set parametres/thresholds and word-frequency statistics are gathered from curated corpora.) There is also more recent work targeting web pages specifically (Xafopoulos et al. 2004, Martins and Silva 2005, Lins and Gonçalves 2004), that address the proper treatment of HTML tags.

Whereas the language identification problem has variously been labelled 'easy' and 'solved' (McNamee 2005), it depends on whether one sets the goal higher than distinguishing non-minimal noise-free samples of European languages. Some recent articles (Murthy and Kumar 2006, da Silva and Lopes 2006b,a) identify practical problems where this is not so. For instance, as far as we can ascertain, the best systems in van Noord's Online Summary[5] minimally require some 20 characters of text to make a judgment at all. Nor are they capable of realizing that a sample text is a concatenation of two languages. For example, The Xerox MLTT Language Identifier[6] classifies the sentence 'good fish prefer their snake' correctly as English, the sentence 'fina fiskar sprattlar inte ofta' correctly as Swedish, but the concatenation of the two is classified as Norwegian (even though there is actually no legal Norwegian word in either sentence).

As indicated already, the present method seeks to tackle also smaller sample texts, which is crucial in order to be able to track whether a text is a composition of words from several languages. While the classic n-gram approaches have found that a good $n = 3$, i.e., that salient morphemes can be approximated as being exactly 3 characters, a more elegant alternative is to hold this variable, so that salient affixes can have any length in any language. Furthermore, we wish to extend the testing scope, as present published testing has been only on a rather small set of European languages.

[5] `http://odur.let.rug.nl/~vannoord/TextCat/competitors.html` accessed the 25th of May 2005.

[6] `http://www.xrce.xerox.com/competencies/content-analysis/tools/guesser` accessed 20 Jan 2007.

5.3 Definitions and Preliminaries

Start with a finite non-empty alphabet Σ. The following terminology and notation will be used.

word: a non-empty finite string over Σ. Thus the set of all possible words can be denoted Σ^+. Lowercase w with subscripts will be used for variables over words. A word will be enclosed in quotes if confusion could arise otherwise.

sentence: a finite non-empty tuple of words $\langle w_1, w_2, \ldots, w_n \rangle$. Commas and brackets will be omitted when no confusion can arise. However, variables that range over tuples, e.g. $\langle l \rangle$, will always be written with brackets.

S_Σ : let $S_\Sigma = \{\langle w_1 w_2 \ldots w_n \rangle | w_i \in \Sigma^+, n \in \mathbf{N}\}$ denote the set of all possible sentences.

language: a probability distribution over sentences $L : S_\Sigma \to [0, 1]$, i.e., such that $\sum_{\langle s \rangle} L(s) = 1$.

training corpus: a finite sequence of sentences. However, we will never make use of the order of sentences, or order or words in the sentences, so a training corpus may be equated with its bag of words. Thus, if T is a training corpus, let $f_T(w)$ denote the frequency of the word w in T. Also, use $W_T = \{w | f_T(w) \geq 1\}$ for the *set* of words in the training corpus.

names and variables: Unless we are talking about existing natural languages, e.g., English, natural numbers $1, 2, \ldots$ will be used for language names. $\Sigma_1, \Sigma_2, \ldots$ will be used for their corresponding alphabets, with $\Sigma = \bigcup_i \Sigma_i$ for the mother alphabet. $L_1, L_2, \ldots$ will be used for languages, i.e., probability distributions, and coindexed $T_1, T_2, \ldots$ for training corpora (where T_i is assumed to be sampled from L_i).

The idea is of course that sentences which are illegal or ill-formed in some natural language will have zero probability and legal sentences will have a non-zero probability corresponding to their relative frequency. A natural way to see how a natural language should correspond to such a formal probabilistic language is to consider ever increasing amounts of natural language text and let the probability of each sentence be its limiting relative frequency. This correspondence requires that this limit actually exists for all sentences. If there are natural languages that do not live up to this, or which cannot be modelled so with an acceptable level of discrepancy, they should not be thought of as languages in our terminology.

Our notion of language is a generalization of the more common formalization of natural language as a *set* of sentences. We actually need this greater flexibility in order for language identifiers to exploit the fact that some words (and thus some sentences) which are legal in several natural languages may be distinguished by their different levels of frequency. It also provides a framework for gracious treatment of new words and proper names which are so ubiquitous

in open domain natural language text (such as newspaper text) that they should not be "abstracted away". With the probability model we have the power to say that any word is possible in any language, for example as a proper name, but it is more probable that an instance of, e.g., 'the' is from English than in some other language where it may have occurred as a proper name.

5.4 A Fine-Grained Model of Language Identification

From the input of a training corpus, the proposed model characterizes a language using the following two components:

Frequency dictionary: Stores each seen word and its (relative) frequency. The frequency of seen words is a very powerful predictor of a language.

Unsupervised affix detection: Salient affixes are extracted (in an unsupervised manner), which form the basis for a probabilistic guessing of previously unseen words.

These two components are combined into a *word emission probability* distribution that aims to predict how likely a language is to have emitted a given word. In principle, a collection of such probability distributions are sufficient to make up a standard case of language identifier that always outputs exactly one language. However, we shall also use another component, a *language holdback bias*, to enable intuitively correct identification of text that is concatenated from several languages.

Word Emission Probability

A frequency dictionary FD_l is built simply as:

$$FD_l(w) = \frac{f_{T_l}(w)}{\sum_{w' \in \Sigma} f_{T_l}(w')}$$

Following Hammarström (2006a) – recounted in Section 3 – we use an unsupervised algorithm to gather information on the salient affixes for a given language. The algorithm uses W_l as its input and outputs a probability distribution on character strings that aims to say whether a given segment is likely to be a characteristic prefix or suffix for the language at hand. To be more precise, the probability distribution aims to capture the notion of morpheme probability that one arrives at if: 1. A linguist does a morphemic segmentation of the word types (not words tokens) occurring in a corpus, 2. The frequencies of the individual morphemes, in prefix or suffix position, are interpreted as probabilities. For example, *-qvj* would likely get zero probability in an English corpus. An example output, adapted from Hammarström (2006a), is given in Table 20, sorted on highest probability. The outcome of the algorithm for languages which do not have any morphology at all is a fairly even spread of probability mass over initial and final characters of the words of the language in question. For reasons

Swedish		English		Swahili	
för-	0.097	*-ed*	0.132	*-a*	0.100
-en	0.086	*-eth*	0.109	*wa-*	0.095
-na	0.036	*-iah*	0.099	*ali-*	0.065
-ade	0.035	*-ly*	0.090	*nita-*	0.059
-a	0.034	*-ings*	0.068	*aka-*	0.049
-ar	0.033	*-ing*	0.062	*ni-*	0.046
-er	0.033	*-ity*	0.059	*ku-*	0.044
-as	0.032	*-edst*	0.058	*ata-*	0.042
-s	0.031	*-ites*	0.046	*ha-*	0.032
-de	0.031	*-s'*	0.036	*a-*	0.031
...	...	...	...	...	...

Table 20. Comparative figures for prefix vs. suffix detection for three sample languages.

of space, the reader is referred to the said paper for a discussion of the inner workings and alternative algorithms.

As mentioned, the output from the affix extraction is a probability distribution over affixes. What we need is a probability distribution over words, in which any word ending in some salient suffix should have nonzero probability. One quite reasonable way to achieve this is to assign geometrically decreasing probabilities for longer and longer words. Thinking in this way, we let all observed (in W_l) word lengths get the probability mass proportional to the number of observed words with such lengths, and unseen word lengths get geometrically decreasing probability. Thus, to get a well-defined probability distribution over words based on the affix probability distribution, we multiply together the word-length mass for w with the highest (not necessarily longest!) matching, if any, affix probability, for a given word w. The details aren't interesting, but use $A_l(w)$ to denote the just described affix-based probability distribution.

Putting the affix detection together with the frequency dictionary to make an emission probability involves a related kind of estimatate. How much probability mass should be assigned to seen vs. unseen words? There are probably many similar alternatives, but here we have simply guessed that unseen words are like hapax words, and assigned the probability mass proportions to be like the proportion of hapax words: $\alpha_l = \frac{|\{w \in W_l \mid f_{T_l}(w)=1\}|}{|W_l|}$.

We are now ready to define emission probability:

$$P_l(w) = \begin{cases} (1-\alpha_l) \cdot FD_l(w) & \text{if } w \in W_l \\ \alpha_l \cdot A_l(w) & \text{if } w \notin W_l \end{cases}$$

It can happen that there is more mass given to an unseen word than to a (rare) seen word, even within one particular language. In fact, proportions vary quite wildly between languages, as can be seen in Table 21 with figures computed on the translations of the same bible text.

Language	$\|T\|$	$\|W\|$	α	$argmax_w(FD(w))$	
Greenlandic	382188	107918	0.706	*taava* (then)	0.00857
Swedish	758773	26825	0.407	*och* (and)	0.05566
Haitian creole	904915	7796	0.335	*yo* (PL/they)	0.05531

Table 21. Some indications as to the widely differing identification cues for three languages; the polysynthetic Greenlandic versus the almost isolating Haitian creole.

Language Holdback Bias

If we have $L_1, \ldots, L_n$ languages, the previous subsection shows how to construct the corresponding $P_1, \ldots, P_n$ probability distributions over words. Next, we shall define a family of probability measures over *sequences of words.* There will be one probability distribution for each language tuple of the same length as the sequence to be measured:

$$P_{l_1 l_2 \ldots l_m}(w_1 w_2 \ldots w_m) = \prod_i P_{l_i}(w_i)$$

Given a sequence of words we could then naively decide which language(s) it most probably belonged to by listing each tuple of the appropriate length and computing which tuple has the highest probability of having generated the sequence of words. However, for several reasons, such an approach is not advisable. First, with n languages there are n^m language tuples so it would not be tractable to enumerate them all. Second, the probability measures so defined, the output will be the concatenation of the most probable language for each word individually. This is probably not what we want since many words that are legal in several languages differ in frequency. Consider a sequence of a million words indisputably belonging to language L_1, and, interspersed inside, a word that is legal in both L_1 and L_2 but slightly more common in L_2. The naive language identifier would yield L_2 disregarding the suggestive surrounding million words of L_1. While it is technically not impossible that it is a concatenation of the two languages, a human would never see it as that. Third, it is not clear how to see if an input sequence is non-trivially legal in more than one way (i.e., there are several satisfactory language tuples). Either we insert some kind of threshold which would be hard to know how to set, or we have to say that pretty much all tuples are satisfactory identification of the sequence only with some degree variation.

For the first problem, it is easy to see that not all tuples need to be enumerated to get the maximally probable one (if we want only this one, rather than the probabilities for all). As defined, the emission probabilities depend only on a particular word, not anything else in the sequence, so maximas can be computed locally in the sequence and glued together as in any standard application of dynamic programming. For the second and third problem, we shall propose a refinement of the strategy that obviates the need for any thresholds.

We propose that a machine language identifier like ours should have a *bias* towards minimizing the number of times we change languages in an identification sequence. To be more precise, the prior probability that a sequence should switch language c times should decrease exponentially in c. Also, other things being equal, the longer the sequence the stronger the bias should be, i.e., it should be less likely that a million word sequence should switch language once somewhere within it, than that a two-word sequence should switch language (once) within it. This is the way to say that having seen a million words of language L_1 counts for more than having seen just one word of L_1. We do not see any basis for this to be a sequential property, e.g., that language switches are significantly more (or less) likely after or before certain words, wherefore a (H)MM-modeling technique offers no advantage.

Formally, let $C(l_1 l_2 \ldots l_m) = |\{i | l_i \neq l_{i+1}\}|$ denote the number of times a change in language occurs in a language sequence. Clearly, we have $0 \leq c \leq m-1$. Let $\langle l \rangle = l_1 l_2 \ldots l_m$ be an arbitrary language tuple under consideration and $c = C(\langle l \rangle)$ its number of switches. Now, for any language identifier parametrized on c and m, we wish the bias, regardless of the particular languages at hand, to ensure that:

$$\frac{P(c,m)}{P(c+k,m)} \geq 2^k \qquad \text{for all } k \geq 0, m$$

$$P(c,m) > P(c,m+k) \qquad \text{for all } k \geq 1, c$$

A simple fulfilment of these is the following **Language Holdback Bias** function $B(c,m)$:

$$B(c,m) = \frac{1}{m^c} \cdot \frac{1}{\sum_{0 \leq i \leq m-1} \frac{1}{m^i}}$$

There of course alternative bias functions that also fulfill the desiderata, but this is the simplest one. Now, with the bias function defined we are ready to present our full definition of the output of the now rather sophisticated language identifier.

$$ID(w_1 \ldots w_m) = \begin{array}{l} \text{the set of all tuples } \langle l \rangle = l_1 \ldots l_m \\ \text{such that for all } \langle l' \rangle \\ B(C(\langle l \rangle), m) \cdot P_{\langle l \rangle}(w_1 \ldots w_m) \geq \\ B(C(\langle l' \rangle), m) \cdot P_{\langle l' \rangle}(w_1 \ldots w_m) \end{array}$$

The formula conveys the following: look for tuples with as few cuts (i.e., minimal c) as possible, that are such that they have higher probability, the bias respected, than any other tuple with *more* cuts. This is the key feature which eliminates the need for a threshold. Thus, for example, a word sequence will be said to be of language L_l iff it has higher probability than any division of the sequence into two parts of different languages (or three parts etc). There may be several such languages, but hardly all, so the yield will be a strong prediction.

The following more procedural reformulation of the identification function may be easier to understand. It should also make it clear that language identification is still polynomial in the sequence length, since there are still no dependencies between the word-probabilities.

1. Find minimal c such that there exists a tuple $\langle l \rangle$ with $C(\langle l \rangle) = c$ and:

$$B(c, m) \cdot P_{\langle l \rangle}(w_1 \ldots w_m) \geq$$
$$B(C(\langle l' \rangle), m) \cdot P_{\langle l' \rangle}(w_1 \ldots w_m)$$
$$\text{for all } \langle l' \rangle \text{ with } C(\langle l \rangle) > c$$

2. Output all tuples $\langle l \rangle$ with $C(\langle l \rangle) = c$ and:

$$B(c, m) \cdot P_{\langle l \rangle}(w_1 \ldots w_m) \geq$$
$$B(C(\langle l' \rangle), m) \cdot P_{\langle l' \rangle}(w_1 \ldots w_m)$$
$$\text{for all } \langle l' \rangle \text{ with } C(\langle l \rangle) > c$$

5.5 Examples

Example 1: The kings hon walikusoma

Consider the sequence *the kings hon walikusoma* which consists of *the*, which is of course the English definite article; *kings* is the well-known English lexical item which does occur in the training corpus – it also happens to end in *-s* which is a very common Swedish inflectional ending (but there is no lexical item 'king' or 'kings' in Swedish); *hon* is a Swedish personal pronoun, abundantly occurring in the Swedish training corpus; and *walikusoma* is a well formed Swahili word whose individual morphemes all individually occur abundantly in the Swahili training corpus – but the semantically odd well-formed word 'walikusoma' does not occur in the training corpus (it would mean 'they read you').

The individual word-probabilities as well as a selection of the more interesting tuple-probabilities for the sequence as a whole, are shown in Table 22. As can be seen, the $P_{eng,eng,swe,swa}$ value beats all tuples with zero or one switches. It also happens to beat all tuples with three switches and it is the only such tuple. Therefore, in this case, the output will be exactly English, English, Swedish, Swahili.

Example 2: The kings are there

The complicated interaction seen in the previous example does not disturb the "normal" easy class of classifications. Table 23 shows the word-probabilities for the almost trivial sentence *the kings are there*. There is a certain zero-switch tuple which is way ahead of the others. As it also beats all one-switch tuples (and no other zero-switch tuple does), it will be the output of the identifier.

Example 3: De la

There are instances where there are several "winning" tuples, though informal tests show that this is not achieved very often. The sequence *de la* is very

	'the'	'kings'	'hon'	'walikusoma'
English	0.051522	0.000286	0.000003	0.000004
Swedish	0.000002	0.000040	0.000916	0.000043
Swahili	0.000218	0.000000	0.000000	0.000317

All one-language tuples

$P_{eng,eng,eng,eng}$	1.350e-016
$P_{swe,swe,swe,swe}$	2.468e-018
$P_{swa,swa,swa,swa}$	1.878e-025

Some top one-switch tuples

$P_{eng,swe,swe,swe}$	2.034e-014
$P_{eng,eng,swe,swe}$	1.465e-013
$P_{eng,eng,eng,swa}$	3.008e-015

The top two-switch tuple

$P_{eng,eng,swe,swa}$	2.701e-013

Table 22. Example 1: $P_l(w)$ for a set of languages and some interesting words, followed by a selection of the more interesting tuple-probabilities.

	'the'	'kings'	'are'	'there'
English	0.051522	0.000286	0.002812	0.002065
Swedish	0.000002	0.000040	0.000006	0.000035
Swahili	0.000218	0.000000	0.000004	0.000006

$P_{eng,eng,eng,eng}$	8.5467629403443202e-011
$P_{swe,swe,swe,swe}$	1.2961894211016589e-020
$P_{swa,swa,swa,swa}$	2.5363460513704776e-023
...	

Table 23. Example 2: $P_l(w)$ for a set of languages and some words that are very easy to classify, followed by examples to indicate that the dominance of a certain zero-switch tuple over some others.

	'de'	'la'		
French	0.029172	0.016325	$P_{fre,fre}$	0.0003174886
English	0.000000	0.000000	$P_{spa,spa}$	0.0003227756
Swedish	0.008400	0.000001	$P_{spa,fre}$	0.0001844997
Swahili	0.000000	0.001517	...	...
Spanish	0.033905	0.014280		

Table 24. Example 3: $P_l(w)$ for a set of languages and two words, followed by a selection of the more interesting tuple-probabilities.

common to both Spanish and French. In English it is not common at all. In Swedish *de* is a personal pronoun so it enjoys a certain frequency, whereas *la* is not a word in (bible) Swedish. Similarly, *la* is a negator in Swahili and is therefore fairly frequent. Table 24 shows the relevant probabilities. The output will be only the tuples spa, spa and fre, fre, because tuples such as swe, swa and spa, fre lose out because of the bias, favouring few switches.

5.6 Evaluation and Discussion

Three extensive tests were performed using a parallel corpus of the bible in 32 languages, which contains languages from the isolating Maori to the record holding polysynthetic Greenlandic (Dryer 2005). In order to get a sufficiently cross-language comparable evaluation, size and randomness were equalized between languages the following way. A random verse from each chapter was selected (there are 1209 chapters in the bible). This was done once for the whole language set. Of course, these verses were removed from the training data. A random word from each selected verse was selected. This word-selection was done separately for each language. For each language, we thus get a set of randomly selected words E_l. Though 1209 word-selections were made for each language, many selections happened to select the same word. Thus the size of the E_l-sets varied from 350 (for Maori) to 974 (for Greenlandic). The descrepancy is not disturbing. Words are not entities of the same kind across languages, but our classifier operates on the granularity of words, and the desiderata is an evaluation of 'accuracy per (randomly selected) word'. An alternative, e.g., selecting 1000 unique words of each language would have made interpretation of the result difficult, because for Maori, it is likely that most of the 1000 words would have been *seen* words, occurring in other verses, whereas the opposite is the case for Greenlandic.

If E is a set of tuples (possibly one-word tuples), drawn for language l, we define the accuracy $R_E(l)$ of a language identifier ID:

$$R_E(l) = \frac{|\{\langle x\rangle | ID(\langle x\rangle) = l \text{ and } \langle x\rangle \in E\}|}{|E|}$$

One-word classification: The R_{E_l} was calculated for each of the 32 languages. Since the input sequence is of length 1, there will never be any

cuts, so the language identifier was set to output the language with highest probability of having emitted the input word. The E_l-sets as defined above may contain words that are "impossible" predict where they were taken from, on the basis of the word alone. For example, let's say a word w is legal in two languages but much more common in l_1 than l_2. If it happened to be drawn from L_{l_2}, it is hard to see how this can be predicted. However, we computed figures on the possible influence of this issue, and it turned out to be minor. Therefore, the results in Table 25 stand, but could be adjusted upwards by very small percentages.

Verse classification: To check how accurate the identifier was on longer segments, we chose to test on segments of roughly the length of a verse. Verses, in fact, happen to be around 100 characters long on average. From the 1209 verses selected (as above), those 100 verses thereof whose number of characters were closest to the average verse length of that language, were selected for testing. Denoting these 100-verse sets by V_l, the verse-classification accuracy R_{V_l} was calculated. This score, as well as data on average verse length, can be seen in Table 25.

4-tuple multilingual classification: A set of 1000 mixed language 4-tuples were built from $E_1, \ldots, E_{32}$ as follows.

1. Pick a random language l and pick two random words from that E_l.
2. Precede it with a random word from a random language $E_{l'}$.
3. Add a random word from a random language $E_{l''}$ at the end.

The results of this test was 193 (**19.3%**) fully correctly identified 4-tuples and 204 (**20.4%**) with exactly one word misclassified.

Some figures are low, not surprisingly for languages with a lot of morphology, but overall we hold the results are very reasonable given the exceedingly difficult test problems of one-word and multi-language classification. It is very easy to make mistakes on single words when there are so many languages in the pool – the accuracies are much higher if the number of competing languages is halved.

Unfortunately, we cannot contrast the verse-test with figures from competing state-of-the-art systems, as none of the systems known to us give enough details (on thresholds and such) to reconstruct a fair version of the classifier.

A matter requiring further commentary is the use of a bias function to do the job a scalar threshold value does in related work. (Human language identifiers, having the ability to assess syntactic and semantic coherence, need not use either.) Conceptually, the bias function employed is nothing other than a complex system of thresholds, in terms of growth behaviour (exponential, linear etc.) rather than scalar values. Arguably, this is an elegance improvement, although it comes with the cost of being harder to understand, compute and analyse. Also, in the experiments reported above, the bias function approach experimentally outperforms a simple systems of scalar threshold values. For example, through supervised training we have tried tuning one single threshold

Language	1-word	Verse	$\overline{V}$
Haitian Creole	0.839	1.00	101.79
Zarma	0.781	1.00	99.45
Kekchi	0.720	1.00	148.78
English	0.678	1.00	104.19
Maori	0.665	1.00	107.73
Hindi	0.607	1.00	119.50
Hausa	0.605	1.00	94.10
Afrikaans	0.594	1.00	103.34
Danish	0.580	1.00	89.30
Cebuano	0.573	1.00	129.48
Icelandic	0.550	1.00	95.58
Swedish	0.547	1.00	107.20
Adamawa Fulfulde	0.539	1.00	96.57
German	0.533	1.00	103.52
Albanian	0.523	1.00	114.80
Spanish	0.511	1.00	95.83
French	0.507	1.00	101.83
Swahili	0.494	1.00	105.03
Slovene	0.488	1.00	100.12
Polish	0.487	1.00	144.52
Portuguese	0.481	1.00	98.41
Esperanto	0.473	1.00	97.80
Italian	0.473	1.00	116.80
Catalan	0.450	1.00	109.70
Dutch	0.415	1.00	109.36
Lithuanian	0.396	1.00	104.99
Hungarian	0.386	1.00	102.10
Latin	0.366	0.99	112.54
Turkish	0.348	0.95	93.43
Finnish	0.345	0.99	107.88
Malayalam	0.276	0.88	128.65
Greenlandic	0.222	0.87	126.99

Table 25. Accuracies for the one-word and verse tests as well as average verse length in characters ($\overline{V}$).

value for all experiments, one threshold value individually for each language, different threshold values for different classification tasks (i.e., one for multi-language classification and one for single language classification) and so on, resulting in generally lower accuracy on the same test set (obviously, there is little room for presenting and discussing figures from these tests here). Nevertheless, it remains possible that some other, yet undiscovered, system of scalar thresholds is superior to the bias function.

5.7 Conclusions

We have described a new model with considerable elegance for language identification on small, possibly mixed languages segments. We have also added significantly to the set of published evaluations of a language identification system with a balanced cross-language test. For larger input texts the new model has excellent accuracy, but it is bigger and slower in practice than the existing state-of-the-art systems.

6 Application 2: Poor Man's Stemming: Unsupervised Recognition of Same-stem Words

Abstract

We present a new fully unsupervised human-intervention-free algorithm for stemming for an open class of languages. Since it does not rely on existing large annotated data collections or other linguistic resources than raw text it is especially attractive for low-density languages. The stemming problem is formulated as a decision whether two given words are variants of the same stem and requires that, if so, there is a concatenative relation between the two. The underlying theory makes no assumptions on whether the language uses a lot of morphology or little, whether it is prefixing or suffixing, or whether affixes are long or short. It does however make the assumption that 1. salient affixes have to be frequent, 2. words essentially are variable length sequences of random characters, and furthermore 3. that a heuristic on what constitutes a systematic affix alteration is valid. Tested on four typologically distant languages, the stemmer shows promising results in an evaluation against a human made gold standard.

6.1 Introduction

The problem at hand can be described as follows:

Input: An unlabeled corpus of an arbitrary natural language and two arbitrary words w_1, w_2 from that language

Output: A YES/NO answer as to whether w_1 and w_2 are morphological variants of one and the same stem (according to traditional linguistic analysis).

Restrictions: We consider only concatenative morphology and assume that the corpus comes already segmented on the word level.

The relevance of the problem is that of stemming as applied in Information Retrieval (IR). The issues of stemming in IR has been discussed at length elsewhere and need not be repeated here. It suffices to say that, though not uncontroversial, stemming continues to be a feature of modern IR systems for languages like English (e.g., Google[7]), and is likely to be of crucial importance for languages which make more use of morphology (cf. Pirkola 2001).

The reasons for attacking the problem in an unsupervised manner include advantages in elegance, economy of time and money (no annotated resources required), and the fact that the same technology may be used on new languages. The latter two reasons are especially important in the context of resource-scarce languages.

Our proposed unsupervised same-stem decision algorithm proceeds in two phases. In the first phase, a ranked list of salient affixes are extracted from an

[7] According to `http://www.google.com/help/basics.html` accessed 20 March 2006.

unlabeled text corpus of a language. In the second phase, an input word pair is aligned to shortlist affixes that could potentially be added to a common stem to alternate between the two. Crucially, this shortlist of affix alternations is analyzed to check whether they form a *systematic* alternation in the language as a whole (i.e., not just in the pair at hand). This analysis depends strongly on the ranked affix list from the first phase.

6.2 Same-Stem Decision Desiderata and Heuristics

We use the technique described in Section 3 to extract a list of salient affixes for a given language from raw text data for that language. However, having a list of salient affixes is not sufficient to parse a given word into stem and affix(es). For example, *sing* happens to end in the most salient suffix yet it is not composed of *s* and *ing* because crucially, there is no **s*, **sed* etc. Thus to parse a given word we have to look at additional evidence beyond the word itself, such as the existence of other inflections of potentially the same stem as the given word, or further, look at inflections of other stems which potentially share an affix with the given word. We use the technique described in Section 4 to grow a paradigm for a given affix, given only raw text data as additional input.

The problem at hand, namely, to decide if two given words w_1, w_2 share a common stem (in the linguistic sense) is easier than parsing one word since we have evidence from two words and since we just have to answer yes/no and not also have to give the actual stem. Essentially, there are four interesting kinds of situations the same-stem-decider must face:

1. w_1 and w_2 do share the same stem and have a salient affix each, e.g., *played* vs. *playing*.

2. w_1 and w_2 do share the same stem but one of them has the "zero" affix, e.g., *play* vs. *playing*.

3. w_1 and w_2 do not share the same stem (linguistically) but do share some initial segment, e.g., *playing* vs. *plough*.

4. w_1 and w_2 do not share the same stem (linguistically) and do not share any initial segment, e.g., *playing* vs. *song*.

Number 4 is trivial to decide in the negative. Number 1 is also easy to affirm using a list of salient affixes, whereas the special case of number 2 requires some care. The real difficulty lies in predicting a negative answer for case number 3 (while, of course, at the same time predicting a positive for cases 1 and 2). We will go for an extended discussion of this matter below.

Consider two words $w_1 = xs_1$ and $w_2 = xs_2$ that share some non-empty initial segment x. Except for chance resemblances, which by definition are rare, we would like to say that w_1 and w_2 belong to the same stem iff:

1. s_1 and s_2 are well-segmented salient suffixes in the language, e.g., *-w* and *-lt* for *saw* and *salt* are **not**; and

2. s_1 and s_2 must systematically contrast in the language, that is, there must be a large set of stems which can take both s_1 and s_2. For example, the word pair *sting* and *station* align to *-ing* and *-ation* which are both salient suffixes but they do **not** systematically contrast.

So far, corresponding to the desiderata 1 and 2, we have the Z_W-score that ranks well-segmented salient suffixes and the $VI(P)$-score that ranks how well a set of suffixes contrast. We can put them together as:

$$A(P) = VI(P) \sum_{s \in P} Z_W(s) \tag{7}$$

As a convention we set $Z_W('') = 0$.

The $A(P)$-score gives a score rather than a YES/NO-decision, which is the actual task. We could say that, at this point, the user has to supply a threshold value. However, instead, we devise another heuristic that obviates the need for a threshold at all. The resulting system thus supplies a YES/NO answer to the same-stem decising problem without any human interaction.

Recall the writing convention $w_1 = xs_1$ and $w_2 = xs_2$. Instead of having a threshold we may conjecture that:

w_1, w_2 have the same stem iff $s_1 \in G^*(s_2)$ and $s_2 \in G^*(s_1)$

For example, this predicts that *sting* and *station* are not the same stem because neither $G^*(ing) = \{'', e, ed, er, es, ing, s\}$ contains 'ation' nor does $G^*(ation) = \{ate, ated, ating, ation, ations\}$ contain 'ing'. From our experience this test is quite powerful. However, there are of course cases where it predicts wrongly, due to the greedy nature of the G^*-calculation, e.g., $G^*(ing)$ does not contain 'ers'. Moreover, if one of the affixes is the empty affix, we need a special fix (see below).

6.3 Same-stem Decision Algorithm

We can now put all pieces together to define the full algorithm as shown in Table 26.

If one of s_1, s_2 is the empty string then step 3 and 4 should be restated as follows (using s to denote the non-empty one of the two). The maximization value in step 3 should be modified to: $\frac{Z_W(s)}{1+place('',H_s)}$. Step 4 should be modified to: answer YES/NO acccordingly as $'' \in G^*(s)$.

The bad news is that the computation of the G^*:s tends to be slow due to the summing and sorting of typically very long (50 000-ish items) lists. On my standard PC with a Python implementation it typically takes 30 seconds to decide whether two words share the same stem.

6.4 Evaluation

Several authors, e.g., Goldsmith et al. (2001) and Melucci and Orio (2003), have evaluated their stemming algorithms on Information Retrieval performance.

Input: A text corpus C and two words w_1, w_2

Step 1. Calculate Z_W as in Table 8

Step 2. Form the set of candidate alignment pairs as:

$$C(w_1, w_2) = \{(s1, s2) | xs_1 = w_1 \text{ and } xs_2 = w_2\}$$

Step 3. If $C(w_1, w_2)$ is empty then answer NO, otherwise pick the best candidate pair as:

$$argmax_{(s1,s2) \in C(w_1,w_2)} A(\{s_1, s_2\})$$

Step 4. For the winning pair, answer YES/NO acccordingly as $s_1 \in G^*(s_2)$ and $s_2 \in G^*(s_1)$

Table 26. Summary of same-stem decision algorithm. If one of s_1, s_2 is the empty string then steps 3 and 4 should be modified as follows (using s to denote the non-empty one of the two). The maximization value in step 3 changes to: $\frac{Z_W(s)}{1+place('', H_s)}$. Step 4 changes to: answer YES/NO acccordingly as $'' \in G^*(s)$.

While IR is the undoubtedly the major application area, we feel that evaluating on retrieval performance does not answer all relevant questions of stemming performance. For instance, a stemmer may make conflations and miss conflations that simply did not affect the test queries. In fact, one may get different best stemmers depending on the test collection. There is also difference as to whether the whole document collection, an abstract of each document or just the query is stemmed.

We find it more instructive to test stemming separately against a stemming gold standard and assess the relevance of stemming for IR by testing the stemming gold standard on IR performance. If stemming turns out to be relevant for IR, then researchers should continue to develop stemming algorithms towards the gold standard. In the other case, one wonders whether IR-improving term conflation methods should rightly be called stemmers.

In order to assess the cross-linguistic applicability of our stemming algorithm we have chosen languages spanning the spectrum of morphological typology – from isolating to highly suffixing – Maori, English, Swedish and Kuku Yalanji (Dryer 2005). As training data we used only the set of words from a bible translation to emphasize the applicability to resource-scarce languages. Table 27 contains information on the bible versions used.

For the four languages we devised a stemming gold standard, consulting descriptive materials for Maori (Bauer et al. 1993, Williams 1971) and Kuku Yalanji (Patz 2002, Hershberger and Hershberger 1982), languages not generally known to the author. So as not to let the test set be dominated by too many simple test cases, the selection of test set cases was done as follows:

1. Select a random word w_1 from W for the corresponding language

Language	Language Type	Corpus	Scope
Maori	Isolating	British & Foreign Bible Society 1996	NT & OT
English	Mildly Suffixing	King James 1977	NT & OT
Swedish	Suffixing	Svenska Bibelsällskapet 1917	NT & OT
Kuku Yalanji	Strongly Suffixing	Summer Institute of Linguistics 1985	NT & OT Parts

Table 27. Summary of Bible corpora used in evaluation.

	same-stem		diff.-stem	
Language	Correct	Total	Correct	Total
Maori	10	13	100	100
English	97	100	100	100
Swedish	96	100	100	100
Kuku Yalanji	94	100	100	100

Table 28. Evaluation results of same-stem decisions given 100 test pairs for each language (see text).

2. Select a random number i in $0 \leq i \leq |w_1| - 1$
3. Select a random word w_2 from the subset of words from $W \setminus \{w_1\}$ sharing i initial characters with w_1
4. Mark the pair w_1, w_2 to be of the same stem according to traditional linguistic analysis

This was repeated until 200 pairs of words for each language had been selected, 100 same-stem and 100 not same-stem. Except for Maori where we could only really find 13 same-stem cases this way, all involving active-passive alternating verbs (described in detail in Sanders 1990).

The evaluation results are shown in Table 28.

The errors fall into just one major type, in which the algorithm is too cautious to conflate; it is when two words do share the stem but where one of the suffixes is rather uncommon (including the case where it is actually composite) and therefore it is not in the grow-set of the other suffix; for example Swedish *skap-ade-s* (passive past) and *skap-are-n-s* (agent-noun definite genitive). We also expected false positives in the form of random resemblances involing short words and short affixes; e.g., *as* versus *a* but no such cases seem to have occurred in the test set in any of the languages.

We have done attempted a comparison with other existing stemmers, mainly because they tend not be aimed at an open set of languages and those which are, are really not fully supervised and we fear we might not do justice to them in setting parametres (see the subsection below on related work). The widely known Porter stemmer (Porter 1980) for English scores exactly the same result for English as our stemmer, which suggests than an unsupervised approach may come very close to explicity human-informed stemmers. Many other stemmers, however, are superior to ours in the sense that they can stem a single word

correctly whereas ours requires a pair of words to make a decision. This is especially relevant when large bodies of data needs to be stem-indexed as it would take quadratic time (in the number of words) in our setting.

6.5 Related Work

A full survey of stemming algorithms for specific languages or languages like English has more or less fully been done elsewhere (the technology becoming relatively mature cf. Erjavec and Džeroski 2004, Frakes and Fox 2003, Goldsmith et al. 2001, Melucci and Orio 2003, Rogati et al. 2003, Hull 1996, Galambos 2004, Flenner 1994 and references therein). We will focus instead on unsupervised approaches for a wider class of languages.

Melucci and Orio (2003) present a very elegant unsupervised stemming model. While training does not require any manually annotated data, some architectural choices depending on the language still has to be supplied by a human. If this can be overcome in an easy way, it would be very interesting to test their Baum-Welch training approach versus the explicit heuristics in this paper, especially on a wider scope of languages than given in their paper. The unsupervised stemmer outlined in Goldsmith et al. (2001) actually requires a lot of parametres to be tweaked humanly and mainly targets languages with one-slot morphology.

Other systems for unsupervised learning of morphology which do not explicitly do stemming could easily be transformed into stemmer (Jacquemin 1997, Yarowsky and Wicentowski 2000, Ćavar et al. 2004a, Brent et al. 1995, Déjean 1998a, Snover et al. 2002, Argamon et al. 2004, Goldsmith 2001, Gaussier 1999, Sharma et al. 2002, Oliver 2004). All of these systems, however, require some parameter tweaking as it is and perhaps one more if transformed to stemmers, so there is still work outstanding before they can be compared on equal grounds to the stemmer described here. Given that they use essentially the same kind of evidence, it is likely that some of them, especially Creutz and Lagus (2007), will reach just as competitive results on the same task.

6.6 Conclusion

We have presented a fully unsupervised human-intervention-free algorithm for stemming for an open class of languages showing very promising accuracy results. Since it does not rely on existing large data collections or other linguistic resources than raw text it is especially attractive for low-density languages. Although polynomial in time, it appears rather slow in practice and may not be suitable for stemming huge text collections. Future directions include investigating whether there is a speedier shortcut and better, more systematic, approach to layered morphology, i.e., languages which allow affixes to be stacked.

Acknowledgements

The author has benefited much from discussions with Bengt Nordström. We also wish to extend special thanks to ASEDA for granting access to electronic versions of the Kuku Yalanji bible texts.

7 Application 3: Poor Man's Word-Segmentation: Unsupervised Morphological Analysis for Indonesian

Abstract

We present a partially new fully unsupervised algorithm for morphological segmentation of a arbitrary natural language with only one-slot concatenative morphology. The behaviour of the algorithm is examined in detail for Indonesian as it is a good approximation of such a language. The underlying theory makes no assumptions on whether the language is prefixing or suffixing, or whether affixes are long or short. It does however make the assumption that 1. salient affixes have to be frequent, 2. words essentially are variable length sequences of random characters, and furthermore 3. that a heuristic on what constitutes a systematic affix alteration is valid. The only input required is raw unannotated text and there are no thresholds or parameters that need human tuning. Since there no reliance on existing large data collections or other linguistic resources than raw text, the approach is especially attractive for low-density languages. We will discuss the pressing question whether unsupervised approaches are advantageous over the alternative approach with human-built rules and lexica, especially as it pertains to languages like Indonesian, which do not have much morphology in the first place.

7.1 Introduction

Indonesian is a language with very little morphology, but nevertheless, there is some. Thus, the first step in the computational treatment of Indonesian is (a tool for) morphological parsing. A traditional approach, be it for Indonesian or another language, is to build a lexicon and morphological rules by hand. We may call this the *manual* approach. In the present paper, we explore another approach, namely to induce a morphological parser from raw (unannotated) text data and, in fact, without any human intervention at all. This will be called the *unsupervised* approach. We will discuss the advantages and disadvantages of both alternatives extensively, with Indonesian as the case at hand.

7.2 Problem Statement

The problem at hand can be described as follows:

Input: An unlabeled text corpus of a natural language where

(a) The corpus comes already segmented on the word level.

(b) The language only has one-slot concatenative morphology.

Output: The same text corpus annotated with morphological divisions

Any language with an orthography that shows word-divisions, e.g., with spaces, passes the (a)-restriction right away. So Indonesian, in its contemporary orthography is immediately applicable, while, e.g., Thai is not. The (b)-restriction mandates that the language in question only has concatenative morphology, and furthermore, that affixes cannot be stacked after another, i.e., that there is only maximally one 'slot' for suffixes and maximally one 'slot' for prefixes. In other words, agglutinative languages are not the target of the present problem. Indonesian passes this restriction, or nearly so, as verbal prefixes (like *di-*) cannot be stacked, and the only inflectional suffixes are the personal endings (like *-nya*). On the other hand, multiple affixes do occur somewhat less commonly where the inner suffix is derivational, e.g., *-an-mu* or *-kan-lah*. Similarly, English would be a good one-slot language except for combinations of derivational and inflectional affixes, such as *-ation-s* and latinate prefix stacking such as *dis-en-*.

Considering languages which have concatenative morphology at all, Indonesian is as close to a one-slot language as one may hope to find (Dryer 2005). For this reason, Indonesian provides an excellent case for study of unsupervised versus manual approaches to morphological parsing. (The reason for having the one-slot restriction at all, is, that the unsupervised approaches can be considerably less complex.)

7.3 Manual versus Unsupervised Methods

In the manual approach to morphological parsing, typically, a human implements the following:

(a) A set of morphological rules for parsing and generation

(b) A lexicon of stems

There are many different frameworks for the implementation of rules and lexicon, e.g., XFST (Beesley and Karttunen 2003) or FM (Forsberg 2007) to name a few. However, the choice of framework has little bearing on the issues in this paper, wherefore we disregard this matter further.

The advantages of the manual approach is that one gets a transparent high-accuracy analyser that segments and labels morphologically complex words according to traditional linguistic analysis.

The disadvantages of the manual approach is the human labour required, both for writing the morphological rules and for building the lexicon. On the other hand, in some cases, a finished lexicon-resource is accessible or purchasable as a module, e.g., via the publisher of a dictionary, and then this major part of the human labour is covered already. Another option to reduce the human labour in lexicon-building is to "extract" a lexicon from raw-text. Given morphological paradigms and search constraints, an algorithm extracts a list of stems that are evidenced to belong to a given paradigm in the raw text. A human only needs to skim such a list of extracted lexical items to weed out false positives (see Forsberg et al. 2006, Lindén 2008, Adler et al. 2008, Carlos et al. 2009 and

references therein). Furthermore, any lexicon-based approach to morphological analysis will fail on out-of-lexicon words which inevitably occur in open-domain text (such as newspaper text). As we shall see, there is considerable overlap between out-of-lexicon guessing, lexicon extraction and unsupervised methods.

The unsupervised approach has the advantage that it eliminates the human labour and there is a certain elegance in that the same techniques can be used for different languages.

Naturally, the disadvantage of the unsupervised approaches is that they do not reach full accuracy of morphological analysis. Also, in addition to clear errors, one may expect a certain amount of morpheme segmentations which are slightly at odds with traditional linguistic analysis but not necessarily erroneous. Furthermore, the segmented morphemes are not labeled (or a separate less-than-perfect module is required to learn appropriate labels – see, e.g., Schone and Jurafsky (2001b)) and for many, but not all, applications such labels are in demand.

Various other factors, such as module size or analysis speed, no longer play significant roles, and thus do not contrast between the two alternative approaches. The key issue is how much behind in accuracy unsupervised approaches actually fare – a matter that has so far been rather unclear (see also Section 7.4). One of the main goals of the present paper is to clarify the situation.

A recent survey of published work on morphological analysis for so-called low-density languages shows that the vast majority are of the manual kind (Hammarström 2009b). While the contrast between manual and unsupervised approaches are mostly relevant for low-density languages, the division is not so simple. Perhaps as a result of the human labour required, most manually constructed morphological analysers are **not** freely available.

7.4 Previous Work on Unsupervised Morphological Analysis

A full but concise survey up to late 2007 can be found in Hammarström (2007b). Some relevant work have appeared since, notably Tepper and Xia (2008), Monson (2009), Johnson (2008). A wide variety of heuristics and models have been employed, which are not of particular importance for this paper, though the interested reader is advised to consult the surveys of (Roark and Sproat 2007, Goldsmith pear).

The relevant issue is the fact that nearly all approaches published so far – exceptions include Golcher (2006) and Hammarström (2006b) – have a little supervision in the form of thresholds or parameters that have to been set and tuned by a human. This, along with the fact that many approaches target a different (related, but not identical) morphological segmentation problem makes most of this work not directly applicable to the problem considered in the present paper. For these reasons, we will focus on a particular line of work on unsupervised morphological segmentation, which is free from any kind of parameters or thresholds. The (very important) comparison between different unsupervised

or little supervised systems is better handled in the annual MorphoChallenge controlled competition[8].

7.5 Poor Man's Word-Segmentation

In this section we will describe a fully unsupervised algorithm for word-segmentation as applied to Indonesian. The work is an extension of Hammarström (2006b) included in Section 6. The following is an outline of the steps and components in the model and its application.

Model: A model is compiled from raw text data

> **Affix Extraction:** Extracts a list of affixes ranked according to salience, i.e., how likely they are to be 'real' suffixes in the language in question
>
> **Affix Alternation:** Given an affix, finds the set of suffixes that systematically appears on the same stems as the given affix
>
> **Affix Purging:** Given a ranked list of salient affixes (as in the extraction component above), weed out the 'unnecessary' ones, namely those which are covered by a more salient affix

Application: The model is applied to a seen or unseen word and analyses it

> **True/Random-Ending Heuristic:** A word that ends in a salient affix that survived purging is either truly composed of a stem and this affix, or it just happens to end in (or begin with) a character sequence that is identical to a salient affix. The true/random-ending heuristic makes this decision, using affix alternation as a component.

All the above components rely on a number of heuristics which are tailored to the Indonesian language type. The heuristics will be discussed in their due place.

We will illustrate the methods on suffixes, but it is obvious that the same procedures can be used to target prefixes but looking at the word from the opposite direction.

As our input raw text corpus we have used the Indonesian bible (American Bible Society 2003), solely because of electronic availability, reproduceability and comparability with other languages. The Indonesian bible amounts to 657 112 word tokens and 15 251 word types (all words lowercased).

Affix Extraction

We use the technique described in Section 3 to extract a list of salient affixes for a given language from raw text data for that language. For Indonesian we get the top 30 plus bottom 3 suffixes as shown in Table 29. The results of this first step do not immediately appear useful, but will be after some postprocessing.

[8] The next edition is the 2009 one `http://www.cis.hut.fi/morphochallenge2009/` accessed 1 May 2009.

-anmu	87195.4	*-mu*	6350.9
-nya	73694.0	*-kannya*	5981.8
-anku	51923.8	*-an*	5931.3
-lah	39535.9	*-arlah*	5702.8
-kanlah	27933.2	*-atlah*	5148.6
-hnya	20915.8	*-anlah*	4522.9
-inya	19677.7	*-nglah*	4121.3
-atnya	18361.2	*-anglah*	3990.4
-kan	18318.4	*-akanlah*	3882.9
-tnya	14237.1	*-hlah*	3658.6
-iel	10000.7	*-ah*	3596.1
-snya	9635.2	*-nku*	3359.6
-rlah	9351.8	*-ya*	3203.6
-annya	7802.0	...	0.0
-ilah	7534.3	*-aadil*	0.0
-anya	7005.4	*-aadai*	0.0
-nmu	6531.7	*-aaan*	0.0

Table 29. Top 30 and bottom 3 extracted suffixes for an Indonesian bible corpus.

For example, when a shorter string has higher score than its extension, e.g., *-nya* versus *-tnya*, the longer one should obviously be discarded. It is also worth noting that in cases where Indonesian does show stacked affixes, we get somewhat confused outcomes; sometimes the composition is seen as one simplex segment as in *-anmu* and sometimes the last layer has a higher score than the compostion as in *-lah* versus *-kanlah*.

Affix Alternation

We use the technique described in Section 4 to grow paradigms given only raw text data as input. To give the reader a feeling for the outcome for Indonesian specifically, Table 30 shows two examples of quotient functions (sorted on highest value) and Table 31 gives two examples of V_P-lists. The latter contrasts the very common paradigm $\{-nya, -mu, -'', -ku\}$ with the nonsense paradigm $\{nya, s, a, ya\}$ consisting only of individually frequent suffixes. The ranks of the members of P_1 to the left are [0, 1, 2, 3], and for P_2 to the right the ranks are [24, 32, 50, 79]. Two growth-examples are shown in Table 32, one which attains a perfect 1.0 score and one in which the original member is not a well-segmented to begin with (the pattern such cases exhibits will be exploited for purging).

Affix Purging

Now, if we return to the ranked list of suffixes in Z_W as of Table 29. As mentioned, one purging heuristic is to discard longer affixes which have a tail whose score is higher. This may be achieved by keeping only those suffixes

y	$H_{nya}(y)$	y	$H_{mu}(y)$
nya	1.000	*mu*	1.000
"	0.913	"	0.943
mu	0.261	*nya*	0.749
ku	0.153	*ku*	0.393
kan	0.071	*kan*	0.063
lah	0.057	*lah*	0.059
an	0.043	*an*	0.056
i	0.039	*kanlah*	0.045
kanlah	0.034	*i*	0.040
kah	0.016	*kah*	0.022
ilah	0.015	*annya*	0.022
annya	0.014	*ilah*	0.019
kannya	0.011	*anmu*	0.015
t	0.008	*n*	0.014
k	0.008	*m*	0.014
anmu	0.008	*k*	0.012
n	0.007	*ng*	0.011
m	0.007	*mulah*	0.011
h	0.007	*t*	0.009
s	0.005	*anku*	0.009
ng	0.005	*wan*	0.008
l	0.005	*ya*	0.007
inya	0.004	*ta*	0.007
ya	0.004	*s*	0.007
...	...	...	...

Table 30. Sample quotient functions/lists for *nya* and *mu*.

y	$V_{P_1}(y)$	y	$V_{P_2}(y)$
"	2.789	"	1.025
nya	1.652	*mu*	0.273
mu	1.004	*ku*	0.166
ku	0.572	*snya*	0.098
lah	0.243	*i*	0.092
kan	0.231	*kan*	0.089
an	0.197	*lah*	0.076
i	0.157	*an*	0.069
kanlah	0.137	*n*	0.064
annya	0.075	*anya*	0.058
ilah	0.068	*slah*	0.057
kah	0.065	*skan*	0.053
n	0.049	*k*	0.053
anmu	0.047	*ng*	0.052
m	0.043	*san*	0.048
t	0.037	*t*	0.045
k	0.036	*kanlah*	0.044
anku	0.033	*m*	0.043
ng	0.032	*si*	0.042
h	0.031	*r*	0.042
mulah	0.031	*h*	0.042
ya	0.031	*l*	0.030
ta	0.029	*amu*	0.030
s	0.027	**nya**	0.027
...	...	...	...

Table 31. Example ranks for $P_1 = \{nya, mu, ", ku\}$ and $P_2 = \{nya, s, a, ya\}$.

P	$VI(P)$
('nya',)	0.000
('', 'nya')	0.333
('', 'mu', 'nya')	0.750
('', 'ku', 'mu', 'nya')	1.000

P	$VI(P)$
('s',)	0.0
('s', 'snya')	0.077
('s', 'smu', 'snya')	0.273
('s', 'sku', 'smu', 'snya')	0.667
('s', 'san', 'sku', 'smu', 'snya')	0.833
('s', 'san', 'skan', 'sku', 'smu', 'snya')	0.882

Table 32. Example iterations of G^*('nya') and G^*('s').

s	*-nya*	*-s*
$G^*(s)$	$\{'', ku, mu, nya\}$	$\{s, san, skan, sku, smu, snya\}$
$\overline{s}P$	$\{'', ku, mu, nya\}$	$\{s\}$
sP	$\{\}$	$\{an, kan, ku, mu, nya\}$
$\sum_{x \in \overline{s}P} Z_W(x)$	80790.7	405.5
$\sum_{x \in sP} Z_W(x)$	0.0	105040.5

Table 33. The purging calculation for example suffixes *-nya* (remains) and *-s* (purged out).

which are best parse for at least one word.

$$U'_W = \{s | s = argmax_{s' \lhd w} Z_W(s') \text{ for some } w \in W\}$$

This purging heuristic is not sufficient, as a certain amount spurious suffixes – albeit with low Z_W-score – remain, e.g., *-s*. At this point, one could introduce a threshold value to weed out the rest of the spurious suffixes. However, it turns out that there is another heuristic that does the same job, without human intervention.

The behaviour of the G^*('s')-set shown in Table 32 is typical for spurious suffixes. Indeed, if s is a spurious suffix, $G(s)$ is likely to consist of s-prefixed to members of a 'true' paradigm. We may therefore posit the following criterion for true-suffixness of a suffix s. Split $G^*(s)$ into $sP = \{x | sx \in G^*(s), |x| > 0\}$, i.e., the members which consist of s followed by some non-empty string versus the rest $\overline{s}P = \{x | x \in G^*(s), x \neq sy, |y| > 0\}$. Note that s itself belongs to $\overline{s}P$ rather than sP and that sP contains the "tail"-strings, stripped of their initial s. If the sum of Z_W-values for $\overline{s}P$ is strictly larger than the the sum of Z_W-values for sP, then the s is a true suffix, otherwise not. Table 33 shows two examples.

The final purged set of suffixes may accordingly be defined as:

$$U''_W = \{s \in U'_W | \sum_{x \in \overline{s}P} Z_W(x) > \sum_{x \in sP} Z_W(x)\}$$

The U''_W for Indonesian boils down to the follwing suffixes:

> {anmu, nya, anku, lah, kan, nmu, mu, nku, i, ezer, zabad, inadab, ihud, nadab, arif, obab, ezib, ilene, laf, ilo, ore, e}

The existence of some remaining spurious affixes, such as *-ilene* makes little difference, as these affixes are very infrequent and do not significantly diminish segmentation accuracy. A real error, however, is where Indonesian deviates from being a one-slot language. The lack of *-an*, which is purged out because *-mu* and *-ku* are frequently attached to it.

True/Random-Ending Heuristic

A word that ends in a salient affix that survived purging is either truly composed of a stem and this affix, or it just happens to end in (or begin with) a character

sequence that is identical to a salient affix. The true/random-ending heuristic makes this decision, using affix alternation as a component. Consider the word 'gadisnya' where the *-nya* is a true occurrence of the suffix *-nya*, and the word 'hanya' which just happens to end in the *-nya* character sequence. How can we (heuristically) distinguish the two cases when there is no stem lexicon to tell us that there is a stem *gadis-* but no stem *ha-*?

The intution is the following. If we count all the words in W which begin with *ha-*, they are rather many (247 to be exact). It is not too surprising that one of them would just by chance continue as *-nya*, and, crucially, none of the other 246 continuations are *-"*, *-mu* or *-ku*, i.e., the $G^*(nya)$ affixes which systematically alternate with *-nya*. On the other hand, if we look that the words in W which begin with *gadis-*, there are only two more than 'gadisnya', namely 'gadis' and 'gadismu', i.e., with *-"* and *-mu*! To turn this into a formal criterion, given a word $w = xs$ such that $s \in U''_W$, let $C_W(x) = \{y | xy \in W\}$ be the set of "continuations" of x. Extend the notion of final frequency to a set of suffixes P as $f_W(P) = |\{w \in W | w = yz \text{ for some } z \in P\}|$ and let $\alpha_W(P) = f_W(P)/|W|$. The heart of the matter is how much is inside $C_W(x) \cap G^*(s)$ versus how much is outside $C_W(x) \setminus G^*(s)$. Given $w = xs$, if s is just a random continuation of an initial segment x, then many items in $C_W(x) \cap G^*(s)$ will be hard to explain, and if s is truly a well-segmented then too many random continuations in $C_W(x) \setminus G^*(s)$ will overshadow this fact. In other words, the $w = xs$ should be segmented iff the following ratio ≥ 1:

$$\frac{(1 - \alpha_W(G^*(s)))^{|C_W(x) \setminus G^*(s)|}}{\alpha_W(G^*(s))^{|C_W(x) \cap G^*(s)|}}$$

Or equivalently, iff the following subtraction ≥ 0:

$$|C_W(x) \setminus G^*(s)| \cdot \log(1 - \alpha_W(G^*(s))) - \\ (|C_W(x) \cap G^*(s)|) \cdot \log \alpha_W(G^*(s))$$

To better illustrate the calculation, an example with actual numbers is shown in Table 34.

In summary, the following is a procedure for segmenting (or not) a word w (given a training wordset W):

1. Calculate U''_W
2. Consider any $s \triangleleft w$ such that $s \in U''_W$. If there is no such s, then output w (no segmentation)
3. Otherwise, break $w = xs$ and apply the true/random-ending heuristic and output x (segmentation appropriate) or w (no segmentation) accordingly.

7.6 Evaluation

Evaluation was made on a small hand-made test set of 100 word types. 100 word types were selected from W at random and manually segmented both

w	'gadisnya'	'hanya'
x	*gadis-*	*ha-*
s	*-nya*	*-nya*
$G^*(s)$	$\{'', ku, mu, nya\}$	$\{'', ku, mu, nya\}$
$\alpha_W G^*(s)$	0.205	0.205
$C_W(x)$	$\{'', nya, mu\}$	$\{nya, sratmu, ncurkanlah, rta,$
$C_W(x) \cap G^*(s)$	$\{'', nya, mu\}$	$\{nya\}$
$C_W(x) \setminus G^*(s)$	$\{\}$	$\{sratmu, ncurkanlah, rta, \ldots\}$
$\lvert C_W(x) \setminus G^*(s)\rvert \cdot \log(1 - \alpha_W(G^*(s)))$	0.0	-56.9
$(\lvert C_W(x) \cap G^*(s)\rvert) \cdot \log \alpha_W(G^*(s))$	-4.8	-1.61

Table 34. The True/Random-Ending Heuristic as applied to the words 'gadisnya' (segmentation appropriate) and 'hanya' (no segmentation).

prefix-wise and suffix-wise. For example, 'direncanakannya' was segmented 'direncana-kan-nya' and 'mengerutkan' segmented to 'meng-erut-kan' (in all cases where the first character of the root changes as a prefix is added, we arbitrarily chose to define the correct segmentation border so that the result of the mutation belongs to the prefix part). A total of 64 segmentations were found on the 100 words. The unsupervised algorithm was applied to the 100 words, once suffix-wise and once prefix-wise. 58 of segmentations were appropriately found, none spurious and 6 missed (i.e., full precision but 58/64=90.6% recall). All 6 of the missed segmentations were words with stacked affixes such as *-kan-lah*, or with final *-an*.

We know of no purely unsupervised stemming approach to Malay/Indonesian. On the other hand, a relatively large number of descriptions of work on manual, supervised and semi-supervised Malay/Indonesian stemmers/analyzers have appeared in the literature (Pisceldo et al. 2008, Adriani et al. 2007, Abdullah et al. 2009, Tai et al. 2000, Ahmad et al. 1996, Ranaivo 2001, Ranaivo-Malançon 2004, Indradjaja and Bressan 2003). Unfortunately, none of these stemmer/analyzers appears to be available for online processing or as a downloadable program, and, though well-described, require a fair amount of manual labour to reproduce. Therefore we are unable to present accuracy figures for comparison.

7.7 Discussion

The outlined algorithm is admittedly a complex path of rather untransparent heuristics whose properties we are, at this stage, not able to prove in a mathematically rigourous manner. Nevertheless, all of them have a clear intuition and therefore some of its virtues and errors are easily explainable. Also, as promised, there are no human thresholds or parameters whatsoever.

The accuracy is promising, but we do expect false positives to turn up in larger test sets, such as oversegmentations of *-i* and segmentation with one of the very rare spurious suffixes that survived purging. Undersegmentation, as a result of the one-slot straitjacket, is the main error. Future work, naturally, will

focus on extending the heuristics to the multi-slot case. Our feeling is that this is possible, with the same level of accuracy, at the cost of an even more complex web of heuristics.

The present experiment aims to show that, at least for a certain type of languages, unsupervised approaches are competitive accuracywise, and likely to be favoured in a labour/accuracy stand-off. On the other hand, Indonesian morphology is so simple that rule-writing would so take little time, that a hybrid system of hand-written rules and lexicon-less heuristics borrowed from the present approach, would be a serious competitor as well.

7.8 Conclusion

We have presented a partially novel unsupervised algorithm for morphological segmentation of a arbitrary natural language with only one-slot concatenative morphology. The algorithm achieves high accuracy on Indonesian, a language with little, but mostly one-slot, morphology. An extension of the algorithm to deal systematically with multi-slot morphology is a priority for future work. The presented experiment with Indonesian clarifies the position of unsupervised methods for morphological analysis for low-density languages as an alternative to traditional manual implementation of rules and lexicon.

8 Application 4: Bootstrapping Language Description: The case of Mpiemo (Bantu A, Central African Republic)

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.

8.1 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.

8.2 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)[9]. 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.*

[9] Cf. the journal/newsletter Language Archives News `http://www.mpi.nl/LAN/`

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.[10] 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,b) 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. 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.

8.3 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).

[10] 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 (No Author Stated 2006b,a).

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, 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 35.

Bandi hɛ ri kɛ gwɔbi i ri bɛ dɛ gɔ: Hi nɔ mɛligi, hi kɛ bɛ sombi Mpanja, hi jɔɔ̀ pèà gɔ, ha nɛ Kamil hó ri ké. Hí jɔɔ̀ pèà gó, Kamil nɔ mɛlándi. Hí kè̀ jɔɔ̀ téri sómbi, a nó mélígí, à wá tí sómbi ya. Mè̀ rì yé nyè̀ mè̀kɔ̀gì. À lán mélígí má tí sombi yà gó. Hi kwàn, hí sàà, hí ké bé 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 35. Sample snippet of Mpiemo text.

8.4 Bootstrapping Experiments

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 (2007b) 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 2007b) 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.[11] As for prefixes, the algorithm found the segmentations listed in Table 36. 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 36. 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 (presented in Section 4). Together with prefix extraction, we get a ranked list of <stem, prefix-set> pairs. The top pairs are shown in Table 37. 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.

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

Table 37. Top pairs in paradigm induction.

[11] 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.

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 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.

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.

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 choice 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.

8.5 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.

8.6 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 sim-

ple, 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.

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.*

References

Abdullah, M. T., Ahmad, F., Mahmod, R., and Sembok, T. M. T. (2009). Rules frequency order stemmer for Malay language. *International Journal of Computer Science and Network Security*, 9(2):433–438.

Adler, M., Goldberg, Y., Gabay, D., and Elhadad, M. (2008). Unsupervised lexicon-based resolution of unknown words for full morphological analysis. In *Proceedings of ACL-08: HLT*, pages 728–736, Columbus, Ohio. Association for Computational Linguistics.

Adriani, M., Asian, J., Nazief, B., Tahaghoghi, S. M. M., and Williams, H. E. (2007). Stemming Indonesian: A confix-stripping approach. *ACM Transactions on Asian Language Information Processing (TALIP)*, 6(4):1–33.

Ahmad, F., Yusoff, M., and Sembok, T. M. T. (1996). Experiments with a stemming algorithm for malay words. *Journal of the American Society for Information Science*, 47(12):909–918.

American Bible Society (1988). *Turkish Bible.* Tulsa, Oklahoma: American Bible Society.

American Bible Society (1999). *Bib La [Haitian Creole Bible].* Tulsa, Oklahoma: American Bible Society.

American Bible Society (2003). *Alkitab [Indonesian Bible].* Tulsa, Oklahoma: American Bible Society.

Andreev, N. D. (1965). Opyt statistiko-kombinatornogo vydeleniya pervogo morfologičeskogo tipa v Vengerskom yazyke. In Andreev, N. D., editor, *Statistiko-kombinatornoe modelirovanie Yazykov*, pages 205–211. Moscow: Akademia Nauk SSSR.

Andreev, N. D. (1967). *Statistiko-kombinatornye metody v teoretičeskom i prikladnom yazykovednii.* Leningrad: Nauka.

Arabsorkhi, M. and Shamsfard, M. (2006). Unsupervised discovery of persian morphemes. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2006, April 3-7, 2006, Trento, Italy: Demonstrations Session*, pages 175–178. The Association for Computer Linguistics.

Argamon, S., Akiva, N., Amir, A., and Kapah, O. (2004). Efficient unsupervised recursive word segmentation using minimum description length. In *Proceedings of COLING 2004*, pages 1058–1064, Geneva, Switzerland. COLING.

Atwell, E. and Roberts, A. (2005). Combinatory hybrid elementary analysis of text. In Kurimo, M., Creutz, M., and Lagus, K., editors, *Proceedings of MorphoChallenge 2005*, pages 37–41.

Baayen, H. R. (2001). *Word Frequency Distributions*, volume 18 of *Text, Speech, and Language Technology*. Kluwer, Dordrecht.

Bacchin, M., Ferro, N., and Melucci, M. (2002a). The effectiveness of a graph-based algorithm for stemming. In Lim, E. P., Foo, S., Khoo, C. S. G., Chen, H., Fox, E. A., Urs, S. R., and Thanos, C., editors, *ICADL '02: Proceedings of the 5th International Conference on Asian Digital Libraries*, volume 2555 of *Lecture Notes in Computer Science*, pages 117–128. Springer-Verlag, Berlin.

Bacchin, M., Ferro, N., and Melucci, M. (2002b). University of Padua at CLEF 2002: Experiments to evaluate a statistical stemming algorithm. In *Working Notes for CLEF 2002: Cross-Language Evaluation Forum Workshop*, pages 161–168. Rome.

Bacchin, M., Ferro, N., and Melucci, M. (2005). A probabilistic model for stemmer generation. *Information Processing and Management*, 41(1):121–137.

Baeza-Yates, R. and Ribeiro-Neto, B. (1997). *Modern Information Retrieval*. Addison-Wesley.

Baroni, M. (2000). *Distributional Cues in Morpheme Discovery: A Computational Model and Empirical Evidence*. PhD thesis, University of California, Los Angeles.

Baroni, M. (2003). Distribution-driven morpheme discovery: A computational/experimental study. *Yearbook of Morphology*, pages 213–248.

Baroni, M., Matiasek, J., and Trost, H. (2002). Unsupervised discovery of morphologically related words based on orthographic and semantic similarity. In *Proceedings of the Workshop on Morphological and Phonological Learning of ACL/SIGPHON-2002*, pages 48–57.

Batchelder, E. O. (1997). *Computational evidence for the use of frequency information in discovery of the infant's first lexicon*. PhD thesis, City University of New York.

Bati, T. B. (2002). Automatic morphological analyser: An experiment using unsupervised and autosegmental approach. Master's thesis, Addis Ababa University, Ethiopia.

Bauer, W., Parker, W., and Evans, T. K. (1993). *Maori.* Descriptive Grammars. London & New York: Routledge.

Beermann Hellan, D. (2007). Development of linguistic documentation tools under the umbrella of NUFU. Presentation at the Year of African Languages Symposium, April 2007, Gothenburg.

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

Belkin, M. and Goldsmith, J. (2002). Using eigenvectors of the bigram graph to infer morpheme identity. In *Morphological and Phonological Learning: Proceedings of the 6th Workshop of the ACL Special Interest Group in Computational Phonology (SIGPHON)*, pages 41–47, Philadelphia. Association for Computational Linguistics.

Bernhard, D. (2005a). Segmentation morphologique à partir de corpus. In *Actes de TALN & RÉCITAL 2005*, volume 1, pages 555–564. ATALA, Dourdan, France.

Bernhard, D. (2005b). Unsupervised morphological segmentation based on segment predictability and word segments alignment. In Kurimo, M., Creutz, M., and Lagus, K., editors, *Unsupervised segmentation of words into morphemes – Challenge 2005*, pages 18–22.

Bernhard, D. (2006). *Apprentissage de connaissances morphologiques pour lÃĆÂŠacquisition automatique de ressources lexicales.* PhD thesis, Université Joseph Fourier – Grenoble I.

Bernhard, D. (2007a). Apprentissage non supervisé de familles morphologiques par classification ascendante hiérarchique. In *Actes de la 14e conférence sur le Traitement Automatique des Langues Naturelles, TALN 2007*, volume 1, pages 367–376. Toulouse, France.

Bernhard, D. (2007b). Simple morpheme labelling in unsupervised morpheme analysis. In Nardi, A. and Peters, C., editors, *Working Notes for the CLEF 2007 Workshop, 19-21 September, Budapest, Hungary.*

Biemann, C. (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.

Biemann, C. and Teresniak, S. (2005). Disentangling from babylonian confusion - unsupervised language identification. In Gelbukh, A. F., editor, *Computational Linguistics and Intelligent Text Processing, 6th International Conference, CICLing 2005, Mexico City, Mexico, February 13-19, 2005, Proceedings*, volume 3406 of *Lecture Notes in Computer Science*, pages 773–784. Springer.

Blomqvist, J. and Jastrup, P. (1998). *Grekisk Grammatik: Graesk grammatik.* København: Akademisk Forlag, 2 edition.

Bordag, S. (2005a). Two-step approach to unsupervised morpheme segmentation. In Kurimo, M., Creutz, M., and Lagus, K., editors, *Unsupervised segmentation of words into morphemes – Challenge 2005*, pages 23–27.

Bordag, S. (2005b). Unsupervised knowledge-free morpheme boundary detection. In *Proceedings of Recent Advances in Natural Language Processing 2005 (RANLP '05)*. Borovets, Bulgaria.

Bordag, S. (2007a). *Elements of Knowledge-free and Unsupervised lexical acquisition.* PhD thesis, University of Leipzig, Leipzig.

Bordag, S. (2007b). Unsupervised and knowledge-free morpheme segmentation and analysis. In Kurimo, M., Creutz, M., and Lagus, K., editors, *Unsupervised segmentation of words into morphemes – Challenge 2007.*

Bordag, S. (2007c). Unsupervised and knowledge-free morpheme segmentation and analysis. In Nardi, A. and Peters, C., editors, *Working Notes for the CLEF 2007 Workshop, 19-21 September, Budapest, Hungary.*

Borin, L. (1991). *The Automatic Induction of Morphological Regularities.* PhD thesis, University of Uppsala.

Borin, L. (1997). Parole-korpusen vid språkbanken, göteborgs universitet. `http://spraakbanken.gu.se` accessed the 11th of Febuary 2004.

Borin, L. (2009). One in the bush: Low-density language technology. Technical report, Gothenburg: GU-ISS-09-1, Research Reports from the Department of Swedish, University of Gothenburg.

Borin, L. and Saxena, A. (2004). Grammar, incorporated. In Henriksson, P. J., editor, *CALL for the Nordic languages*, volume 30 of *Copenhagen Studies in Languages*, pages 125–146. Samfundslitteratur.

Brent, M. (1993). Minimal generative explanations: A middle ground between neurons and triggers. In *Proceedings of the fifteenth annual conference of the Cognitive Science Society: June 18 to 21, 1993, Institute of Cognitive Science, University of Colorado, Boulder*, pages 28–36. Lawrence Erlbaum Associates.

Brent, M. R. (1999). An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning*, 34:71–105.

Brent, M. R., Murthy, S., and Lundberg, A. (1995). Discovering morphemic suffixes: A case study in minimum description length induction. In *Fifth International Workshop on Artificial Intelligence and Statistics*, pages 482–490. Fort Lauderdale, Florida: Society for Artificial Intelligence and Statistics.

Brenzinger, M. (2007). Language endangerment throughout the world. In Brenzinger, M., editor, *Language Diversity Endangered*, volume 181 of *Trends in Linguistics: Studies and Monographs*, pages ix–xviii. Mouton de Gruyter.

British & Foreign Bible Society (1953). *Maandiko matakatifu ya Mungu yaitwaya Biblia, yaani Agano la kale na Agano jipya, katika lugha ya Kiswahili [Swahili Bible]*. London: British & Foreign Bible Society.

British & Foreign Bible Society (1996). *Maori Bible*. London: British & Foreign Bible Society.

Carlos, C. S., Choudhury, M., and Dandapat, S. (2009). Large-coverage root lexicon extraction for Hindi. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 121–129, Athens, Greece. Association for Computational Linguistics.

Caseiro, D. (1999). Automatic language identification bibliography. `http://www.phys.uni.torun.pl/kmk/projects/ali-bib.html` accessed the 25th of May 2005.

Ćavar, D., Herring, J., Ikuta, T., Rodrigues, P., and Schrementi, G. (2004a). On induction of morphology grammars and its role in bootstrapping. In Jäger, G., Monachesi, P., Penn, G., and Wintner, S., editors, *Proceedings of Formal Grammar 2004*, pages 47–62.

Ćavar, D., Herring, J., Ikuta, T., Rodrigues, P., and Schrementi, G. (2004b). On statistical parameter setting. In *Proceedings of the First Workshop on Psycho-computational Models of Human Language Acquisition, 28-29 August 2004, Geneva, Switzerland (Held in cooperation with COLING-2004)*, pages 9–16.

Ćavar, D., Herring, J., Ikuta, T., Rodrigues, P., and Schrementi, G. (2006). On unsupervised grammar induction from untagged corpora. In Kaszubski, P., editor, *PSiCL: Poznan' Studies in Contemporary Linguistics*, volume 41, pages 57–71. Poznan', Poland: Adam Mickiewicz University.

Ćavar, D., Rodrigues, P., and Schrementi, G. (2005). Unsupervised morphology induction for part-of-speech tagging. *U. Penn Working Papers in Linguistics*, 10(1).

Cavnar, W. B. and Trenkle, J. M. (1994). N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175, Las Vegas, US.

Chan, E. (2006). Learning probabilistic paradigms for morphology in a latent class model. In *Proceedings of the Eighth Meeting of the ACL Special Interest Group on Computational Phonology and Morphology at HLT-NAACL 2006*, pages 69–78. Association for Computational Linguistics, New York City, USA.

Cho, S. and Han, S.-S. (2002). Automatic stemming for indexing of an agglutinative language. In Yakhno, T., editor, *Advances in Information Systems*, volume 2457 of *Lecture Notes in Computer Science*, pages 154–165. Springer-Verlag, Berlin.

Clark, A. S. (2001). *Unsupervised Language Acquisition: Theory and Practice.* PhD thesis, University of Sussex.

Creutz, M. (2003). Unsupervised segmentation of words using prior distributions of morph length and frequency. In *Proceedings of the ACL 2003*, pages 280–287. Sapporo, Japan.

Creutz, M. (2006). *Induction of the Morphology of Natural Language: Unsupervised Morpheme Segmentation with Application to Automatic Speech Recognition.* PhD thesis, Helsinki University of Technology, Espoo, Finland.

Creutz, M. and Lagus, K. (2002). Unsupervised discovery of morphemes. In *Proceedings of the 6th Workshop of the ACL Special Interest Group in Computational Phonology (SIGPHON), Philadelphia, July 2002*, pages 21–30. Association for Computational Linguistics.

Creutz, M. and Lagus, K. (2004). Induction of a simple morphology for highly-inflecting languages. In *Proceedings of the 7th Meeting of the ACL Special Interest Group in Computational Phonology (SIGPHON)*, pages 43–51. Barcelona.

Creutz, M. and Lagus, K. (2005a). Inducing the morphological lexicon of a natural language from unannotated text. In *Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR '05), 15-17 June, Espoo, Finland*, pages 106–113. Espoo.

Creutz, M. and Lagus, K. (2005b). Morfessor in the Morpho Challenge. In Kurimo, M., Creutz, M., and Lagus, K., editors, *Unsupervised segmentation of words into morphemes – Challenge 2005*, pages 12–17.

Creutz, M. and Lagus, K. (2005c). Unsupervised morpheme segmentation and morphology induction from text corpora using morfessor 1.0. Technical report, Publications in Computer and Information Science, Report A81, Helsinki University of Technology.

Creutz, M. and Lagus, K. (2007). Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing*, 4(1,3):1–33.

Creutz, M., Lagus, K., Lindén, K., and Virpioja, S. (2005a). Morfessor and hutmegs: Unsupervised morpheme segmentation for highly-inflecting and compounding languages. In *Proceedings of the Second Baltic Conference on Human Language Technologies, Tallinn, 4 - 5 April*, pages 107–112. Tallinn, Estonia.

Creutz, M., Lagus, K., and Virpioja, S. (2005b). Unsupervised morphology induction using morfessor. In Yli-Jyrä, A., Karttunen, L., and Karhumäki, J., editors, *Finite State Methods in Natural Language Processing: 5th International Workshop, FSMNLP 2005, Helsinki, Finland, September 1-2, 2005. Revised Papers*, volume 4002 of *Lecture Notes in Computer Science*, pages 300–301. Springer-Verlag, Berlin.

Creutz, M. and Lindén, K. (2004). Morpheme segmentation gold standards for Finnish and English. Technical report, Publications in Computer and Information Science, Report A77, Helsinki University of Technology.

Cromm, O. (1997). Affixerkennung in deutschen wortformen: Ein nichtlexikalisches segmentierungsverfahren nach N. D. Andreev. *LDV-Forum*, 14(2):4–13.

da Silva, J. F. and Lopes, G. P. (2006a). Identification of document language is not yet a completely solved problem. In *CIMCA '06: Proceedings of the International Conference on Computational Intelligence for Modelling Control and Automation and International Conference on Intelligent Agents Web Technologies and International Commerce*, pages 212–219, Washington, DC, USA. IEEE Computer Society.

da Silva, J. F. and Lopes, J. G. P. (2006b). Identification of document language in hard contexts. In *Proceedings of the SIGIR 2006 Workshop on New Directions in Multilingual Information Access, Seattle, USA.*

Damashek, M. (1995). Gauging Similarity with n-Grams: Language-Independent Categorization of Text. *Science*, 267(5199):843–848.

Dang, M. T. and Choudri, S. (2005). Simple unsupervised morphology analysis algorithm (sumaa). In Kurimo, M., Creutz, M., and Lagus, K., editors, *Proceedings of MorphoChallenge 2005*, pages 47–51.

Dasgupta, S. (2007). Toward language-independent morphological segmentation and part-of-speech induction. Master's thesis, The University of Texas at Dallas.

Dasgupta, S. and Ng, V. (2006). Unsupervised morphological parsing of bengali. *Language Resources and Evaluation*, 3-4:311–330.

Dasgupta, S. and Ng, V. (2007). High-performance, language-independent morphological segmentation. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 155–163, Rochester, New York. Association for Computational Linguistics.

Dasgupta, S. and Ng., V. (2007). Unsupervised word segmentation for bangla. In *Proceedings of the 5th International Conference on Natural Language Processing (ICON 2007)*. Hyderabad, India.

de Kock, J. and Bossaert, W. (1969). Towards an automatic morphological segmentation. In *International conference on computational linguistics, COLING, 1-4 September 1969, Sånga-Säby, Sweden*, volume 60, pages 10–11. Stockholm: Forskningsgruppen fÃűr kvantitativ lingvistik.

de Kock, J. and Bossaert, W. (1974). *Introducción a la lingüística automática en las lenguas Románicas*, volume 202 of *Biblioteca románica hispánica 2: Estudios y ensayos*. Gredos, Madrid.

de Kock, J. and Bossaert, W. (1978). *The Morpheme: An Experiment in Quantitative and Computational Linguistics*. Van Gorcum, Amsterdam.

De Pauw, G. and Wagacha, P. W. (2007). Bootstrapping morphological analysis of Gĩkũyũ using maximum entropy learning. In *Proceedings of the 8th Annual Conference of the International Speech Communication Association (INTERSPEECH 2007), Antwerp, Belgium, August 27-31, 2007*, pages 1517–1520. ISCA.

Déjean, H. (1998a). *Concepts et algorithmes pour la découverte des structures formelles des langues*. PhD thesis, Université de Caen Basse Normandie.

Déjean, H. (1998b). Morphemes as a necessary concept for structures discovery from untagged corpora. In *NeMLaP3/CoNLL98 Workshop on Paradigms and Grounding in Language Learning*, pages 295–298. Association for Computational Linguistics, Philadephia.

Deligne, S. (1996). *Modèles de séquences de longueurs variables: application au traitement du langage écrit et de la parole*. PhD thesis, École Nationale Supérieure des Télécommunications, Paris.

Deligne, S. and Bimbot, F. (1997). Inference of variable-length linguistic and acoustic units by multigrams. *Speech Communication*, 23(3):223–241.

Demberg, V. (2007). A language-independent unsupervised model for morphological segmentation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 920–927, Prague, Czech Republic. Association for Computational Linguistics.

Dryer, M. S. (2005). Prefixing versus suffixing in inflectional morphology. In Comrie, B., Dryer, M. S., Gil, D., and Haspelmath, M., editors, *World Atlas of Language Structures*, pages 110–113. Oxford University Press.

Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern Classification*. Wiley, New York, 2 edition.

Dunning, T. (1994). Statistical identification of language. Technical report, Techical Report MCCS-94-273, Computing Research Lab (CRL), New Mexico State University.

Erjavec, T. and Džeroski, S. (2004). Machine learning of morphosyntactic structure: Lemmatizing slovene words. *Applied Artificial Intelligence*, 18:17–41.

Faulk, R. D. and Gustavson, F. G. (1990). Segmenting discrete data representing continuous speech input. *IBM Systems Journal*, 29(2):287–296.

Flenner, G. (1992). *Ein quantitatives Morphsegmentierungsverfahren für spanische Wortformen.* PhD thesis, Georg-August-Universität Göttingen.

Flenner, G. (1994). Ein quantitatives morphsegmentierungssystem für spanische wortformen. In Klenk, U., editor, *Computatio Linguae II: Aufsätze zur algorithmischen und Quantitativen Analyse der Sprache*, volume 83 of *Zeitschrift für Dialektologie und Linguistik: Beihefte*, pages 31–62. Franz Steiner, Stuttgart.

Flenner, G. (1995). Quantitative morphsegmentierung im spanischen auf phonologischer basis. *Sprache und Datenverarbeitung*, 19(2):63–78.

Forsberg, M. (2007). *Three Tools for Language Processing: BNF Converter, Functional Morphology, and Extract.* PhD thesis, Chalmers University of Technology, Gothenburg.

Forsberg, M., Hammarström, H., and Ranta, A. (2006). Lexicon extraction from raw text data. In Salakoski, T., Ginter, F., Pyysalo, S., and Pahikkala, T., editors, *Advances in Natural Language Processing: Proceedings of the 5th International Conference, FinTAL 2006 Turku, Finland, August 23-25, 2006*, volume 4139 of *Lecture Notes in Computer Science*, pages 488–499. Springer-Verlag, Berlin.

Frachtenberg, L. J. (1920). *Alsea texts and myths*, volume 67 of *Bureau of American Ethnology Bulletin.* Washington, D.C.: Smithsonian Institution.

Frakes, W. B. and Fox, C. J. (2003). Strength and similarity of affix removal stemming algorithms. *SIGIR Forum*, 37(1):26–30.

Francis, N. W. and Kucera, H. (1964). Brown corpus. Department of Linguistics, Brown University, Providence, Rhode Island. 1 million words.

Freitag, D. (2005). Morphology induction from term clusters. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 128–135, Ann Arbor, Michigan. Association for Computational Linguistics.

Galambos, L. (2004). *Multilingual Stemmer in Web Environment.* PhD thesis, Faculty of Mathematics and Physics, Charles University in Prague.

Gammon, E. (1969). Quantitative approximations to the word. In *International conference on computational linguistics, COLING, 1-4 September 1969, Sånga-Säby, Sweden*, volume 10, pages 1–28. Stockholm: Forskningsgruppen fÃűr kvantitativ lingvistik.

Gaussier, É. (1999). Unsupervised learning of derivational morphology from inflectional lexicons. In Kehler, A. and Stolcke, A., editors, *Proceedings of the workshop on Unsupervised Learning in Natural Language Processing at the 37th Annual Meeting of the Association for Computational Linguistics (ACL-1999)*, pages 24–30. Association for Computational Linguistics, Philadephia.

Gelbukh, A. F., Alexandrov, M., and Han, S.-Y. (2004). Detecting inflection patterns in natural language by minimization of morphological model. In Sanfeliu, A., Trinidad, J. F. M., and Carrasco-Ochoa, J. A., editors, *Proceedings of Progress in Pattern Recognition, Image Analysis and Applications, 9th Iberoamerican Congress on Pattern Recognition, CIARP '04*, volume 3287 of *Lecture Notes in Computer Science*, pages 432–438. Springer-Verlag, Berlin.

Gippert, J., Himmelmann, N. P., and Mosel, U., editors (2006). *Essentials of language documentation*, volume 178 of *Trends in linguistics: Studies and Monographs*. Mouton de Gruyter.

Golcher, F. (2006). Statistical text segmentation with partial structure analysis. In *Proceedings of KONVENS 2006*, pages 44–51. Universität Konstanz.

Goldsmith, J. (2000). Linguistica: An automatic morphological analyzer. In Okrent, A. and Boyle, J., editors, *Proceedings from the Main Session of the Chicago Linguistic Society's thirty-sixth Meeting*, pages 125–139. Chicago Linguistics Society, Chicago.

Goldsmith, J. (2001). Unsupervised learning of the morphology of natural language. *Computational Linguistics*, 27(2):153–198.

Goldsmith, J., Higgins, D., and Soglasnova, S. (2001). Automatic language-specific stemming in information retrieval. In Peters, C., editor, *Cross-Language Information Retrieval and Evaluation: Proceedings of the CLEF 2000 Workshop*, Lecture Notes in Computer Science, pages 273–283. Springer-Verlag, Berlin.

Goldsmith, J., Hu, Y., Matveeva, I., and Sprague, C. (2005). A heuristic for morpheme discovery based on string edit distance. Techincal Report of Computer Science Department, University of Chicago.

Goldsmith, J. A. (2006). An algorithm for the unsupervised learning of morphology. *Natural Language Engineering*, 12(4):353–371.

Goldsmith, J. A. ([to appear]). Segmentation and morphology. In Clark, A., Fox, C., and Lappin, S., editors, *Handbook of Computational Linguistics and Natural Language Processing*. Oxford: Blackwell.

Goldwater, S. (2007). *Nonparametric Bayesian Models of Lexical Acquisition*. PhD thesis, Brown University.

Goldwater, S., Griffiths, T., and Johnson, M. (2005). Interpolating between types and tokens by estimating power-law generators. In *Advances in Neural Information Processing Systems 18 [Neural Information Processing Systems, NIPS 2005, December 5-8, 2005, Vancouver, British Columbia, Canada]*.

Goodman, S. A. (2008). Morphological induction through linguistic productivity. In *Working Notes for the CLEF 2008 Workshop, 17-19 September, Aarhus, Denmark*.

Gordon, A. (1999). *Classification*, volume 82 of *Monographs on Statistics and Applied Probability*. CRC Press, 2 edition.

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

Grefenstette, G. (1995). Comparing two language identification schemes. In Bolasco, S., Lebart, L., and Salem, A., editors, *The proceedings of 3rd International Conference on Statistical Analysis of Textual Data (JADT 95), Rome, Italy, Dec. 1995*.

Grünwald, P. D. (2007). *The minimum description length principle*. Adaptive computation and machine learning. Cambridge, Massachusetts: MIT Press.

Hadouche, F. (2002). Détection de relations morphologiques en corpus basée sur les cooccurrences. Master's thesis, DESS, Centre de Recherche en Ingénierie Multilingue, CRIM, France.

Hafer, M. A. and Weiss, S. F. (1974). Word segmentation by letter successor varieties. *Information and Storge Retrieval*, 10:371–385.

Hammarström, H. (2005). A new algorithm for unsupervised induction of concatenative morphology. In Yli-Jyrä, A., Karttunen, L., and Karhumäki, J., editors, *Finite State Methods in Natural Language Processing: 5th International Workshop, FSMNLP 2005, Helsinki, Finland, September 1-2, 2005. Revised Papers*, volume 4002 of *Lecture Notes in Computer Science*, pages 288–289. Springer-Verlag, Berlin.

Hammarström, H. (2006a). A naive theory of morphology and an algorithm for extraction. In Wicentowski, R. and Kondrak, G., 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.

Hammarström, H. (2006b). Poor man's stemming: Unsupervised recognition of same-stem words. In Ng, H. T., Leong, M.-K., Kan, M.-Y., and Ji, D., 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.

Hammarström, H. (2007a). A survey and classification of methods for (mostly) unsupervised learning of morphology. In *NODALIDA 2007, the 16th Nordic Conference of Computational Linguistics, Tartu, Estonia, 25-26 May 2007*. NEALT.

Hammarström, H. (2007b). 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.

Hammarström, H. (2009a). Poor man's word-segmentation: Unsupervised morphological analysis for Indonesian. In *Proceedings of the Third International Workshop on Malay and Indonesian Language Engineering (MALINDO)*. Singapore: ACL.

Hammarström, H. (2009b). A survey of computational morphological resources for low-density languages. Submitted.

Harris, Z. S. (1955). From phoneme to morpheme. *Language*, 31(2):190–222.

Harris, Z. S. (1968). Recurrent dependence process: Morphemes by phoneme neighbours. In *Mathematical structures of language*, volume 21 of *Interscience tracts in pure and applied mathematics*, pages 24–28. Interscience, New York.

Harris, Z. S. (1970). Morpheme boundaries within words: Report on a computer test. In Harris, Z. S., editor, *Papers in Structural and Transformational Linguistics*, volume 1 of *Formal Linguistics Series*, pages 68–77. D. Reidel, Dordrecht. Original Zellig Harris 1967 Morpheme boundaries within words: Report on a computer test. In Transformations and Discourse Analysis Papers 73. Department of Linguistics, University of Pennsylvania.

Henson, B. (2007). *The Phonology and Morphosyntax of Kol.* PhD thesis, University of California at Berkeley.

Hershberger, H. D. and Hershberger, R. (1982). *Kuku-Yalanji dictionary*, volume 7 of *Work Papers of SIL - AAB. Series B.* Summer Institute of Linguistics, Darwin.

Hirsimäki, T., Creutz, M., Siivola, V., and Kurimo, M. (2003). Unlimited vocabulary speech recognition based on morphs discovered in an unsupervised manner. In *Proceedings of Eurospeech 2003, Geneva*, pages 2293–2996. Geneva, Switzerland.

Hu, Y., Matveeva, I., Goldsmith, J., and Sprague, C. (2005a). Refining the SED heuristic for morpheme discovery: Another look at Swahili. In *Proceedings of the Workshop on Psychocomputational Models of Human Language Acquisition*, pages 28–35, Ann Arbor, Michigan. Association for Computational Linguistics.

Hu, Y., Matveeva, I., Goldsmith, J., and Sprague, C. (2005b). Using morphology and syntax together in unsupervised learning. In *Proceedings of the Workshop on Psychocomputational Models of Human Language Acquisition*, pages 20–27, Ann Arbor, Michigan. Association for Computational Linguistics.

Hughes, B., Baldwin, T., Bird, S., Nicholson, J., and MacKinlay, A. (2006). Reconsidering language identification for written language resources. In *Proceedings 5th International Conference on Language Resources and Evaluation (LREC2006)*, pages 485–488. European Language Resources Association (ELRA).

Hull, D. A. (1996). Stemming algorithms: A case study for detailed evaluation. *Journal of the American Soicety for Information Science*, 47(1):70–84.

Hyuk-Cha, S., Yoon, S., and Tappert, C. C. (2005). Enhancing binary feature vector similarity measures. Technical report, Technical Report 210, School of Science and Information Systems, Pace University.

Indradjaja, L. S. and Bressan, S. (2003). Automatic learning of stemming rules for the Indonesian language. In *Language, Information and Computation: Proceedings of the 17th Pacific Asia Conference, 1-3 October, 2003, Sentosa, Singapore*, pages 62–68. COLIPS.

Jacquemin, C. (1997). Guessing morphology from terms and corpora. In *Proceedings, 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '97), Philadelphia, PA*, pages 155–165.

Janßen, A. (1992). Segmentierung französischer wortformen ohne lexikon. In Klenk, U., editor, *Computatio Linguae: Aufsätze zur algorithmischen und Quantitativen Analyse der Sprache*, volume 73 of *Zeitschrift für Dialektologie und Linguistik: Beihefte*, pages 74–95. Franz Steiner, Stuttgart.

Johnsen, L. G. (2005). Morphological learning as principled argument. In Kurimo, M., Creutz, M., and Lagus, K., editors, *Proceedings of MorphoChallenge 2005*, pages 33–36.

Johnson, H. and Martin, J. (2003). Unsupervised learning of morphology for English and Inuktitut. In *HLT-NAACL 2003, Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, May 27 - June 1, Edmonton, Canada*, volume Companion Volume - Short papers, pages 43–45.

Johnson, M. (2008). Unsupervised word segmentation for Sesotho using adaptor grammars. In *Proceedings of the Tenth Meeting of ACL Special Interest Group on Computational Morphology and Phonology*, pages 20–27, Columbus, Ohio. Association for Computational Linguistics.

Jordan, C., Healy, J., and Keselj, V. (2005). Swordfish: Using ngrams in an unsupervised approach to morphological analysis. In Kurimo, M., Creutz, M., and Lagus, K., editors, *Proceedings of MorphoChallenge 2005*, pages 42–46.

Jordan, C., Healy, J., and Keselj, V. (2006). Swordfish: an unsupervised ngram based approach to morphological analysis. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 657–658, New York, NY, USA. ACM.

Juola, P. (2006). Language identification, automatic. In Brown, K., editor, *Encyclopedia of Language and Linguistics*, volume 6, pages 508–510. Elsevier, Amsterdam, 2 edition.

Juola, P., Hall, C., and Boggs, A. (1994). Corpus-based morphological segmentation by entropy changes. In Monaghan, A., editor, *Third Conference on the Cognitive Science of Natural Language Processing*. Dublin City University.

Katrenko, S. (2004). Towards unsupervised learning of morphology applied to Ukrainian. In i Alemany, L. A. and Egré, P., editors, *Student Session: 16th European Summer School in Logic, Language and Information, Nancy, France, 9-20 August, 2004*, pages 138–148. FoLLI.

Kazakov, D. (1997). Unsupervised learning of naïve morphology with genetic algorithms. In Daelemans, W., Weijters, T., and van der Bosch, A., editors, *ECML'97 – Workshop Notes on Empirical Learning of Natural Language Tasks*, pages 105–112, Prague. University of Economics.

Kazakov, D. and Manandhar, S. (1998). A hybrid approach to word segmentation. In Page, C. D., editor, *Proceedings of the 8th International Workshop on Inductive Logic Programming (ILP-98) in Madison, Wisconsin, USA*, volume 1446 of *Lecture Notes in Artificial Intelligence*, pages 125–134. Springer-Verlag, Berlin.

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

Keshava, S. and Pitler, E. (2005). A simpler, intuitive approach to morpheme induction. In Kurimo, M., Creutz, M., and Lagus, K., editors, *Proceedings of MorphoChallenge 2005*, pages 28–32.

King James (1977). *The Holy Bible, containing the Old and New Testaments and the Apocrypha in the authorized King James version*. Nashville, New York: Thomas Nelson.

Klenk, U. (1991). Verfahren der segmentierung von wörtern in morphe: Mit einer untersuchung zum Spanischen. In und Dieter Seelbach, J. R., editor, *Romanistische Computerlinguistik: Theorien und Implementationen*, volume 266 of *Linguistische Arbeiten*, pages 197–206. Niemeyer, Tübingen.

Klenk, U. (1992). Verfahren morphologischer segmentierung und die wortstruktur des Spanischen. In Klenk, U., editor, *Computatio Linguae: Aufsätze zur algorithmischen und Quantitativen Analyse der Sprache*, volume 73 of *Zeitschrift für Dialektologie und Linguistik: Beihefte*, pages 110–124. Franz Steiner, Stuttgart.

Klenk, U. and Langer, H. (1989). Morphological segmentation without a lexicon. *Literary and Linguistic Computing*, 4(4):247–253.

Kohonen, O., Virpioja, S., and Klami, M. (2008). Allomorfessor: Towards unsupervised morpheme analysis. In *Working Notes for the CLEF 2008 Workshop, 17-19 September, Aarhus, Denmark.*

Kontorovich, L., Don, D., and Singer, Y. (2003). A markov model for the acquisition of morphological structure. Technical report, CMU-CS-03-147, School of Computer Science, Carnegie Mellon University.

Kruengkrai, C., Srichaivattana, P. andSornlertlamvanich, V., and Isahara, H. (2005). Language identification based on string kernels. In *IEEE International Symposium on Communications and Information Technology, 2005. ISCIT 2005*, volume 2, pages 926–929.

Kurimo, M., Creutz, M., and Turunen, V. (2007a). Overview of Morpho Challenge in CLEF 2007. In Nardi, A. and Peters, C., editors, *Working Notes for the CLEF 2007 Workshop, 19-21 September, Budapest, Hungary.*

Kurimo, M., Creutz, M., and Turunen, V. (2007b). Unsupervised morpheme analysis evaluation by IR experiments – Morpho Challenge 2007. In Nardi, A. and Peters, C., editors, *Working Notes for the CLEF 2007 Workshop, 19-21 September, Budapest, Hungary.*

Kurimo, M., Creutz, M., and Varjokallio, M. (2007c). Unsupervised morpheme analysis evaluation by a comparison to a linguistic gold standard – Morpho Challenge 2007. In Nardi, A. and Peters, C., editors, *Working Notes for the CLEF 2007 Workshop, 19-21 September, Budapest, Hungary.*

Kurimo, M., Creutz, M., and Varjokallio, M. (2008). Morpho Challenge evaluation using a linguistic gold standard. In Peters, C., Jijkoun, V., Mandl, T., Müller, H., Oard, D. W., and Penas, A., editors, *Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers*, volume 5152 of *Lecture Notes in Computer Science*, pages 864–872. Springer-Verlag, Berlin, Heidelberg.

Kurimo, M. and Turunen, V. (2008). Unsupervised morpheme analysis evaluation by IR experiments – Morpho Challenge 2008. In *Working Notes for the CLEF 2008 Workshop, 17-19 September, Aarhus, Denmark.*

Kurimo, M. and Varjokallio, M. (2008). Unsupervised morpheme analysis evaluation by a comparison to a linguistic gold standard – Morpho Challenge 2008. In *Working Notes for the CLEF 2008 Workshop, 17-19 September, Aarhus, Denmark.*

Ladefoged, P. (2005). *Vowels and Consonants.* Oxford: Blackwell, 2 edition.

Langer, H. (1991). *Ein automatisches Morphsegmentierungsverfahren für deutsche Wortformen.* PhD thesis, Georg-August-Universität zu Göttingen.

Lefebvre, C. (2004). *Issues in the study of Pidgin and Creole languages*, volume 70 of *Studies in Language Companion Series.* Amsterdam: John Benjamins.

Lehmann, H. (1973). *Linguistische Modellbildung und Methodologie.* Max Niemeyer Verlag, Tübingen.

Leizarraga, J. (1571). *Iesus Krist Gure Iaunaren Testamentu Berria [New Testament in Basque].* Roxellan: Pierre Hautin, Inprimizale.

Lindén, K. (2008). A probabilistic model for guessing base forms of new words by analogy. In Gelbukh, A. F., editor, *Proceedings of CICLing-2008: 9th International Conference on Intelligent Text Processing and Computational Linguistics*, volume 4919 of *Lecture Notes in Computer Science*, pages 106–116. Springer.

Lins, R. D. and Gonçalves, Jr., P. (2004). Automatic language identification of written texts. In *SAC '04: Proceedings of the 2004 ACM symposium on Applied computing*, pages 1128–1133, New York, NY, USA. ACM Press.

Maho, J. (2003). A classification of the Bantu languages: An update of Guthrie's referential system. In Nurse, D. and Philippson, G., editors, *The Bantu Languages*, Routledge Language Family Series, pages 639–651. London & New York: Routledge.

Majumder, P., Mitra, M., and Pal, D. (2008). Bulgarian, Hungarian and Czech stemming using YASS. In Peters, C., Jijkoun, V., Mandl, T., Müller, H., Oard, D. W., and Penas, A., editors, *Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers*, pages 49–56. Springer-Verlag, Berlin.

Majumder, P., Mitra, M., Parui, S. K., Kole, G., Mitra, P., and Datta, K. (2007). YASS: Yet another suffix stripper. *ACM Transactions on Information Systems*, 25(4):18:1–20.

Martin, T., Baker, B., Wong, E., and Sridharan, S. (2006). A syllable-scale framework for language identification. *Computer Speech & Language*, 20(2-3):276–302.

Martins, B. and Silva, M. J. (2005). Language identification in web pages. In *SAC '05: Proceedings of the 2005 ACM symposium on Applied computing*, pages 764–768, New York, NY, USA. ACM Press.

Mayfield, J. and McNamee, P. (2003). Single n-gram stemming. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 415–416, New York, NY, USA. ACM.

McNamee, P. (2005). Language identification: a solved problem suitable for undergraduate instruction. *Journal of Computing Sciences in Colleges*, 20(3):94–101.

McNamee, P. (2008). Retrieval experiments at Morpho Challenge 2008. In *Working Notes for the CLEF 2008 Workshop, 17-19 September, Aarhus, Denmark.*

McNamee, P. and Mayfield, J. (2007). N-gram morphemes for retrieval. In Nardi, A. and Peters, C., editors, *Working Notes for the CLEF 2007 Workshop, 19-21 September, Budapest, Hungary.*

Medina Urrea, A. (2000). Automatic discovery of affixes by means of a corpus: A catalog of Spanish affixes. *Journal of Quantitative Linguistics*, 7(2):97–114.

Medina Urrea, A. (2003). *Investigación cuantitativa de afijos y clíticos del español de México: Glutinometría en el* Corpus del Español Mexicano Contemporáneo. PhD thesis, El Colegio de México, México, D.F.

Medina-Urrea, A. (2006). Affix discovery by means of corpora: Experiments for Spanish, Czech, Ralámuli and Chuj. In Mehler, A. and Köhler, R., editors, *Aspects of Automatic Text Analysis*, volume 209 of *Studies in Fuzziness and Soft Computing*, pages 277–299. Springer, Berlin.

Medina Urrea, A. (2006). Towards the automatic lemmatization of 16th century Mexican Spanish: A stemming scheme for the CHEM. In Gelbukh, A. F., editor, *Computational Linguistics and Intelligent Text Processing, 7th International Conference, CICLing 2006, Mexico City, Mexico, February 19-25, 2006, Proceedings*, volume 3878 of *Lecture Notes in Computer Science*, pages 101–104. Springer-Verlag, Berlin.

Medina-Urrea, A. (2008). Affix discovery based on entropy and economy measurements. In Gaylord, N., Palmer, A., and Ponvert, E., editors, *Computational Linguistics for Less-Studied Languages*, volume X of *Texas Linguistics Society*, pages 99–112. Stanford: CSLI.

Medina Urrea, A. and Díaz, E. C. B. (2003). Características cuantitativas de la flexión verbal del Chuj. *Estudios de Lingüística Aplicada*, 38:15–31.

Melucci, M. and Orio, N. (2003). A novel method for stemmer generation based on Hidden Markov Models. In *CIKM '03: Proceedings of the twelfth international conference on Information and knowledge management*, pages 131–138, New York, NY, USA. ACM Press.

Monson, C. (2004). A framework for unsupervised natural language morphology induction. In van der Beek, L. and andDaniel Midgley, D. G., editors, *ACL 2004: Student Research Workshop*, pages 67–72, Barcelona, Spain. Association for Computational Linguistics.

Monson, C. (2009). *ParaMor: From paradigm structure to natural language morphology induction.* PhD thesis, Carnegie Mellon University.

Monson, C., Carbonell, J., Lavie, A., and Levin, L. (2007a). ParaMor: Finding paradigms across morphology. In Nardi, A. and Peters, C., editors, *Working Notes for the CLEF 2007 Workshop, 19-21 September, Budapest, Hungary.*

Monson, C., Carbonell, J., Lavie, A., and Levin, L. (2007b). ParaMor: Minimally supervised induction of paradigm structure and morphological analysis. In *Proceedings of Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology*, pages 117–125, Prague, Czech Republic. Association for Computational Linguistics.

Monson, C., Carbonell, J., Lavie, A., and Levin, L. (2008a). ParaMor and Morpho Challenge 2008. In *Working Notes for the CLEF 2008 Workshop, 17-19 September, Aarhus, Denmark.*

Monson, C., Lavie, A., Carbonell, J., and Levin, L. (2004). Unsupervised induction of natural language morphology inflection classes. In *SIGPHON 2004: Proceedings of the Seventh Meeting of the ACL Special Interest Group in Computational Phonology*, pages 52–61, Barcelona, Spain. Association for Computational Linguistics.

Monson, C., Lavie, A., Carbonell, J., and Levin, L. (2008b). Evaluating an agglutinative segmentation model for ParaMor. In *Proceedings of the Tenth Meeting of ACL Special Interest Group on Computational Morphology and Phonology*, pages 49–58, Columbus, Ohio. Association for Computational Linguistics.

Monson, C., Llitjós, A. F., Ambati, V., Levin, L., Lavie, A., Alvarez, A., Aranovich, R., Carbonell, J., Frederking, R., Peterson, E., and Probst, K. (2008c). Linguistic structure and bilingual informants help induce machine translation of lesser-resourced languages. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, pages 2854–2859. Marrakech, Morocco.

Moon, T., Erk, K., and Baldridge, J. (2009). Unsupervised morphological segmentation and clustering with document boundaries. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 668–677, Singapore. Association for Computational Linguistics.

Murthy, K. N. and Kumar, G. B. (2006). Language identification from small text samples. *Journal of Quantitative Linguistics*, 13(1):57–80.

Muthusamy, Y. K. and Spitz, L. A. (1997). Automatic language identification. In Cole, R. A., editor, *Survey of the State of the Art in Human Language Technology*, chapter 8.7. Center for Spoken Language Understanding CSLU, Carnegie Mellon University, Pittsburgh, PA.

Naradowsky, J. and Goldwater, S. (2009). Improving morphology induction by learning spelling rules. In *International Joint Conference on Artificial Intelligence*.

Nash, D. G. (1980). *Topics in Warlpiri Grammar*. PhD thesis, Massachusetts Institute of Technology. Also published by Garland 1986.

No Author Stated (2006a). *Godokono Hido Tabo: Aramia River Tabo Testament*. Port Moresby: Bible Society of Papua New Guinea.

No Author Stated (2006b). *Godokono Wade Tabo: Fly River Tabo New Testament*. Port Moresby: Bible Society of Papua New Guinea.

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

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

Nunzio, G. D., Ferro, N., Melucci, M., and Orio, N. (2004). Experiments to evaluate probabilistic models for automatic stemmer generation and query word translation. In Peters, C., Braschler, M., Gonzalo, J., and Kluck, M., editors, *Proceedings of the Cross-Language Evaluation Forum (CLEF): Methodology and Metrics (CLEF 2003)*, volume 3237 of *Lecture Notes in Computer Science*, pages 220–235. Springer-Verlag, Berlin.

Oliver, A. (2004). *Adquisició d'informació lèxica i morfosintàctica a partir de corpus sense anotar: aplicació al rus i al croat*. PhD thesis, Universitat de Barcelona.

Pandey, A. K. and Siddiqui, T. J. (2008). An unsupervised Hindi stemmer with heuristic improvements. In *AND '08: Proceedings of the second workshop on Analytics for noisy unstructured text data*, pages 99–105, New York, NY, USA. ACM.

Patz, E. (2002). *A Grammar of the Kuku Yalanji Language of North Queensland*, volume 257 of *Pacific Linguistics*. Canberra: Research School of Pacific and Asian Studies, Australian National University.

Pereira, F. C. N., Tishby, N., and Lee, L. (1993). Distributional clustering of english words. In *31st Meeting of the Association for Computational Linguistics*, pages 183–190.

Pirkola, A. (2001). Morphological typology of languages for IR. *Journal of Documentation*, 57(3):330–348.

Pirrelli, V., Calderone, B., Herreros, I., and Virgilio, M. (2004). Non-locality all the way through: Emergent global constraints in the italian morphological lexicon. In *SIGPHON 2004: Proceedings of the Seventh Meeting of the ACL Special Interest Group in Computational Phonology*, pages 52–61, Barcelona, Spain. Association for Computational Linguistics.

Pirrelli, V. and Herreros, I. (2007). Learning morphology by itself. In Booij, G., Ducceschi, L., Fradin, B., Guevara, E., Ralli, A., and Scalise, S., editors, *Proceedings of the Fifth Mediterranean Morphology Meeting (MMM5) Fréjus 15-18 September 2005*, pages 269–290. Università degli Studi di Bologna.

Pisceldo, F., Mahendra, R., Manurung, R., and Arka, I. W. (2008). A two-level morphological analyser for the Indonesian language. In *Proceedings of the 2008 Australasian Language Technology Association Workshop (ALTA 2008)*, pages 142–150. Hobart, Australia.

Poon, H., Cherry, C., and Toutanova, K. (2009). Unsupervised morphological segmentation with log-linear models. In *Proceedings of NAACL '09: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 209–217, Morristown, NJ, USA. Association for Computational Linguistics.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 14(3):130–137.

Poutsma, A. (2002). Applying Monte Carlo techniques to language identification. In Mariët, T., Nijholt, A., and Hondorp, H., editors, *Computational Linguistics in the Netherlands 2001: Selected Papers from the Twelfth CLIN Meeting*, volume 45 of *Language and Computers - Studies in Practical Linguistics*, pages 179–189. Rodopi, Amsterdam/New York, NY.

Powers, D. M. W. (1998). Reconciliation of unsupervised clustering, segmentation and cohesion. In *NeMLaP3/CoNLL '98 Workshop on Paradigms and Grounding in Language Learning*, pages 307–310. ACL.

Prager, J. M. (2000). Linguini: Language identification for multilingual documents. *Journal of Management Information Systems*, 16(3):71–102.

Ranaivo, B. (2001). *Reconnaissance automatique de l'affixation en Malais.* PhD thesis, Institut National des Langues et Civilisations Orientales, Paris, France.

Ranaivo-Malançon, B. (2004). Computational analysis of affixed words in Malay language. UTMK, USM, Malaysia.

Redington, M., Chater, N., and Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science*, 22(4):425–469.

Redlich, A. N. (1993). Redundancy reduction as a strategy for unsupervised learning. *Neural Computation*, 5(2):289–304.

Roark, B. and Sproat, R. W. (2007). Machine learning of morphology. In *Computational approaches to morphology and syntax*, volume 4 of *Oxford surveys in syntax and morphology*, pages 116–136. Oxford University Press.

Rodrigues, P. and Ćavar, D. (2005). Learning arabic morphology using information theory. *Proceedings from the Annual Meeting of the Chicago Linguistic Society*, 41(2):49–58.

Rodrigues, P. and Ćavar, D. (2007). Learning arabic morphology using statistical constraint-satisfaction models. In Benmamoun, E., editor, *Perspectives on Arabic Linguistics: Papers from the annual symposium on Arabic Linguistics Volume XIX: Urbana, Illinois, April 2005*, volume 289 of *Current Issues in Linguistic Theory*, pages 63–75. Amsterdam: John Benjamins.

Rogati, M., McCarley, S., and Yang, Y. (2003). Unsupervised learning of arabic stemming using a parallel corpus. In *ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 391–398, Morristown, NJ, USA. Association for Computational Linguistics.

Rosenthal, F. (1995). *A grammar of biblical Aramaic*, volume 5 of *Porta linguarum Orientalium*. Wiesbaden: Harrassowitz, 6 edition.

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

Sanders, G. (1990). On the analysis and implications of Maori verb alternations. *Lingua*, 80:149–196.

Saxena, A. and Borin, L., 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.

Schone, P. (2001). *Toward Knowledge-Free Induction of Machine-Readable Dictionaries*. PhD thesis, University of Colorado.

Schone, P. and Jurafsky, D. (2000). Knowledge-free induction of inflectional morphologies using latent semantic analysis. In *Conference on Natural Language Learning 2000 (CoNLL-2000), Lisbon, Portugal*.

Schone, P. and Jurafsky, D. (2001a). Knowledge-free induction of inflectional morphologies. In *Proceedings of the North American Chapter of the Association for Computational Linguistics, Pittsburgh, PA, 2001*, pages 183–191.

Schone, P. and Jurafsky, D. (2001b). Language-independent induction of part of speech class labels using only language universals. In *"Machine Learning: Beyond Supervision", Workshop at IJCAI-2001, Seattle, WA, August 2001*.

Schütze, H. (1993). Part-of-speech induction from scratch. In *Proceedings of the 31st conference on Association for Computational Linguistics*, pages 251–258, Morristown, NJ, USA. Association for Computational Linguistics.

Sharma, U. and Das, R. (2002). Classification of words based on affix evidence. In Sangal, R. and Bendre, S. M., editors, *International Conference on Natural Language Processing, ICON-2002, Mumbai, December 18-21, 2002*, pages 31–39. Vikas Publishing House Pvt Ltd., New Delhi.

Sharma, U., Kalita, J., and Das, R. (2002). Unsupervised learning of morphology for building lexicon for a highly inflectional language. In *Proceedings of the 6th Workshop of the ACL Special Interest Group in Computational Phonology (SIGPHON), Philadelphia, July 2002*, pages 1–10. Association for Computational Linguistics.

Sharma, U., Kalita, J., and Das, R. (2003). Root word stemming by multiple evidence from corpus. In *Proceedings of the 6th International Conference on Computational Intelligence and Natural Computation (CINC), Cary, North Carolina, September 2003*, pages 1593–1596.

Sibun, P. and Reynar, J. C. (1996). Language identification: Examining the issues. In *5th Symposium on Document Analysis and Information Retrieval*, pages 125–135, Las Vegas, Nevada, U.S.A.

Snover, M. G. (2002). An unsupervised knowledge free algorithm for the learning of morphology in natural languages. Master's thesis, Department of Computer Science, Washington University.

Snover, M. G. and Brent, M. R. (2001). A bayesian model for morpheme and paradigm identification. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL-2001)*, pages 482–490. Morgan Kaufmann Publishers.

Snover, M. G. and Brent, M. R. (2003). A probabilistic model for learning concatenative morphology. In Becker, S., Thrun, S., and Obermayer, K., editors, *Advances in Neural Information Processing Systems 15*, pages 1513–1520. MIT Press, Cambridge, MA.

Snover, M. G., Jarosz, G. E., and Brent, M. R. (2002). Unsupervised learning of morphology using a novel directed search algorithm: Taking the first step. In *Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning*, pages 11–20. Association for Computational Linguistics.

Snyder, B. and Barzilay, R. (2008). Unsupervised multilingual learning for morphological segmentation. In *Proceedings of ACL-08: HLT*, pages 737–745, Columbus, Ohio. Association for Computational Linguistics.

Spiegler, S., Golénia, B., Shalonova, K., Flach, P., and Tucker, R. (2008). Learning the morphology of Zulu with different degrees of supervision. In *Spoken Language Technology Workshop, 2008 (SLT 2008)*, pages 9–12. IEEE.

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

Summer Institute of Linguistics (1985). Bible: New testament and old testament selctions in kuku-yalanji.

Summer Institute of Linguistics (2001). *Bible: selections in Warlpiri*. Canberra: Document 0650 of the Aboriginal Studies Electronic Data Archive (ASEDA), AIATSIS (Australian Institute of Aboriginal and Torres Strait Islander Studies).

Svenska Bibelsällskapet (1917). *Gamla och Nya testamentet: de kanoniska böckerna [Swedish Bible]*. Stockholm: Norstedt.

Tai, S. Y., Ong, C. S., and Abullah, N. A. (2000). On designing an automated Malaysian stemmer for the Malay language. In *IRAL '00: Proceedings of the fifth international workshop on on Information retrieval with Asian languages*, pages 207–208, New York, NY, USA. ACM.

Takci, H. and Sogukpinar, I. (2004). Centroid-based language identification using letter feature set. In Gelbukh, A., editor, *Computational Linguistics and Intelligent Text Processing: 5th International Conference, CICLing 2004 Seoul, Korea, February 15-21, 2004 Proceedings*, volume 2945 of *Lecture Notes in Computer Science*, pages 640–648. Springer-Verlag, Berlin.

Tepper, M. (2007). Knowledge-lite induction of underlying morphology: A hybrid approach to learning morphemes using context-sensitive rewrite rules. Master's thesis, University of Washington.

Tepper, M. and Xia, F. (2008). A hybrid approach to the induction of underlying morphology. In *Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP 2008)*, pages 17–24, Hyderabad, India. Asian Federation of Natural Language Processing.

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

Thornell, C. (2004a). Minioritetsspråket Mpiemos sociolingvistiska kontext. *Africa & Asia*, 5:167–191.

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

Thornell, C. (2008). *"Boulettes de graines de courge, pêche, hospitalité ...": Enregistrements transcrits et annotées pour une documentation du Mpiemo (langue Bantoue de la République Centrafricaine et du Cameroun)*, volume 25 of *Wortkunst und Dokumentartexte in afrikanischen Sprachen*. Köln: Rüdiger Köppe.

Thornell, C. and Nagano-Madsen, Y. (2004). Preliminaries to the phonetic structure of the Bantu language Mpiemo. *Africa & Asia: Göteborg working papers on Asian and African languages and literatures*, 4:163–180.

Traill, A. (1994). *A !Xóõ Dictionary*, volume 9 of *Quellen zur Khoisan-Forschung/Research in Khoisan Studies*. Köln: Rüdiger Köppe.

ur Rehman, K. and Hussain, I. (2005). Unsupervised morphemes segmentation. In Kurimo, M., Creutz, M., and Lagus, K., editors, *Proceedings of MorphoChallenge 2005*, pages 52–56.

van der Voort, H. (2007). Theoretical and social implications of language documentation and description on the eve of destruction in Rondônia. In Austin, P. K., Bond, O., and Nathan, D., editors, *Proceedings of Conference on Language Documentation and Linguistic Theory*, pages 251–259. London: SOAS.

Wicentowski, R. (2002). *Modeling and Learning Multilingual Inflectional Morphology in a Minimally Supervised Framework*. PhD thesis, Johns Hopkins University, Baltimore, MD.

Wicentowski, R. (2004). Multilingual noise-robust supervised morphological analysis using the wordframe model. In *Proceedings of the ACL Special Interest Group on Computational Phonology (SIGPHON)*, pages 70–77.

Williams, H. W. (1971). *A dictionary of the Maori language*. Wellington: GP Books, 7 edition.

Xafopoulos, A., Kotropoulos, C., Almpanidis, G., and Pitas, I. (2004). Language identification in web documents using discrete HMMs. *Pattern Recognition*, 37(3):583–594.

Xanthos, A. (2007). *Apprentissage automatique de la morphologie: Le cas des structures racine-schème*. PhD thesis, Université de Lausanne. Published 2008 by Peter Lang AG (Sciences pour la Communication 88).

Xanthos, A., Hu, Y., and Goldsmith, J. (2006). Exploring variant definitions of pointer length in mdl. In *Proceedings of the Eighth Meeting of the ACL Special Interest Group on Computational Phonology and Morphology at HLT-NAACL 2006*, pages 32–40. Association for Computational Linguistics, New York City, USA.

Yarowsky, D. and Wicentowski, R. (2000). Minimally supervised morphological analysis by multimodal alignment. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL-2000)*, pages 207–216.

Yvon, F. (1996). *Prononcer par analogie: motivation, formalisation et evaluation*. PhD thesis, École Nationale Supérieure des Télécommunications, Paris.

Zeman, D. (2007). Unsupervised acquiring of morphological paradigms from tokenized text. In Nardi, A. and Peters, C., editors, *Working Notes for the CLEF 2007 Workshop, 19-21 September, Budapest, Hungary.*

Zeman, D. (2008a). Unsupervised acquiring of morphological paradigms from tokenized text. In Peters, C., Jijkoun, V., Mandl, T., Müller, H., Oard, D. W., and Penas, A., editors, *Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers*, pages 892–899. Springer-Verlag, Berlin.

Zeman, D. (2008b). Using unsupervised paradigm acquisition for prefixes. In *Working Notes for the CLEF 2008 Workshop, 17-19 September, Aarhus, Denmark.*

Ziegler, D.-V. (1991). *The Automatic Identification of Languages Using Linguistic Recognition Signals.* PhD thesis, University of New York at Buffalo.

Zweigenbaum, P., Hadouche, F., and Grabar, N. (2003). Apprentissage de relations morphologiques en corpus. In Daille, B., editor, *Actes de TALN 2003*, pages 285–294. Batz-sur-mer, France.

Chapter II

A Survey of Computational Morphological Resources for Low-Density Languages

Hammarström, H. (2009). A Survey of Computational Morphological Resources for Low-Density Languages *Submitted.*

A Survey of Computational Morphological Resources for Low-Density Languages

Harald Hammarström
Department of Computer Science and Engineering
Chalmers University of Technology
and University of Gothenburg
SE-412 96 Göteborg, Sweden
harald2@chalmers.se

Abstract

The present paper is a survey of published work on computational morphological resources for so-called low-density languages. To avoid confusion, we introduce the term low-affluence language for a language whose speakers have low economic power. We assert empirically the appreciable observation that computational morphological resources emerge for languages with higher affluence, and we can now also account for the manner in which this happens and for the exceptions to the rule. The survey collects a wide array of work on computational morphology for low-density languages and as such outlines cross-linguistic trends in computational morphology.

1 Introduction

The present paper looks at computational morphological resources for so-called low-density languages. A survey of this kind is relevant for understanding how language resources come about [for low-density languages], which in turn is relevant for the broader questions of language survival [of low-density languages]. It is also relevant for the design and application of unsupervised NLP tools, which potentially offer their highest added value for low-density languages. These issues have been discussed in more detail in several works (Yli-Jyrä 2005, Streiter et al. 2006, Berment 2004, David and Maxwell 2008, Sornlertlamvanich 2008, Singh 2008) but there is so far no empirically grounded understanding of how language resources for low-density languages emerge. For this reason, in this paper we focus on a comprehensive survey, with the aim of a better understanding of the past and future dynamics of language resources.

The term low-density languages is increasingly employed to denote a set of languages distinct from the most powerful and affluent languages. To some,

to have a low density of language resources is the defining property of being a low-density language. In this view, a language which receives a fair amount of resources automatically ceases to be a low-density language. To others, low density languages are a class which are underprivileged vis-a-vis the big players, and an absolute increase in density – as long as it fails to disrupt the skewing – does not change the status of a language. To avoid confusion, we introduce a new term "low-affluence language"[1], which singles out the set of languages of interest but is oblivious as to whether there exist resources for a given language. It is then an empirical question which low-affluence language have resources of various kinds. Accordingly, the present paper surveys which languages have (a published description of) a computational morphological analyser of some sort.

We also aim to gather together work which is otherwise scattered in very disparate fora to facilitate comparison – presumably important lessons from one language can be taught to another language in a similar situation.

We focus on resources for morphological analysis since it is one of the bottom layers of analysis of written language data (cf. Beesley 2004, Berment 2004:18-26 and Oflazer 2009). The layer below, namely raw text data, has already been addressed in several other initiatives (Baldwin et al. 2006, Xia and Lewis 2009, Lewis 2003, Xia and Lewis 2008, Scannell 2007, Mallikarjun 2003). It is clear that the languages for which raw text data appears on the web[2] is much larger than those for which there is a published description of a morphological analyser.

2 Low-Affluence Languages

For this survey to be practical we need a clear definition of what a low-density language is, rather than a sliding scale or a prototypical set of properties.

We suggest that the set of languages of interest may be characterized in terms of the economic power of its speakers. Hence, in analogy with Gross National Product (GNP), we define the Gross Language Product (GLP), of a language as the total market value of all final goods and services produced by the speakers of the language within a calendar year. Since there is no detailed data available to compute this statistic accurately for a large number of languages, we use the following formula to estimate GLP:

$$GLP(L) = \sum_C S_C(L) \cdot GNP\text{-per-capita}(C)$$

Where $S_C(L)$ denotes the number of L1 speakers of language L in country C. For example, our data claims that Finnish has 4 700 000 speakers in Finland (GNP-per-capita = 46769), 200 000 in Sweden (GNP-per-capita = 50415) and

[1] For other choices of terminology see Streiter et al. (2006:2-3).

[2] According to a popular article *Weaving a Web of linguistic diversity*, `http://www.guardian.co.uk/GWeekly/Story/0,3939,427939,00.html`, 2001-01-25, retrieved 2006-09-12, this number is about 1 000. Though it is not clear how this figure was computed.

34 100 (GNP-per-capita = 14800) in Russia. So $GLP(Finnish) = 4700000 \cdot 46769 + 200000 \cdot 50415 + 34100 \cdot 14800 = 230\ 401\ 980\ 000$. Since most languages are spoken only in a few countries, $S_C(L) = 0$ for most countries C, thus contributing nothing to the sum.

We employ the convenient fiction that the GNP-per-capita for a country is indicative also of each subpopulation of speakers of that country. We strongly suspect, however, that for most minority languages of a country, the GNP-per-capita of the speakers of the minority languages is lower than the average of the country as a whole. Furthermore, we only have systematic data on L1 speaker numbers but, obviously, the economic power of L2 speakers is relevant as well. However, the cases where L2-economies appear to be the most significant[3] are among high-affluent languages anyway, so although their economic power is underestimated by the above formula, they are still too far from being low in affluence. Data for GNP per capita from Central Intelligence Agency of the United States (2007) and speaker numbers from Lewis (2009) are accessible. The GNP-figures used are not PPP-adjusted since the prices for NLP related services appear to have little to do with local prices for basic commodities. Table 1-2 shows the top 140 affluent languages according to the GLP-metric just explained.

For the purposes of this paper, we will put the threshold of low-density at 100 billion dollars of GLP. This threshold was chosen based on nothing more than its large number of zeroes and the fact that a convenient number of non-low-affluence languages emerge. With this setting, there are currently 44 non-low-affluence languages, and all the rest, beginning with Ukrainian at rank #45, are low-affluence languages.[4] Of the 44 high-affluence languages, 25 are predominantly European, 16 are Asian and 3 (English, Spanish, Portuguese) are European in origin but have more speakers in the Americas.

The top 44 high-affluence languages all have fair amounts of NLP infrastructure (including the less obvious Catalan, Farsi, Romanian, Hungarian and Bengali) except for an important class of languages with the following properties:

- They are not popularly written
- In the country where they are spoken, there is a standardized close relative which is the preferred language for written communication

The high-affluence languages which fall into this class are Bavarian-Mainfränkisch-Swiss German (under German), Min Nan-Wu-Yue-Hakka-Kinyu (under Mandarin Chinese), Galician (under Spanish), Lombard-Piemontese-Napoletano-Calabrese-Sicilian-Venetian (under Italian), Javanese (under Indonesian), Vlaams (under Dutch) and Najdi Arabic (under Modern Standard Arabic).

[3] In addition to extinct languages, there are important exceptions, such as Hausa, Swahili and Indonesian, where the L2-economy appears to make up a major fraction.

[4] Of course, the language-dialect divisions are debatable. We will assume the Ethnologue divisions for the present paper (Lewis 2009).

#	Language	iso-639-3	GLP	Total Pop.	GNP-per-capita
1	English	eng	14 570 119 604 622	326 959 888	44 562.4
2	Japanese	jpn	4 162 642 000 000	121 000 000	34 402.0
3	Spanish	spa	4 117 723 821 500	327 380 860	12 577.7
4	German, Standard	deu	3 408 803 154 660	84 959 210	40 122.8
5	Arabic, Standard	arb	2 807 780 000 000	206 000 000	13 630.0
6	French	fra	2 775 416 019 700	67 661 960	41 018.8
7	Chinese, Mandarin	cmn	2 146 387 252 580	845 033 030	2 540.0
8	Italian	ita	2 061 118 900 800	56 638 620	36 390.6
9	Russian	rus	1 809 937 945 460	125 102 940	14 467.5
10	Portuguese	por	1 345 089 888 980	174 307 980	7 716.7
11	Dutch	nld	966 883 126 740	21 309 290	45 373.7
12	Korean	kor	885 367 820 000	64 739 000	13 675.9
13	Bavarian	bar	570 287 492 000	13 259 000	43 011.3
14	Turkish	tur	445 343 958 400	47 777 700	9 321.1
15	Swedish	swe	412 626 274 000	8 206 000	50 283.4
16	Polish	pol	412 358 750 120	36 998 360	11 145.3
17	Catalan-Valencian-Balear	cat	404 298 076 000	11 351 000	35 617.8
18	Norwegian	nor	392 520 800 000	4 640 000	84 595.0
19	Chinese, Min Nan	nan	382 848 478 700	46 915 100	8 160.4
20	German, Swiss	gsw	337 411 118 000	6 469 000	52 158.1
21	Lombard	lmo	336 656 463 000	9 133 000	36 861.5
22	Greek	ell	334 090 819 150	11 526 360	28 984.9
23	Danish	dan	312 161 238 200	5 478 830	56 975.8
24	Vlaams	vls	267 949 458 000	6 132 000	43 696.9
25	Napoletano-Calabrese	nap	255 217 050 000	7 050 000	36 201.0
26	Finnish	fin	230 401 980 000	4 934 100	46 695.8
27	Mainfränkisch	vmf	197 946 650 000	4 910 000	40 315.0
28	Chinese, Wu	wuu	189 918 445 380	77 201 820	2 460.0
29	Hindi	hin	177 285 437 200	180 469 200	982.3
30	Sicilian	scn	174 850 830 000	4 830 000	36 201.0
31	Czech	ces	158 690 740 000	9 290 000	17 081.8
32	Javanese	jav	157 822 354 630	84 600 970	1 865.4
33	Romanian	ron	157 093 940 480	23 118 480	6 795.1
34	Hungarian	hun	156 132 196 260	12 253 140	12 742.2
35	Chinese, Yue	yue	148 402 526 750	54 471 530	2 724.4
36	Bengali	ben	121 632 915 000	180 624 200	673.4
37	Chinese, Hakka	hak	116 456 948 110	29 976 560	3 884.9
38	Galician	glg	113 090 635 000	3 185 000	35 507.2
39	Piemontese	pms	112 585 110 000	3 110 000	36 201.0
40	Hebrew	heb	112 330 850 000	4 850 000	23 161.0
41	Arabic, Najdi Spoken	ars	112 117 150 000	9 450 000	11 864.2
42	Chinese, Jinyu	cjy	110 655 000 000	45 000 000	2 459.0
43	Farsi, Western	pes	109 305 988 000	22 455 000	4 867.7
44	Venetian	vec	106 593 680 000	6 230 000	17 109.7
45	Ukrainian	ukr	96 173 972 800	31 260 800	3 076.5
46	Chinese, Xiang	hsn	88 524 000 000	36 000 000	2 459.0
47	Arabic, Egyptian Spoken	arz	83 580 000 000	52 500 000	1 592.0
48	Malay	zlm	82 378 737 000	10 181 000	8 091.4
49	Occitan	oci	81 924 544 560	2 048 310	39 996.1
50	Arabic, Hijazi Spoken	acw	81 785 879 400	6 023 900	13 576.8
51	Saxon, Upper	sxu	80 630 000 000	2 000 000	40 315.0
52	Arabic, Algerian Spoken	arq	80 478 000 000	20 400 000	3 945.0
53	Tatar	tat	79 181 967 200	5 350 800	14 798.1
54	Tamil	tam	79 091 044 720	66 787 482	1 184.2
55	Thai	tha	76 275 200 000	20 200 000	3 776.0
56	Arabic, Gulf Spoken	afb	75 108 676 000	2 339 000	32 111.4
57	Emiliano-Romagnolo	eml	73 129 640 100	2 020 100	36 201.0
58	Ligurian	lij	69 710 940 000	1 925 100	36 211.5
59	Telugu	tel	67 720 800 000	69 600 000	973.0
60	Arabic, South Levantine Spoken	ajp	67 391 610 000	6 010 000	11 213.2
61	Slovak	slk	67 261 405 800	4 894 800	13 741.4
62	Marathi	mar	66 164 000 000	68 000 000	973.0
63	Sunda	sun	62 730 000 000	34 000 000	1 845.0
64	Zulu	zul	58 707 636 000	10 344 500	5 675.2
65	Limburgish	lim	58 661 100 000	1 300 000	45 123.9
66	Azerbaijani, South	azb	56 677 840 000	12 360 000	4 585.5
67	Thai, Northeastern	tts	56 640 000 000	15 000 000	3 776.0
68	Urdu	urd	56 428 350 000	59 126 000	954.3
69	Vietnamese	vie	54 281 199 800	66 200 200	819.9
70	Panjabi, Western	pnb	53 307 830 000	62 510 000	852.7
71	Croatian	hrv	51 418 268 570	4 666 300	11 019.0
72	Chinese, Gan	gan	50 655 400 000	20 600 000	2 459.0
73	Chinese Sign Language	csl	49 278 360 000	20 040 000	2 459.0
74	Walloon	wln	48 885 760 000	1 120 000	43 648.0
75	Bulgarian	bul	47 578 541 000	8 781 750	5 417.8

Table 1. 140 top-affluence languages, ranked by GLP [Page 1(2)].

#	Language	iso-639-3	GLP	Total Pop.	GNP-per-capita
76	Kurdish, Northern	kmr	47 420 205 000	8 493 000	5 583.4
77	Xhosa	xho	45 530 524 000	7 808 000	5 831.2
78	Gujarati	guj	44 508 492 600	45 959 800	968.4
79	Slovene	slv	44 161 046 190	1 851 190	23 855.4
80	Tswana	tsn	42 204 873 500	4 521 700	9 333.8
81	Indonesian	ind	42 066 000 000	22 800 000	1 845.0
82	Arabic, North Levantine Spoken	apc	41 684 300 000	12 700 000	3 282.2
83	Arabic, Moroccan Spoken	ary	40 814 800 000	18 800 000	2 171.0
84	Kazakh	kaz	39 924 356 800	7 535 600	5 298.0
85	Arabic, Libyan Spoken	ayl	38 312 692 000	4 321 000	8 866.6
86	Filipino	fil	38 250 000 000	25 000 000	1 530.0
87	Bhojpuri	bho	37 933 946 000	38 546 000	984.1
88	Awadhi	awa	36 868 913 000	38 261 000	963.6
89	Pidgin, Nigerian	pcm	34 920 000 000	30 000 000	1 164.0
90	Malayalam	mal	34 798 470 000	35 410 000	982.7
91	Kannada	kan	34 346 900 000	35 300 000	973.0
92	Guadeloupean Creole French	gcf	34 089 600 000	848 000	40 200.0
93	Arabic, Mesopotamian Spoken	acm	34 054 400 000	14 500 000	2 348.5
94	Azerbaijani, North	azj	33 962 663 000	7 243 000	4 689.0
95	Okinawan, Central	ryu	33 851 568 000	984 000	34 402.0
96	Belarusan	bel	33 339 520 000	6 940 000	4 803.9
97	Swabian	swg	33 017 985 000	819 000	40 315.0
98	Tagalog	tgl	32 895 000 000	21 500 000	1 530.0
99	Maithili	mai	31 971 100 000	34 700 000	921.3
100	Lithuanian	lit	31 748 960 000	2 960 000	10 726.0
101	Oriya	ori	30 844 100 000	31 700 000	973.0
102	Arabic, Tunisian Spoken	aeb	30 645 000 000	9 000 000	3 405.0
103	Arabic, Sa'idi Spoken	aec	30 248 000 000	19 000 000	1 592.0
104	Luxembourgeois	ltz	29 030 843 150	320 010	90 718.5
105	Friulian	fur	28 743 594 000	794 000	36 201.0
106	Afrikaans	afr	28 506 527 300	4 903 900	5 813.0
107	Hawai'i Creole English	hwc	27 575 400 000	600 000	45 959.0
108	Gronings	gos	27 462 288 000	592 000	46 389.0
109	Turkmen	tuk	27 251 477 160	5 930 920	4 594.8
110	Panjabi, Eastern	pan	26 712 796 500	27 119 500	985.0
111	Sotho, Southern	sot	26 107 130 000	6 010 000	4 343.9
112	Chinese, Min Bei	mnp	25 469 408 000	10 304 000	2 471.7
113	Madura	mad	25 123 884 300	13 600 900	1 847.2
114	Chinese, Min Dong	cdo	25 099 442 220	9 134 060	2 747.8
115	Serbian	srp	25 004 841 100	6 379 460	3 919.5
116	Chuvash	chv	24 272 000 000	1 640 000	14 800.0
117	Cebuano	ceb	24 174 000 000	15 800 000	1 530.0
118	Hausa	hau	23 948 696 500	24 864 000	963.1
119	Sotho, Northern	nso	23 897 870 000	4 090 000	5 843.0
120	Basque	eus	23 697 320 000	656 200	36 112.9
121	Welsh	cym	23 340 308 000	533 000	43 790.4
122	Uyghur	uig	22 699 715 500	8 704 500	2 607.8
123	Thai, Northern	nod	22 661 809 200	6 009 400	3 771.0
124	Réunion Creole French	rcf	22 311 000 000	555 000	40 200.0
125	Yoruba	yor	22 304 640 000	19 365 000	1 151.8
126	Sinhala	sin	22 257 112 950	15 500 850	1 435.8
127	Frisian, Western	fry	21 663 663 000	467 000	46 389.0
128	Igbo	ibo	20 952 000 000	18 000 000	1 164.0
129	Gaelic, Irish	gle	20 697 310 000	355 000	58 302.2
130	Bashkort	bak	20 424 000 000	1 380 000	14 800.0
131	Hunsrik	hrx	20 328 000 000	3 000 000	6 776.0
132	Breton	bre	20 100 000 000	500 000	40 200.0
133	Umbundu	umb	20 012 000 000	4 000 000	5 003.0
134	Chechen	che	19 691 935 000	1 333 000	14 772.6
135	Malay, Kedah	meo	19 523 400 000	2 600 000	7 509.0
136	Sindhi	snd	18 440 630 000	21 310 000	865.3
137	Sardinian, Logudorese	src	18 100 500 000	500 000	36 201.0
138	Kabardian	kbd	17 807 775 000	1 629 000	10 931.7
139	Arabic, Sudanese Spoken	apd	17 580 000 000	15 000 000	1 172.0
140	Chhattisgarhi	hne	17 027 500 000	17 500 000	973.0
141	Thai, Southern	sou	16 992 000 000	4 500 000	3 776.0
142	Latvian	lav	16 816 220 000	1 390 000	12 098.0
143	Assamese	asm	16 253 429 000	16 709 000	972.7
144	Arabic, North Mesopotamian Spoken	ayp	16 120 100 000	6 100 000	2 642.6
...					

Table 2. 144 top-density languages, ranked by GLP [Page 2(2)].

3 Survey Methodology

The methodology used for gathering the items in the survey was general knowledge, browsing of the meta-literature, the corpora-list and googling suitably for each of the 100 most affluent low-affluence languages. No restriction as to the meta-language was applied (the author is able read nearly all languages under consideration) but, nevertheless, mostly descriptions written in English turned up.

For languages which have very little morphology we listed some other NLP work at a comparable stage, that is, the next processing level after raw text data. Similarly, work on languages where the orthography forces morphological segmentation, is listed along the same principle.

In some cases, the work found does not cover all of morphology of the language in question. In such cases, the language is "rounded-off" and counted as one having morphological resources. Stemmers are counted along the same principle.

We have adhered to the criterion that the work has to have been published, available as a thesis or available on the web with a fair amount of information. Conference presentations, manuscripts and other items of partial or unavailable reports are not counted.[5]

For short, we use the term CMR for computational morphological resources with a published description of the kind explained.

4 Survey Results

The results of the survey are shown in Tables 3-6, arranged alphabetically and by continent. Table 7 shows a summary of the number of languages found in the survey. The survey lists 80 low-affluence languages for with CMR. We suspect, however, that there is more work on which nothing has been published so far.

5 Discussion

The survey allows us to shed light on the questions posed in the introduction about the emergence of CMR for low-affluence languages.

5.1 Which languages obtain CMR?

The survey outcome confirms the intuition that high GLP is strongly correlated with obtaining CMR – on all continents, the average GLP of the languages with CMR is higher than the average GLP of the continent as a whole.

[5] If such had been included on the basis of their title or the like, a much wider range of languages would have been encountered (see Wedekind and Wedekind 2009, Wedekind 2008, Wedekind et al. 1983, Adegbola 2009, Ridings and Mavhu 2002, Maphosa 2002 for some examples known to the present author).

Language	iso-639-3	GLP	#	Publication(s)
Afrikaans	afr	28 506 527 300	106	de Stadler and Coetzer 1990
Amharic	amh	5 154 640 000	255	Gambäck et al. 2009, Ephrem 2006, Alemayehu and Willett 2002, Amsalu and Gibbon 2005, Alemu Argaw and Asker 2007, Bayou 2000, Bati 2002, Fissaha Adafre 2005
Bamanankan	bam	1 523 003 100	442	Fleisch and Seidel 2006
Dholuo	luo	3 447 920 000	306	Pauw et al. 2007
Gikuyu	kik	5 700 920 000	242	De Pauw and Wagacha 2007
Ekegusii	guz	1 683 403 300	417	Elwell 2008
Ha	haq	406 890 000	790	Harjula 2005
Malagasy, Plateau	plt	2 827 954 700	339	Dalrymple et al. 2006
Ndebele	nbl	3 739 520 000	288	Bosch et al. 2008a
Oshiwambo	kua	2 989 782 000	325	Hurskainen and Halme 2001
Rwanda	kin	2 496 844 000	349	Muhirwe and Trosterud 2008, Muhirwe 2007
Somali	som	4 050 678 000	277	Abdillahi et al. 2006
Sotho, Northern	nso	23 897 870 000	119	Faaß et al. 2009, Prinsloo and Heid 2006, Prinsloo 1994
Sotho, Southern	sot	26 107 130 000	111	Johnson 2008, de Schryver and Pauw 2007
Swahili	swh	424 619 270	777	Hurskainen 1992, Pauw et al. 2006, Pauw and de Schryver 2008
Swati	ssw	8 473 423 400	196	Bosch et al. 2008a
Tamajaq, Tawallammat	ttq	252 591 640	1001	Enguehard and Modi 2009
Tswana	tsn	42 204 873 500	80	Groenewald 2009, Pretorius et al. 2009
Xhosa	xho	45 530 524 000	77	Bosch et al. 2008b, 2003, Pretorius and Bosch 2009, Bosch et al. 2008a
Yoruba	yor	22 304 640 000	125	Finkel and Odejobi 2009
Zulu	zul	58 707 636 000	64	Bosch et al. 2008b, Pretorius and Bosch 2003

Table 3. Low-affluence languages with CMR 1(4): Africa.

Language	iso-639-3	GLP	#	Publication(s)
American Sign Language	ase	91 596 287	1636	Shield and Baldridge 2008
Aymara, Central	ayr	4 465 996 300	269	Beesley 2003
Cayuga	cay	2 633 490	4829	Graham 2007
Chuj, Ixtatán	cnm	136 579 200	1357	Medina Urrea and Díaz 200
Cree, Plains	crk	1 482 847 900	447	Wolfart and Pardo 1979
Inuktitut, Greenlandic	kal	3 296 912 000	311	Trosterud 2008b
Inupiatun, North Alaskan	esi	1 720 048 230	416	Trosterud 2008c
K'iche', West Central	qut	661 750 000	633	Kudlek 1975
Mapudungun	arn	2 658 600 000	343	Monson et al. 2008
Tarahumara, Central	tar	452 045 000	756	Medina-Urrea 2006

Table 4. Low-affluence languages with CMR 2(4): Americas.

The extinct Ancient Greek, Latin, Syriac and Sanskrit serve as cultural heritage languages. If we count the people in these respective cultural sphere and estimate their purchasing powers, these languages too achieve considerable GLP.

Most remaining language with low (or even very low) GLP which nevertheless have CMR, i.e., Akkadian (extinct), Cayuga, American Sign Language, Chuj, Tarahumara, K'iche', Ha, A-Pucikwar and the small Uralic languages all appear to owe their CMR to a dedicated individual or group who "happen" to be interested in the language, for a variety of reasons independent of GLP. Such endeavors tend to be more distant from concrete usage with the speech community. For example, we find work here which was produced in past decades when the practical role of computational processing of language must have been rather opaque. Also, we find proof-of-concept descriptions of other aspects than morphology for this class of languages, e.g., Bender (2008). Except for continental biases, the selection of these languages versus the plethora of other languages of similar low GLP appears to be random. The continent bias is that the Americas are overrepresented, while, if anything, the Pacific region is underrepresented.

Swahili and Tamajaq also have low GLP, but Swahili has a high number of L2 speakers. Tamajaq, on the other hand, is a true example of a low-GLP language with CMR which is not due to an external dedicated individual.

As even with high-affluence languages, languages which are second to a dominant nation-state variety tend lag behind in attaining CMR, despite competitive GLP.

Among the remaining languages that have competitive GLP but lack CMR, the only real eye-raisers are the Nigerian languages Hausa and Igbo (but see Adegbola 2009).

5.2 Who creates CMR?

As noted above, a small number of CMR are the result of the entrepreneurship of a dedicated individual.

Language	iso-639-3	GLP	#	Publication(s)
Akkadian	akk	0	6863	Berthélemy 1998, Kataja and Koskenniemi 1988
Assamese	asm	16 253 429 000	143	Sharma et al. 2002
A-Pucikwar	apq	9 730	6807	Choudhary 2006
Burmese	mya	9 296 300 000	185	Htay and Murthy 2008, Maung and Mikami 2008
Gujarati	guj	44 508 492 600	78	Patel and Gali 2008
Indonesian	ind	42 066 000 000	81	Hammarström 2009, Adriani et al. 2007, Pisceldo et al. 2008, Indradjaja and Bressan 2003
Kannada	kan	34 346 900 000	91	Vikram and Urs 2007, Sharada and Lakshmi 2006
Lao	lao	1 864 455 000	402	Berment 2004
Malay	zlm	82 378 737 000	48	Sankupellay and Valliappan 2006, Abdullah et al. 2009, Ahmad et al. 1996, Tai et al. 2000, Ranaivo-Malançon 2004
Malayalam	mal	34 798 470 000	90	Idicula and David 2007
Meitei	mni	1 341 941 000	468	Singh and Bandyopadhyay 2008, 2006, Choudhury et al. 2004
Marathi	mar	66 164 000 000	62	Devlekar et al. 2006
Mongolian, Halh	khk	3 121 958 000	321	Khaltar and Fujii 2008
Oriya	ori	30 844 100 000	101	Mohanty et al. 2005, Shabadi 2003
Pashto, Northern	pbu	12 475 810 000	162	Khan and Zuhra 2007
Sanskrit	san	2 870 350	4757	Huet 2005, Bharati et al. 2006, Jha et al. 2006
Sinhala	sin	22 257 112 950	126	Herath et al. 1989
Syriac	syc	0	6863	Kiraz 1998, 2001, 2000
Tagalog	tgl	32 895 000 000	98	Nelson 2004
Tamil	tam	79 091 044 720	54	Anandan et al. 2002, Viswanathan et al. 2003
Telugu	tel	67 720 800 000	59	Karthik Kumar et al. 2006, Rama Sree et al. 2008
Thai	tha	76 275 200 000	55	Tongchim et al. 2008
Turkmen	tuk	27 251 477 160	109	Tantuğ et al. 2006
Uyghur	uig	22 699 715 500	122	Ablimit et al. 2008
Urdu	urd	56 428 350 000	68	Humayoun et al. 2007, Hardie 2003, Bögel et al. 2008, Hussain 2008, 2004, Akram et al. 2009
Vietnamese	vie	54 281 199 800	69	Nguyen et al. 2008

Table 5. Low-affluence languages with CMR 3(4): Asia.

Language	iso-639-3	GLP	#	Publication(s)
Basque	eus	23 697 320 000	120	Alegria et al. 1996
Bulgarian	bul	47 578 541 000	75	Slavcheva 2003, Simov et al. 2004, Nakov 2003, Angelov 2008
Czech	ces	158 690 740 000	31	Chrupała 2008, Schmid and Laws 2008
Erzya	myv	9 088 541 500	187	Prószéky and Novák 2005
Estonian	est	15 691 577 000	146	Uibo 2002, Kaalep and Vaino 2001, Müürisep et al. 2003
Faroese	fao	2 749 328 000	341	Trosterud 2008a
Gaelic, Irish	gle	20 697 310 000	129	Uí Dhonnchadha et al. 2003, Sulger 2008
Greek, Ancient	grc	0	6863	Lee 2008
Khanty	kca	201 280 000	1130	Prószéky and Novák 2005
Komi-Zyrian	kpv	3 211 600 000	317	Prószéky and Novák 2005
Icelandic	isl	15 235 200 000	148	Loftsson 2008a,b
Latin	lat	0	6863	Forsberg 2007
Latvian	lav	16 816 220 000	142	Paikens 2008
Lithuanian	lit	31 748 960 000	100	Rimkutė et al. 2007
Mansi	mns	40 700 000	2336	Prószéky and Novák 2005
Nenets	yrk	463 240 000	747	Prószéky and Novák 2005
Nganasan	nio	7 400 000	3923	Prószéky and Novák 2005
Saami, North	sme	1 550 092 300	436	Trosterud 2008d
Serbian	srp	25 004 841 100	115	Krstev et al. 2004, Kešelj and Šipka 2008
Slovene	slv	44 161 046 190	79	Erjavec and Džeroski 2004, Chrupała 2008, Hajič 2000
Udmurt	udm	6 867 200 000	214	Prószéky and Novák 2005
Ukrainian	ukr	96 173 972 800	45	Katrenko 2004, Kovalenko 2002
Welsh	cym	23 340 308 000	121	Chrupała 2008

Table 6. Low-affluence languages with CMR 4(4): Europe.

Continent	# languages with CMR	Average GLP for languages with CMR	# Languages	Average GLP
Africa	21	13 830 066 200.5	2169	499 712 661.1
Americas	10	1 496 900 840.7	1187	272 294 046.4
Asia	26	32 734 534 912.4	2320	1 295 239 804.7
Europe	23	24 682 518 995.0	258	6 229 757 797.5
Oceania	0	-	1380	31 167 413.1
Total	80	20 834 739 866.5	7314	828 867 197.7

Table 7. Numbers of low-affluence languages with CMR per continent, along with continent totals.

Not one of the surveyed publications shows any sign of private sector involvement (though this does not preclude that the work described is made use of in the private sector after the publication). If private companies are interested in CMR for low-affluence languages, they do not publish about this.

The bulk of CMR work comes about through academic support in the respective nation state. Some or all authors of the published articles are either employed at universities in the corresponding nation state and/or acknowledge more specific funding initiatives to develop CMR.

Perhaps surprisingly, there is no Australian Aboriginal language represented in spite of their presence in a high-technological country where there is government support and revitalization efforts (Larkin 2005).

5.3 How are CMR created?

On one end of the spectrum, CMR consist of human-written rules in some formal framework. At the other end, unsupervised methods are used to induce CMR probabilistically from raw text data. A mid-way approach arranges the human to provide annotation of a certain amount of text, from which CMR is abstracted using supervised methods.

Another dimension of interest is that of transfer, i.e., if language A has CMR and language B is similar to A, then CMR for B are built from those of A. Transfer can occur with manually intensive methods, e.g., the source-code for morphological rules is taken over and is subjected to small adjustments, as well as in unsupervised methods, e.g., the same unsupervised method is applied to a new language with similar morphological typology.

Though we cannot go into detail, there is one clear division as to CMR for low-affluence languages, namely that the majority of CMR are human-written morphological rule systems, often in a finite-state framework. In later years, more supervised and unsupervised work has began to surface, probably boosted by increased availability of raw text data.

Relative to the number of opportunities, there is surprisingly little evidence of transfer, though this is visible among the Southern and Eastern Bantu languages, and presumably among the small Uralic languages. It appears to be lacking in such obvious pairs as Faroese-Icelandic, Malay-Indonesian, Dutch-Afrikaans and Lao-Thai. In fact, even for one and the same language, in the cases where we have different lines of work, they tend to be parallel rather than serial. There is insufficient information for a systematic analysis, but one may speculate that the lack of transfer has to do with access limitations to source-code and annotated corpora.

Put bluntly, one strategic lesson that emerges is that, sooner or later, GLP-competitive languages will muster the institutional support for manual labour-intensive CMR, and thus that unsupervised solutions have an edge on languages with lower GLP.

6 Conclusion

Analogous to the Gross National Product (GNP) of a country, one may define the Gross Language Product (GLP) of a language as the sum economic power of its speakers. Low-affluence languages are languages whose speakers muster a low GLP. Through an empirical survey of published work on Computational morphological resources (CMR), we can confirm the intuition that Computational morphological resources surface for languages with high GLP, except those high-GLP languages which are dominated by a related language in their respective nation-state. The rather obvious generalization that languages which are not even popularly written rarely attain CMR is also borne out. Also, a few "lucky" languages have CMR owing to the entrepreneurship of a dedicated individual. The manner in which GLP-competitive languages obtain their CMR is via state support through funded university positions and/or specific language technology development funding. Non-state-channelled driving forces, such as private sector, are not heard of in the survey. So far, most CMR for low-affluence languages has been manually built rule systems but the field is open also for supervised and unsupervised statistical approaches.

Acknowledgements

The author has benefited from commentary by Håkan Burden and Lars Borin. The usual disclaimers apply.

References

Abdillahi, Nimaan, Pascal Nocera & Juan-Manuel Torres-Moreno. 2006. Boites à outils TAL pour les langues peu informatisées: le cas du Somali. In *Journées d'Analyses des Données Textuelles (JADT 06)*, 697-705. Besançon-France.

Abdullah, Muhamad Taufik, Fatimah Ahmad, Ramlan Mahmod & Tengku Mohd Tengku Sembok. 2009. Rules Frequency Order Stemmer for Malay Language. *International Journal of Computer Science and Network Security* 9(2). 433–438.

Ablimit, Mijit, M. Eli & T. Kawahara. 2008. Partly supervised Uighur morpheme segmentation. In *Proceedings of the Oriental-COCOSDA Workshop*, 71-76. Japan.

Adegbola, Tunde. 2009. Building capacities in human language technology for African languages. In *Proceedings of the First Workshop on Language Technologies for African Languages*, 53–58. Athens, Greece: Association for Computational Linguistics.

Adriani, Mirna, Jelita Asian, Bobby Nazief, S. M. M. Tahaghoghi & Hugh E. Williams. 2007. Stemming Indonesian: A confix-stripping approach. *ACM Transactions on Asian Language Information Processing (TALIP)* 6(4). 1–33.

Ahmad, Fatimah, Mohammed Yusoff & Tengku M. T. Sembok. 1996. Experiments with a stemming algorithm for Malay words. *Journal of the American Society for Information Science* 47(12). 909–918.

Akram, Qurat-ul-Ain, Asma Naseer & Sarmad Hussain. 2009. Assas-band, an affix-exception-list based Urdu stemmer. In *Proceedings of the 7th Workshop on Asian Language Resources*, 40–47. Suntec, Singapore: Association for Computational Linguistics.

Alegria, I., X. Artola, K. Sarasola & M. Urkia. 1996. Automatic morphological analysis of Basque. *Literary & Linguistic Computing* 11(4). 193–203.

Alemayehu, Nega & Peter Willett. 2002. Stemming of Amharic words for information retrieval. *Literary and Linguistic Computing* 17. 1–17.

Alemu Argaw, Atelach & Lars Asker. 2007. An Amharic stemmer: Reducing words to their citation forms. In *Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources*, 104–110. Prague, Czech Republic: Association for Computational Linguistics.

Amsalu, Saba & Dafydd Gibbon. 2005. Finite state morphology of Amharic. In *5th Recent Advances in Natural Language Processing*, 47-51. Borovets, Bulgaria.

Anandan, P., K. Saravanan Ranjani Parthasarathi & T. V. Geetha. 2002. Morphological Analyzer for Tamil. In *International Conference on Natural Language Processing, ICON-2002, Mumbai, December 18-21, 2002*, 3-10. Vikas Publishing House Pvt Ltd., New Delhi.

Angelov, Krasimir. 2008. Type-theoretical Bulgarian grammar. In *GoTAL '08: Proceedings of the 6th international conference on Advances in Natural Language Processing*, 52–64. Berlin, Heidelberg: Springer-Verlag.

Baldwin, T., S. G. Bird & B. Hughes. 2006. Collecting low-density language materials on the web. In *Proceedings of 12th Australasian Web Conference (AusWeb06)*. Southern Cross University.

Bati, Tesfaye Bayu. 2002. Automatic morphological analyser: An experiment using unsupervised and autosegmental approach. Addis Ababa University, Ethiopia masters thesis.

Bayou, Abiyot. 2000. Design and development of word parser for Amharic language. Addis Ababa University, Ethiopia masters thesis.

Beesley, Kenneth R. 2003. Finite-State Morphological Analysis and Generation for Aymara. EACL Workshop on Finite State Methods in Natural Language Processing, Budapest, 2003.

Beesley, Kenneth R. 2004. Morphological analysis and generation: A first-step in natural language processing. In *First Steps in Language Documentation for Minority Languages: Computational Linguistic Tools for Morphology, Lexicon and Corpus Compilation, Proceedings of the SALTMIL Workshop at LREC 2004*, 1-8. Lisboa, Portugal.

Bender, Emily M. 2008. Evaluating a crosslinguistic grammar resource: A case study of Wambaya. In *Proceedings of ACL-08: HLT*, 977–985. Columbus, Ohio: Association for Computational Linguistics.

Berment, V. 2004. Méthodes pour informatiser les langues et les groupes de langues «peu dotées». Université Joseph-Fourier, Grenoble I doctoral dissertation.

Berthélemy, François. 1998. A Morphological Analyzer for Akkadian Verbal Forms with a Model of Phonetic Transformations. In Michael Rosner (ed.), *Proceedings of the Workshop on Computational Approaches to Semitic Languages 16th August 1998*. Université de Montreal, Montreal, Quebec, Canada.

Bharati, Akshar, Amba Kulkarni & V. Sheeba. 2006. Building a Wide Coverage Sanskrit Morphological Analyzer: A Practical Approach. In *Proceedings of the First National Symposium on Modeling and Shallow Parsing of Indian Languages MSPIL-06 (An event of NLP Association of India) 2-4 April 2006*, 35-41. Centre for Indian Language Technology, Department of Computer Science and Engineering, Indian Institute of Technology, Bombay, Mumbai, India.

Bögel, Tina, Miriam Butt, Annette Hautli & Sebastian Sulger. 2008. Developing a Finite-State Morphological Analyzer for Urdu and Hindi. In *Proceedings of the Sixth International Workshop on Finite-State Methods and Natural Language Processing (FSMNLP 2007)*. Potsdam.

Bosch, Sonja, Kholisa Podile, Jackie Jones & Mmemezi Mfusi. 2003. Automating Xhosa Morphology. Paper presented at the 12th Biennial International ALASA (African Language Association of Southern Africa) Conference, University of Stellenbosch, 8 July 2003.

Bosch, Sonja, Laurette Pretorius & Axel Fleisch. 2008a. Experimental Bootstrapping of Morphological Analysers for Nguni Languages. *Nordic Journal of African Studies* 17(2). 66–88.

Bosch, Sonja, Laurette Pretorius, Kholisa Podile & Axel Fleisch. 2008b. Experimental Fast-Tracking of Morphological Analysers for Nguni Languages. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, 2588-2595. Marrakech, Morocco.

Central Intelligence Agency of the United States. 2007. *The World Factbook*. U.S. government.

Choudhary, Narayan Kumar. 2006. Developing a Computational Framework for the Verb Morphology of Great Andamanese. Literature & Culture Studies Jawaharlal Nehru University New Delhi, India: Centre for Linguistics School of Language masters thesis.

Choudhury, Sirajul Islam, Leihaorambam Sarbajit Singh, Samir Borgohain & Pradip Kumar Das. 2004. Morphological Analyzer for Manipuri: Design and Implementation. In *Applied Computing: Proceedings of the Second Asian Applied Computing Conference, AACC 2004, Kathmandu, Nepal, October 29-31, 2004* (LNCS 3285), 123-129. Berlin: Springer-Verlag.

Chrupała, Grzegorz. 2008. Towards a Machine-Learning Architecture for Lexical Functional Grammar Parsing. Dublin City University doctoral dissertation.

Dalrymple, Mary, Maria Liakata & Lisa Mackie. 2006. Tokenization and Morphological Analysis for Malagasy. *Computational Linguistics and Chinese Language Processing* 11(4). 315–332.

David, Anne & Michael Maxwell. 2008. Invited talk: Building language resources: Ways to move forward. In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*, 1–2. Hyderabad, India: Asian Federation of Natural Language Processing.

De Pauw, Guy & Peter W. Wagacha. 2007. Bootstrapping morphological analysis of Gĩkũyũ using Maximum Entropy Learning. In *Proceedings of the 8th Annual Conference of the International Speech Communication Association (INTERSPEECH 2007), Antwerp, Belgium, August 27-31, 2007*, 1517-1520. ISCA.

de Schryver, G.-M. & G. De Pauw. 2007. Dictionary Writing System (DWS) + Corpus Query Package (CQP): The case of TshwaneLex. *Lexikos* 17. 226–246.

de Stadler, L. G. & M. W. Coetzer. 1990. A morphological parser for afrikaans. In *Proceedings of the 13th conference on Computational linguistics*, 85–88. Morristown, NJ, USA: Association for Computational Linguistics.

Devlekar, Sushant S., Sachin S. Burange & Pushpak Bhattacharyya. 2006. Rule Governed Marathi POS Tagging. In *Proceedings of the First National Symposium on Modeling and Shallow Parsing of Indian Languages MSPIL-06 (An event of NLP Association of India) 2-4 April 2006*, 69-78. Centre for Indian Language Technology, Department of Computer Science and Engineering, Indian Institute of Technology, Bombay, Mumbai, India.

Elwell, Robert. 2008. Finite State Methods for Bantu Verb Morphology. In Nicholas Gaylord, Alexis Palmer & Elias Ponvert (eds.), *Computational Linguistics for Less-Studied Languages* (Texas Linguistics Society X), 56-67. Stanford: CSLI.

Enguehard, Chantal & Issouf Modi. 2009. Towards an electronic dictionary of Tamajaq language in Niger. In *Proceedings of the First Workshop on Language Technologies for African Languages*, 81–88. Athens, Greece: Association for Computational Linguistics.

Ephrem, Binyam. 2006. Nominal Inflection in Amharic and their Implementation in Computational Grammar. Norweigian University of Science and Technology, Trondheim masters thesis.

Erjavec, Tomaž & Sašo Džeroski. 2004. Machine Learning of Morphosyntactic Structure: Lemmatizing Slovene Words. *Applied Artificial Intelligence* 18. 17–41.

Faaß, Gertrud , Ulrich Heid, Elsabé Taljard & Danie Prinsloo. 2009. Part-of-speech tagging of Northern Sotho: Disambiguating polysemous function words. In *Proceedings of the First Workshop on Language Technologies for African Languages*, 38–45. Athens, Greece: Association for Computational Linguistics.

Finkel, Raphael & Odetunji Ajadi Odejobi. 2009. A computational approach to Yorùbá morphology. In *Proceedings of the First Workshop on Language Technologies for African Languages*, 25–31. Athens, Greece: Association for Computational Linguistics.

Fissaha Adafre, Sisay. 2005. Part of speech tagging for Amharic using conditional random fields. In *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*, 47–54. Ann Arbor, Michigan: Association for Computational Linguistics.

Fleisch, Axel & Frank Seidel. 2006. Cologne Initiative on Natural Language Processing in African Languages. In *Proceedings on the Workshop on Networking the Development of Language Resources for African Languages held in conjunction with the 5th International Conference on Language Resources and Evaluation (LREC), 22 May 2006, Genoa, Italy*, 17-22. European Language Resources Association (ELRA).

Forsberg, Markus. 2007. Three Tools for Language Processing: BNF Converter, Functional Morphology, and Extract. Chalmers University of Technology, Gothenburg doctoral dissertation.

Gambäck, Björn , Fredrik Olsson, Atelach Alemu Argaw & Lars Asker. 2009. Methods for Amharic part-of-speech tagging. In *Proceedings of the First Workshop on Language Technologies for African Languages*, 104–111. Athens, Greece: Association for Computational Linguistics.

Graham, Dougal. 2007. Finite-state parsing of Cayuga morphology. Memorial University of Newfoundland, Canada masters thesis.

Groenewald, Hendrik Johannes. 2009. Using technology transfer to advance automatic lemmatisation for Setswana. In *Proceedings of the First Workshop on Language Technologies for African Languages*, 32–37. Athens, Greece: Association for Computational Linguistics.

Hajič, Jan. 2000. Morphological tagging: data vs. dictionaries. In *Proceedings of the first conference on North American chapter of the Association for Computational Linguistics*, 94–101. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Hammarström, Harald. 2009. Poor man's word-segmentation: Unsupervised morphological analysis for Indonesian. In *Proceedings of the Third International Workshop on Malay and Indonesian Language Engineering (MALINDO)*. Singapore: ACL.

Hardie, A. 2003. The computational analysis of morphosyntactic categories in Urdu. University of Lancaster doctoral dissertation.

Harjula, L. 2005. Morphological Parsing of Tone: An Experiment with Two-Level Morphology on the Ha language. *Nordic Journal of African Studies* 14(4). 452–463.

Herath, S., T. Ikeda, S. Yokoyama, H. Isahara & S. Ishizaki. 1989. Sinhalese morphological analysis: a step towards machine processing of sinhalese. In *IEEE International Workshop on Tools for Artificial Intelligence: Architectures, Languages and Algorithms*, 100-107. IEEE Computer Society Press.

Htay, Hla Hla & Kavi Narayana Murthy. 2008. Myanmar word segmentation using syllable level longest matching. In *The 6th Workshop on Asian Language Resources*, 41-48. Asian Federation of Natural Language Processing.

Huet, Gérard. 2005. A functional toolkit for morphological and phonological processing, application to a Sanskrit tagger. *Journal of Functional Programming* 15(4). 573–614.

Humayoun, M., H. Hammarström & A. Ranta. 2007. Urdu morphology, orthography and lexicon extraction. In Ali Farghaly & Karine Megerdoomian (eds.), *Proceedings of the 2nd Workshop on Computational Approaches to Arabic Script-based Languages*, 59-68. Stanford, California.

Hurskainen, A. 1992. A Two-Level Computer Formalism for the Analysis of Bantu Morphology: An Application to Swahili. *Nordic Journal of African Studies* 1(1). 87–119.

Hurskainen, A. & R. Halme. 2001. Mapping Between Disjoining and Conjoining Writing Systems in Bantu Languages: Implementation on Kwanyama. *Nordic Journal of African Studies* 10. 399–414.

Hussain, Sara. 2004. Finite-State Morphological Analyzer for Urdu. Center for Research in Urdu Language Processing, National University of Computer and Emerging Sciences, Pakistan masters thesis.

Hussain, Sarmad. 2008. Resources for Urdu language processing. In *The 6th Workshop on Asian Language Resources*, 99-100. Asian Federation of Natural Language Processing.

Idicula, Sumam Mary & Peter S. David. 2007. A Morphological Processor for Malayalam Language. *South Asia Research* 27(2). 173–186.

Indradjaja, Lily Suryana & Stéphane Bressan. 2003. Automatic learning of stemming rules for the Indonesian language. In *Language, Information and Computation: Proceedings of the 17th Pacific Asia Conference, 1-3 October, 2003, Sentosa, Singapore*, 62-68. COLIPS.

Jha, Girish N., Sudhir K. Mishra, R. Chandrashekar, Priti Bhowmik, Subash, Sachin Mendiratta & Muktanand. 2006. Towards a Computational Analysis System for Sanskrit. In *Proceedings of the First National Symposium on Modeling and Shallow Parsing of Indian Languages MSPIL-06 (An event of NLP Association of India) 2-4 April 2006*, 25-34. Centre for Indian Language Technology, Department of Computer Science and Engineering, Indian Institute of Technology, Bombay, Mumbai, India.

Johnson, Mark. 2008. Unsupervised word segmentation for Sesotho using adaptor grammars. In *Proceedings of the Tenth Meeting of ACL Special Interest Group on Computational Morphology and Phonology*, 20–27. Columbus, Ohio: Association for Computational Linguistics.

Kaalep, Heiki-Jaan & Tarmo Vaino. 2001. Complete morphological analysis in the linguists toolbox. In *Congressus Nonus Internationalis Fenno-Ugristarum Pars V*, 9-16. Tartu.

Karthik Kumar, Gali, K. Sudheer & P. V. S. Avinesh. 2006. Comparative study of various machine learning methods for Telugu part of speech tagging. In *Proceedings of the NLPAI contest workshop during NWAI-06, SIGAI.* Mumbai.

Kataja, Laura & Kimmo Koskenniemi. 1988. Finite-state description of Semitic morphology: A case study of Ancient Akkadian. In *COLING '88*, 313-315. ACL.

Katrenko, Sophia. 2004. Towards unsupervised learning of morphology applied to Ukrainian. In Laura Alonso i Alemany & Paul Egré (eds.), *Student Session: 16th European Summer School in Logic, Language and Information, Nancy, France, 9-20 August, 2004*, 138-148. FoLLI.

Kešelj, Vlado & Danko Šipka. 2008. A Suffix Subsumption-based Approach to Building Stemmers and Lemmatizers for Highly Inflectional Languages with

Sparse Resources. *Infotheca: Journal of Informatics and Librarianship* IX(1-2). 23.

Khaltar, Badam-Osor & Atsushi Fujii. 2008. A lemmatization method for modern Mongolian and its application to information retrieval. In *Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP 2008)*, 1-8. Hyderabad, India: Asian Federation of Natural Language Processing.

Khan, Mohammad Abid & Fatima Tuz Zuhra. 2007. The Computational Morphology of Pashto Nouns. *South Asian Language Review* XVII(1). 1–27.

Kiraz, George Anton. 1998. Syriac morphology: From a linguistic model to a computational implementation. In R. Lavenant (ed.), *Symposium Syriacum VII: Uppsala University, Department of Asian and African Languages 11 - 14 August 1996* (Orientalia christiana analecta 256). Pontificio Istituto Orientale, Roma.

Kiraz, George Anton. 2000. Multitiered nonlinear morphology using multitape finite automata: a case study on Syriac and Arabic. *Computational Linguistics* 26(1). 77–105.

Kiraz, George Anton. 2001. *Computational Nonlinear Morphology: With Emphasis on Semitic Languages* (Studies in Natural Language Processing). Cambridge University Press.

Kovalenko, Andrej. 2002. Verojatnostnyj morfologičeskij analizator Russkogo i Ukrainskogo jazykov [A probabilistic Morphological Analyzer of Russian and Ukrainian]. *Sistemnyj Administrator [System Administrator]* 1. –.

Krstev, Cvetana, Duško Vitas & Tomaz Erjavec. 2004. Morpho-syntactic descriptions in multext-east - the case of serbian. In *Informatica 28*, 431-436. The Slovene Society Informatika, Ljubljana.

Kudlek, Manfred. 1975. Computer Programs for Generating and Analyzing Quiché verb phrases. In Ernesta Cerulli & Gilda Della Ragione (eds.), *Linguistica – Folklore – Storia americana – Sociologia* (Atti del XL Congresso Internazionale degli Americanisti (Roma – Genova, 3–10 Settembre 1972) 3), 45-54. Tilgher, Genoa.

Larkin, Steve. 2005. National Indigenous Languages Survey Report. Australian Institute of Aboriginal and Torres Strait Islander Studies.

Lee, John. 2008. A nearest-neighbor approach to the automatic analysis of Ancient Greek morphology. In *CoNLL 2008: Proceedings of the Twelfth Conference on Computational Natural Language Learning*, 127–134. Manchester, England: Coling 2008 Organizing Committee.

Lewis, Paul M. (ed.). 2009. *Ethnologue: Languages of the World.* 16th edn. Dallas: SIL International.

Lewis, William D. 2003. Mining and migrating interlinear glossed text. In *Proceedings of the EMELD Workshop on Digitizing and Annotating Texts and Field Recordings.* East Lansing, MI.

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

Loftsson, H. 2008b. Tagging Icelandic text: An experiment with integrations and combinations of taggers. *Nordic Journal of Linguistics* 40(2). 175–181.

Mallikarjun, B. 2003. Corpora in minor languages of India: Some issues. In *Proceedings of the EACL 2003 Workshop on Computational Linguistics for South Asian Languages*, 35-42. Budapest, Hungary: Association for Computational Linguistics.

Maphosa, M. 2002. Word Division and Orthography as Some of the Factors Posing Challenges in the Development of the Ndebele Grammatical Parser. Paper presented at the Seventh International Conference of the African Association for Lexicography, organized by the Dictionary Unit of South African English, Rhodes University, Grahamstown, Republic of South Africa, 8-10 July 2002.

Maung, Zin Maung & Yoshiki Mikami. 2008. A rule-based syllable segmentation of myanmar text. In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*, 51–58. Hyderabad, India: Asian Federation of Natural Language Processing.

Medina-Urrea, Alfonso. 2006. Affix Discovery by Means of Corpora: Experiments for Spanish, Czech, Ralámuli and Chuj. In Alexander Mehler & Reinhard Köhler (eds.), *Aspects of Automatic Text Analysis* (Studies in Fuzziness and Soft Computing 209), 277-299. Springer, Berlin.

Medina Urrea, Alfonso & E. C. Buenrostro Díaz. 2003. Características cuantitativas de la flexión verbal del Chuj. *Estudios de Lingüística Aplicada* 38. 15–31.

Mohanty, S., P. K. Santi & K. P. Das Adhikary. 2005. Analysis and design of Oriya morphological analyser: Some tests with orinet. In *Proceedings of symposium on Indian Morphology, Phonology and Language Engineering.* IIT Kharagpur, India.

Monson, Christian, Ariadna Font Llitjós, Vamshi Ambati, Lori Levin, Alon Lavie, Alison Alvarez, Roberto Aranovich, Jaime Carbonell, Robert Frederking, Erik Peterson & Katharina Probst. 2008. Linguistic Structure and Bilingual Informants Help Induce Machine Translation of Lesser-Resourced Languages. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, 2854-2859. Marrakech, Morocco.

Muhirwe, J. 2007. Computational Analysis of Kinyarwanda Morphology: The Morphological Alternations. In J.M. Kizza, J. Muhirwe, J. Aisbett, K. Getao, V. Mbarika, D. Patel & A.J. Rodrigues (eds.), *Special Topics in Computing and ICT Research: Strengthening the Role of ICT in Development*, 78-87. Kampala: Fountain Publishers. Also in International Journal of Computing and ICT Research 1(1):85-92.

Muhirwe, Jackson & Trond Trosterud. 2008. Finite state solutions for reduplication in Kinyarwanda language. In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*, 73–80. Hyderabad, India: Asian Federation of Natural Language Processing.

Müürisep, Kaili, Tiina Puolakainen, Kadri Muischnek, Mare Koit, Tiit Roosmaa & Heli Uibo. 2003. A new language for constraint grammar: Estonian. In *Recent Advances in Natural Language Processing*, 304-310. Borovets, Bulgaria.

Nakov, Preslav. 2003. Building an inflectional stemmer for Bulgarian. In *CompSysTech '03: Proceedings of the 4th international conference conference on Computer systems and technologies*, 419–424. New York, NY, USA: ACM.

Nelson, Hans J. 2004. A Two-Level Engine for Tagalog Morphology and a Structured XML Output for Pc-Kimmo. Brigham Young University masters thesis.

Nguyen, Thi Minh Huyen, Mathias Rossignol, Hong Phuong Le, Quang Thang Dinh, Xuan Luong Vu & Cam Tu Nguyen. 2008. Word segmentation of Vietnamese texts: a comparison of approaches. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco.

Oflazer, Kemal. 2009. Computational Morphology for Lesser-studied Languages. In S. Nirenburg (ed.), *Language Engineering for Lesser-Studied Languages* (NATO Science for Peace and Security Series: Sub-Series D: Information and Communication Security 21), 135-152. IOS Press.

Paikens, Pēteris. 2008. Lexicon-based morphological analysis of Latvian language. In *Proceedings of the 3rd Baltic Conference on Human Language Technologies, Kaunas, October, 2007*, 235-240. Vilnius: Institute of the Lithuanian Language, Vytautas Magnus University.

Patel, Chirag & Karthik Gali. 2008. Part-of-speech tagging for Gujarati using conditional random fields. In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*, 117–122. Hyderabad, India: Asian Federation of Natural Language Processing.

Pauw, G. De & G.-M. de Schryver. 2008. Improving the Computational Morphological Analysis of a Swahili Corpus for Lexicographic Purposes. *Lexikos* 18. 303–318.

Pauw, G. De, G-M. de Schryver & P.W. Wagacha. 2006. Data-driven part-of-speech tagging of Kiswahili. In *Proceedings of Text, Speech and Dialogue, 9th International Conference* (LNAI 4188), 197-204. Berlin: Springer-Verlag.

Pauw, G. De, P.W. Wagacha & D.A. Abade. 2007. Unsupervised induction of Dholuo word classes using maximum entropy learning. In *Proceedings of the First International Computer Science and ICT Conference (COSCIT 2007)*. Nairobi, Kenya: University of Nairobi.

Pisceldo, Femphy, Rahmad Mahendra, Ruli Manurung & I Wayan Arka. 2008. A two-level morphological analyser for the Indonesian language. In *Proceedings of the 2008 Australasian Language Technology Association Workshop (ALTA 2008)*, 142-150. Hobart, Australia.

Pretorius, L. & S. Bosch. 2009. Exploiting cross-linguistic similarities in Zulu and Xhosa computational morphology. In G. De Pauw, G.M. de Schryver & L. Levin (eds.), *Proceedings of the First Workshop on Language Technologies for African Languages (AfLaT 2009)*, 96-103. Athens, Greece: Association for Computational Linguistics.

Pretorius, Laurette & Sonja E. Bosch. 2003. Finite-State Computational Morphology: An Analyzer Prototype For Zulu. *Machine Translation* 8(3). 195–216.

Pretorius, Rigardt , Ansu Berg, Laurette Pretorius & Biffie Viljoen. 2009. Setswana tokenisation and computational verb morphology: Facing the challenge of a disjunctive orthography. In *Proceedings of the First Workshop on Language Technologies for African Languages*, 66–73. Athens, Greece: Association for Computational Linguistics.

Prinsloo, D.J. & U. Heid. 2006. Creating Word Class Tagged Corpora for Northern Sotho by Linguistically Informed Bootstrapping. In I. Ties (ed.), *Proceedings of the Lesser Used Languages and Computer Linguistics Conference (LULCL), Bolzano, 27-28 October 2005*, 97-115. EURAC Research.

Prinsloo, Daan J. 1994. Lemmatization of verbs in Northern Sotho. *South African Journal of African Languages* 14(2). 93–102.

Prószéky, G. & Attila Novák. 2005. Computational Morphologies for Small Uralic Languages. In Antti Arppe, Lauri Carlson, Krister Lindén, Jussi Piitulainen, Mickael Suominen, Martti Vainio, Hanna Westerlund & Anssi Yli-Jyrä (eds.), *Inquiries into Words, Constraints and Contexts: Festschrift for Kimmo Koskenniemi on his 60th Birthday*, 116-125. CSLI Publications.

Rama Sree, R.J., G. Uma Maheswara Rao & K. V. Madhu Murthy. 2008. Assessment and development of POS tag set for Telugu. In *The 6th Workshop on Asian Language Resources*, 85-88. Asian Federation of Natural Language Processing.

Ranaivo-Malançon, Bali. 2004. Computational Analysis of Affixed Words in Malay Language. UTMK, USM, Malaysia.

Ridings, D. & W. Mavhu. 2002. Problems and Challenges Encountered when Developing a Morphological Parser for the Shona Language. Paper presented at the Seventh International Conference of the African Association for Lexicography, organized by the Dictionary Unit of South African English, Rhodes University, Grahamstown, Republic of South Africa, 8-10 July 2002.

Rimkutė, Erika , Vidas Daudaravičius & Andrius Utka. 2007. Morphological annotation of the Lithuanian corpus. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing*, 94–99. Prague, Czech Republic: Association for Computational Linguistics.

Sankupellay, Mangalam & Subbu Valliappan. 2006. Malay-language stemmer. *Sunway Academic Journal* 3. 147–153.

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

Schmid, Helmut & Florian Laws. 2008. Estimation of conditional probabilities with decision trees and an application to fine-grained pos tagging. In *COLING-08*. ACL.

Shabadi, Kalyani R. 2003. Finite state morphological processing of Oriya verbal forms. In *Proceedings of the EACL 2003 Workshop on Computational Linguistics for South Asian Languages*, 49-56. Budapest, Hungary: Association for Computational Linguistics.

Sharada, B.A. & A. Lakshmi. 2006. Morphological Analyzer as NLP Tool for Kannada . In *Proceedings of the First National Symposium on Modeling and Shallow Parsing of Indian Languages MSPIL-06 (An event of NLP Association of India) 2-4 April 2006*, 8-10. Centre for Indian Language Technology, Department of Computer Science and Engineering, Indian Institute of Technology, Bombay, Mumbai, India.

Sharma, Utpal, Jugal Kalita & Rajib Das. 2002. Unsupervised learning of morphology for building lexicon for a highly inflectional language. In *Proceedings of the 6th Workshop of the ACL Special Interest Group in Computational Phonology (SIGPHON), Philadelphia, July 2002*, 1-10. Association for Computational Linguistics.

Shield, Aaron & Jason Baldridge. 2008. A morphological analyzer for verbal aspect in American Sign Language. In Nicholas Gaylord, Alexis Palmer & Elias Ponvert (eds.), *Computational Linguistics for Less-Studied Languages* (Texas Linguistics Society X), 125-138. Stanford: CSLI.

Simov, Kiril, Petya Osenova, Sia Kolkovska, Elisaveta Balabanova & Dimitar Doikoff. 2004. A Language Resources Infrastructure for Bulgarian. In *Proceedings of LREC 2004*, 1685-1688. Lisbon, Portugal.

Singh, Anil Kumar. 2008. Natural language processing for less privileged languages: Where do we come from? Where are we going? In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*, 7–12. Hyderabad, India: Asian Federation of Natural Language Processing.

Singh, Thoudam Doren & Sivaji Bandyopadhyay. 2006. Word Class and Sentence Type Identification in Manipuri Morphological Analyzer. In *Proceedings of the First National Symposium on Modeling and Shallow Parsing of Indian Languages MSPIL-06 (An event of NLP Association of India) 2-4 April 2006*, 11-17. Centre for Indian Language Technology, Department of Computer Science and Engineering, Indian Institute of Technology, Bombay, Mumbai, India.

Singh, Thoudam Doren & Sivaji Bandyopadhyay. 2008. Morphology driven Manipuri POS tagger. In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*, 91–98. Hyderabad, India: Asian Federation of Natural Language Processing.

Slavcheva, Milena. 2003. Some aspects of the morphological processing of bulgarian. In *Proceedings of the EACL 2003 Workshop on Morphological Processing of Slavic Languages*, 71-77. Budapest, Hungary: Association for Computational Linguistics.

Sornlertlamvanich, Virach. 2008. Invited talk: Cross language resource sharing. In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*, 3–4. Hyderabad, India: Asian Federation of Natural Language Processing.

Streiter, Oliver, Kevin P. Scannell & Mathias Stuflesser. 2006. Implementing NLP projects for noncentral languages: instructions for funding bodies, strategies for developers. *Machine Translation* 20(4). 267–289.

Sulger, Sebastian. 2008. Implementing a Finite-State Morphological Analyzer for Irish: Issues at the Morphology-Syntax Interface. In *Proceedings of the Seventh International Workshop on Finite-State Methods and Natural Language Processing (FSMNLP 2008)*. Konstanz.

Tai, Sock Yin, Cheng Soon Ong & Noor Aida Abullah. 2000. On designing an automated Malaysian stemmer for the Malay language. In *IRAL '00: Proceedings of the fifth international workshop on on Information retrieval with Asian languages*, 207-208. New York, NY, USA: ACM.

Tantuğ, A. Cüneyd, E. Adal & Kemal Oflazer. 2006. Computer Analysis of the Turkmen Language Morphology. In T. Salakoski, F. Ginter, S. Pyysalo & T Pahikkala (eds.), *Advances in Natural Language Processing: Proceedings*

of the 5th International Conference, FinTAL 2006 Turku, Finland, August 23-25, 2006 (LNCS 4139), 186-193. Berlin: Springer-Verlag.

Tongchim, Shisanu, Randolf Altmeyer, Virach Sornlertlamvanich & Hitoshi Isahara. 2008. A Dependency Parser for Thai. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, 136-139. Marrakech, Morocco.

Trosterud, Trond. 2008a. Faroese language technology. `http://giellatekno.uit.no/fao.html` accessed 24 Dec 2008.

Trosterud, Trond. 2008b. Greenlandic language technology. `http://giellatekno.uit.no/kal.html` accessed 24 Dec 2008.

Trosterud, Trond. 2008c. Iñupiaq language technology. `http://giellatekno.uit.no/ipk.html` accessed 24 Dec 2008.

Trosterud, Trond. 2008d. Sámi language technology. `http://giellatekno.uit.no/` accessed 24 Dec 2008.

Uí Dhonnchadha, Elaine, Caoilfhionn Nic Pháidín & Josef Van Genabith. 2003. Design, Implementation and Evaluation of an Inflectional Morphology Finite State Transducer for Irish. *Machine Translation* 18(3). 173–193.

Uibo, Heli. 2002. Experimental Two-Level Morphology of Estonian. In *Proceedings of LREC 2002: Third International Conference on Language Resources and Evaluation*, 1012-1015. Las Palmas, Gran Canaria.

Vikram, T. N. & Shalini R. Urs. 2007. Development of Prototype Morphological Analyzer for he South Indian Language of Kannada. In *Asian Digital Libraries. Looking Back 10 Years and Forging New Frontiers* (LNCS 4822), 109-116. Berlin: Springer-Verlag.

Viswanathan, S., S. Ramesh Kumar, B. Kumara Shanmugam & S. Arulmozi. 2003. A Tamil Morphological Analyser. In Rajeev Sangal, S.M. Bendre & Udaya Narayana Singh (eds.), *Recent Advances in Natural Language Processing: Proceedings of the International Conference Natural ICON-2003*, 31-39. Vikas Publishing House Pvt Ltd., New Delhi.

Wedekind, Charlotte & Klaus Wedekind. 2009. Beja: How well does the current parser handle input texts?. Paper presented at "Aethiopisches Forschungs-Colloquium", Berlin, 12-13 June 2009.

Wedekind, Klaus. 2008. A Report on the Automatic Parsing of Tigre. Paper prepared for the International Workshop 'History and Language of the Tigre-Speaking Peoples', Università di Napoli "L'Orientale", February 8th-9th 2008.

Wedekind, Klaus, Karen Buseman & Gary Unruh. 1983. A computer-tested minimal parsing-machine of Gedeo [Derasa]: Morphological rules and morpheme inventory for the automatic analysis of Gedeo texts. Manuscript available at The Institute of Ethiopian Studies, Addis Abeba, Ethiopia.

Wolfart, H. C. & F. Pardo. 1979. Computer-Aided Philology and Algorithmic Linguistics. *International Journal of American Linguistics* 45(2). 107–122.

Xia, Fei & William D. Lewis. 2008. Repurposing theoretical linguistic data for tool development and search. In *Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP 2008)*. Hyderabad, India: Asian Federation of Natural Language Processing.

Xia, Fei & William D. Lewis. 2009. Applying NLP technologies to the collection and enrichment of language data on the web to aid linguistic research. In *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education (LaTeCH – SHELT&R 2009)*, 51–59. Athens, Greece: Association for Computational Linguistics.

Yli-Jyrä, Anssi. 2005. Toward a Widely Usable Finite-State Morphology Workbench for Less Studied Languages – Part I: Desiderata. *Nordic Journal of African Studies* 14(4). 479–491.

Chapter III

Morphological Lexicon Extraction from Raw Text Data

Forsberg, M., Hammarström, H., and Ranta, A. (2006). Lexicon extraction from raw text data. In Salakoski, T., Ginter, F., Pyysalo, S., and Pahikkala, T., editors, *Advances in Natural Language Processing: Proceedings of the 5th International Conference, FinTAL 2006 Turku, Finland, August 23-25, 2006*, volume 4139 of *Lecture Notes in Computer Science*, pages 488–499. Springer-Verlag, Berlin.

Morphological Lexicon Extraction from Raw Text Data

Markus Forsberg, Harald Hammarström and Aarne Ranta
Department of Computer Science and Engineering
Chalmers University of Technology
and University of Gothenburg
SE-412 96 Göteborg, Sweden
markus,harald2,aarne@chalmers.se

Abstract

We introduce a tool *extract* developed for automatic extraction of lemma-paradigm pairs from raw text data. The tool combines regular expressions containing variables with propositional logic to form search patterns which identify lemmas tagged with their paradigm class. Furthermore, we describe the underlying algorithm of the tool and suggest a method for developing a morphological lexicon.

The tool was primarily developed for morphologies defined in the tool *Functional Morphology* (Forsberg and Ranta 2004), but it is usable for all similar systems that implement a word-and-paradigm description of a morphology.

We demonstrate the usefulness of the tool with a case study on the Canadian Hansards Corpus of French. The result is evaluated in terms of precision of the extracted lemmas and statistics on coverage and rule productiveness. Competitive extraction figures show that human-written rules in a tailored tool is a relatively time-efficient approach to the task at hand.

1 Introduction

A wide-coverage morphological lexicon is a key part of any information retrieval system, machine translation engine and of a variety of other Natural Language Processing applications. The demand is high not only for low-density languages, since existing lexica for major languages are often not publicly available. Moreover, even if they were, running text – especially newspaper and technical texts – will always contain out-of-lexicon words.

Manual development of a full-scale lexicon is a time-consuming task, so it is natural to investigate how the lexicon development can be automated. The situation is usually such that access to large collections of raw language data is

cheap, so cheap that it is tempting to look at ways to exploit the raw data to obtain the sought after high-quality morphological lexicon. Clearly, attempts to fully automatize the process (e.g (Creutz and Lagus 2005, Sharma et al. 2002) – most other systems for unsupervised learning of morphology cannot be used directly to build a lexicon) do not reach the kind of quality we are generally interested in. However, instead of using humans for supervised learning of lexicon extraction in some form, we believe there is a more advantageous placement of the human role. With a suitable tool, humans can use their knowledge to guide a computerized extraction from raw text, with comparatively little time spent. To be more specific, we intend to show that a profitable role for the human is to write intelligent extraction rules, given that there is a sufficiently flexible environment.

The *extract* tool has been developed with this in mind. The idea behind *extract* is simple: start with a large-sized corpus and a description of the word forms in the paradigms with the varying parts, which we refer to as *technical stems*, represented with variables. In the tool's syntax, we could describe the first declension noun of Swedish with the following definition.

```
paradigm decl1 =
  x+"a"
  { x+"a"  & x+"as"  & x+"an"  &  x+"ans" &
    x+"or" & x+"ors" & x+"orna & x+"ornas" } ;
```

Given that all forms in the curly brackets, called the *constraint*, are found for some prefix `x`, the tool outputs the *head* `x+"a"` tagged with the name of the paradigm. E.g., if these forms exist in the text data: `ärta`, `ärtas`, `ärtan`, `ärtans`, `ärtor`, `ärtors`, `ärtorna` and `ärtornas`, the tool will output `decl1 ärta`. Given that we have the lemma and the paradigm class, it is a relatively simple task to generate all word forms.

The paradigm definition has a major drawback: very few lemmas appear in all word forms. It could in fact be relaxed to increase recall without sacrificing precision: to identify a Swedish word as a noun of the first declension it is often enough to find one instance of the four singular forms and one of the four plural forms. The tool offers a solution by supporting propositional logic in the constraint, further described in section 2.1. Various issues of the extraction process are discussed in section 3.

Another problem with the given definition is the lack of control over what the variable `x` might be. Section 2.2 describes how the tool improves this situation by allowing variables to be associated with regular expressions.

The stems of first declension nouns in Swedish are the same for all word forms, but this is not the case for many paradigms, e.g. German nouns with umlaut. Section 2.3 presents the tool's use of multiple variables as a solution to this problem.

⟨*Def* ⟩ ::= `paradigm` ⟨*Name* ⟩ ⟨*VarDef* ⟩ =
⟨*Head* ⟩ `{` ⟨*Logic* ⟩ `}`

| `regexp` ⟨*Name* ⟩ = ⟨*Reg* ⟩

Figure 1. Regexp and paradigm definitions

⟨*Logic* ⟩ ::= ⟨*Logic* ⟩ `&` ⟨*Logic* ⟩`~` ⟨*Logic* ⟩
| ⟨*Logic* ⟩ | ⟨*Logic* ⟩ ⟨*Pattern* ⟩
| ⟨*Logic* ⟩ | `(` ⟨*Logic* ⟩ `)`

Figure 2. Propositional logic grammar

2 Paradigm File Format

A paradigm file consists of two kinds of definitions: `regexp` and `paradigm`, in figure 1.

A `regexp` definition associates a name (`Name`) with a regular expression (`Reg`). A `paradigm` definition consists of a name (`Name`), a set of variable-regular expression associations (`VarDef`), a set of output constituents (`Head`) and constraint (`Logic`).

The basic unit in `Head` and `Logic` is a *pattern* that describes a word form. A pattern consists of a sequence of variables and string literals glued together with the '+' operator. An example of a pattern given previously was `x+"a"`.

Both definitions will be discussed in detail in the following sections.

2.1 Propositional Logic

Propositional logic appears in the constraint to enable a more fine-grained description of what word forms the tool should look for. The basic unit is a pattern, corresponding to a word form, which is combined with the operators `&` (*and*), `|` (*or*), and `~` (*not*).

The syntax for propositional logic is given in figure 2, where *Pattern* refers to one word form.

The addition of new operators allow the paradigm in section 1 to be rewritten with disjunction to reflect that it is sufficient to find one singular and one plural word form.

```
paradigm decl1 =
      x+"a"
      { (x+"a"   | x+"as"  | x+"an"  | x+"ans") &
        (x+"or"  | x+"ors" |x+"orna  | x+"ornas") } ;
```

2.2 Regular Expressions

It was mentioned in section 1 that control over the variable part of a paradigm description was desired. The solution provided by the tool is to enable the user

⟨*Reg* ⟩	::=	⟨*Reg* ⟩ | ⟨*Reg* ⟩	`digit`
	|	⟨*Reg* ⟩ − ⟨*Reg* ⟩	`letter`
	|	⟨*Reg* ⟩ ⟨*Reg* ⟩	`upper`
	|	⟨*Reg* ⟩ `*`	`lower`
	|	⟨*Reg* ⟩ +	`char`
	|	⟨*Reg* ⟩ `?`	⟨*String* ⟩
	|	`eps`	`(` ⟨*Reg* ⟩ `)`
	|	⟨*Char* ⟩	

Figure 3. Regular expression

to associate every variable with a regular expression. The association dictates which (sub-)strings a variable can match. An unannotated variable can match any string, i.e. its regular expression is Kleene star over any symbol.

As a simple example, consider German, where nouns always start with an uppercase letter. This can be expressed as follows.

```
regexp UpperWord = upper letter*;

paradigm n [x:UpperWord] = ... ;
```

The syntax of the tool's regular expressions is given in figure 3, with the normal connectives: union, concatenation, set minus, Kleene star, Kleene plus and optionality. *eps* refers to the empty string, *digit* to $0 - 9$, *letter* to an alphabetic character, *lower* and *upper* to a lowercase respectively an uppercase letter. *char* refers to any character. A regular expression can also contain a double-quoted string, which is interpreted as the concatenation of the characters in the string.

2.3 Multiple Variables

Not all paradigm definitions are as neat as the initial example — phenomena like *umlaut* require an increased control over the variable part. The solution the tool provides is to allow multiple variables, i.e. a pattern may contain more than one variable. This is best explained with an example, where two German noun paradigms are described, both with umlaut. The change of the stem vowel is captured by introducing two variables and by letting the stem vowel be a constant string.

```
regexp Consonant = ... ;

regexp Pre = upper letter*;

regexp Aft = Consonant+ ;

paradigm n2 [F:Pre, ll:Aft] =
  F+"a"+ll
    { F+"a"+ll & F+"ä"+ll+"e" } ;
```

```
paradigm n3 [W:Pre, rt:Aft] =
   W+"o"+rt
    { W+"o"+rt & W+"ö"+rt+"er" } ;
```

The use of variables may reduce the time-performance of the tool, since every possible variable binding is considered. The use of multiple variables should be moderate, and the variables should be restricted as much as possible by their regular expression association to reduce the search space.

A variable does not need to occur in every pattern, but the tool only performs an initial match with patterns containing all variables. The reason for this is efficiency — the tool only considers one word at the time, and if the word matches one of the patterns, it searches for all other patterns with the variables instantiated by the initial match. For obvious reasons, an initial match is never performed under a negation, since this would imply that the tool searches for something it does not want to find.

It is allowed to have repeated variables, i.e. non-linear patterns, which is equivalent to *back reference* in the programming language Perl. An example where a sequence of bits is reduplicated is given. This language is known to be non-context-free (Hopcroft and Ullman 2001).

```
regexp ABs = (0|1)*;

paradigm reduplication [x:ABs] =
   x+x { x+x } ;
```

2.4 Multiple Arguments

The head of a paradigm definition may have multiple arguments to support more abstract paradigms. An example is Swedish nouns, where many nouns can be correctly classified by just detecting the word forms in nominative singular and nominative plural. An example is given below, where the first and second declension is handled with the same paradigm function, where the head consists of two output forms. The constraints are omitted.

```
paradigm regNoun =
flick+"a" flick+"or"
{...} ;
```

```
paradigm regNoun =
pojk+"e" pojk+"ar"
{...} ;
```

2.5 The Algorithm

The underlying algorithm of the tool is presented in pseudo-code notation.

```
let L be the empty lexicon.
let P be the set of extraction paradigms.
let W be all word types in the corpus.
for each w : W
 for each p : P
  for each constraint C with which w matches p
```

```
if W satisfies C with the result H,
 add H to L
```

The algorithm is initialized by reading the word types of the corpus into an array W. A word w *matches* a paradigm p, if it can match any of the patterns in the paradigm's constraint that contains all variables occurring in the constraint. The result of a successful match is an *instantiated constraint* C, i.e. a logical formula with words as atomic propositions. The corpus W *satisfies* a constraint C if the formula is true, where the truth of an atomic proposition a means that the word a occurs in W.

2.6 The Performance of the Tool

The extraction tool is implemented in Haskell. It is available as an open-source free software. A typical example of using the tool, the experiment reported in Section 4 extracted a lexicon of 19,295 lemmas from a corpus of 66,853 word types, by using 43 paradigms. The execution time was 22min 36s on a laptop running Mandrake Linux 9.2 with a 1.4GHz Pentium M. The memory consumption was 46MB.

3 The Art of Extraction

The constraint of a paradigm describes a sub-paradigm, a subset of the word forms, considered to be evidence enough to be able to judge that the lemmas in the head are in that paradigm class. The identification of appropriate sub-paradigms requires good insights into the target language and intuitions about the distributions of the word forms. However, these insights and intuitions may be acquired while using the tool by trial and error.

Lexicon extraction is a balance between *precision*, i.e. the percentage of the extracted lemmas that are correctly classified, and *recall*, i.e. the percentage of the lemmas in the text data that are extracted. Precision, however, is by far the most important, since poor recall can be compensated with more text data, but poor precision requires more human labor.

How about extracting the paradigm descriptions from a set of paradigms automatically? We use the term *minimum-size sub-paradigm* to describe the minimum-sized set of word forms needed to uniquely identify a paradigm P. More formally, a minimum-sized sub-paradigm is a minimum-size set of word forms $P' \subseteq P$ such that for any other paradigm Q, $P' \not\subseteq Q$. It turns out that the problem of finding the minimum-size sub-paradigm for a paradigm P is NP-complete (proof in section 3.1), and thus has a theoretical worst-case complexity exponential in $|P|$. Therefore there is all the more reason to let a human choose which forms to require and also weigh in which forms are likely to be common/uncommon in actual usage.

Also, some natural languages have *overshadowed paradigms*, i.e. paradigms where the form of one paradigm is a subset of another paradigm. For example, in Latin some noun paradigms are overshadowed by adjective paradigms.

The distinction of Latin nouns and adjectives can be done through the use of negation where a second declension noun paradigm is defined by also stating that the feminine endings, which would indicate that it is an adjective, should not be present. If the set of word forms are equal then there is no way to distinguish the words morphologically.

```
paradigm decl2servus =
       serv++"us"
   { serv+"us" & serv+"i" &
      ~(serv+"a" | serv+"ae")};
```

Negation is similar with *negation as failure* in Prolog, with the same problems associated with it. The main problem is that negation rests on the absence, not the presence, of information, which in turn means that the extraction process with negation is non-monotonic: the use of a larger corpus may lead to an extracted lexicon which is smaller. A worst-case scenario is a misspelt or foreign word that, by negation, removes large parts of the correctly classified lemmas in the extracted lexicon.

In most cases, a better alternative to negation is a more careful use of regular expressions, and in the case of Latin nouns, a rudimentary POS tagger that resolves the POS ambiguity may outperform negation.

3.1 Sub-Paradigm Problem is NP Complete

The minimum-size sub-paradigm problem (MSS) is equivalent to the well-known set-cover problem (Garey and Johnson 1979). We repeat the problem definitions for clarity:

SET-COVER: Input: a collection of sets $C = C_1, \ldots, C_n$ all being subsets of some universe $U = \{1, \ldots, k\}$ with $\bigcup C = U$. Goal: Find a minimum-size subcollection $C' \subseteq C$ such that $\bigcup C' = U$.

MSS: Input: A paradigm (set) P and a collection of sets $P_1, \ldots, P_m$ with $P \not\subseteq P_j$ for any j. Goal: Find a minimum-size $P' \subseteq P$ such that $P' \not\subseteq P_j$ for any j.

The following polynomial transformations translate between the two problems:

- To solve SET-COVER with MSS: Let $P = U$ and $P_j = \{i | j \notin C_i\}$ for $j \in P$. An output P' gives a minimum-size cover $C' = \{C_i | i \in P'\}$.
- To solve MSS with SET-COVER: Let $C_i = \{j | i \notin P_j\}$ for all $i \in P$. An output C' gives a minium-size sub-paradigm $P' = \{i | C_i \in C'\}$.

It is instructive to model the problem(s) in terms of a matrix. Rows represent elements and columns stand for sets of these elements. We put a plus at (x, y) if element x in contained in the set at y and a minus otherwise. Now we can think of set-cover as a selection of as few columns as possible such that there

	P_1	P_2
1	+	-
2	-	+
3	-	-

	1	2	3
P_1	-	+	+
P_2	+	-	+

Figure 4. Example correspondence between MSS and SET-COVER.

is a plus in every row in at least one of the selected columns. Likewise, the sub-paradigm problem as selection of as few rows as possible such that there is a minus in every column in at least one of the selected rows.

To see what is going on, consider the following example: Let $P = \{1, 2, 3\}$, $P_1 = \{1\}$ and $P_2 = \{2\}$. Clearly, $P' = \{3\}$ is the minimal solution. The corresponding matrix is given to the left in figure 4. Intuitively, selecting e.g 1 for our P' means that we can assure $P' \not\subset P_2$. Selecting e.g 3 for our P' means that we can assure both $P' \not\subset P_2$ and $P' \not\subset P_2$. Since selecting 3 covers every column with a minus, we are finished.

The transformations simply rotate the matrix and exchange the minus and plus:es. Thus, the corresponding set-cover instance is shown to the right in figure 4, and easily gives $C' = \{3\}$ as the minimal collection (from $C = \{1, 2, 3\}$).

3.2 Manual Verification

Almost all corpora have misspellings which may lead to false conclusions. Added to that are word forms that incidentally coincide. One possible solution to handle misspellings is to only consider words that occur at some frequency. However, that would remove a lot of unusual but correctly spelled words (to an extent which is unacceptable).

Misspellings, foreign words and coincidences are the main reason why manual verification of the extracted lexicon cannot be circumvented even with "perfect" paradigm definitions. However, browse-filtering a high-precision extracted lexicon requires much less time than building the same lexicon by hand. Also, nothing in principle prohibits statistical techniques to be applied in collaboration here. For instance, one can sort the extracted lemmas heuristically according to how many forms and with what frequencies they occur (cf. section 5). In general, this is productive for poly-occurring lemmas but helps little for the (typically many) hapax lemmas.

4 Experiments

We will evaluate our proposed extraction technique with a study of real-world extraction on the Hansards corpus of Canadian French (Germann 2003). All words were manually annotated to enable a thorough evaluation. However, the intended practical usage of the extraction tool is to simply eye-browse the output list for erroneous extractions.

Tokens	15 000 000
Types	66 853
Non-junk types	49 477
Lemmas	27 681

Figure 5. Statistics on the corpus of Canadian French Hansards used in the experiment

The corpus consisted of approximately 15 million running tokens of 66853 types. From these 66853 types we manually removed all junk – foreign words, proper names, misspellings, numeric expressions, abbreviations as well as pronouns, prepositions, interjections and non-derived adverbs – so that a 49477 true lexical items remained. 27681 lemmas account for the 49477 forms, where verb-lemmas tended to occur in more forms than noun- and adjective-lemmas. Of course, not all these lemmas occurred in such forms that their morphological class could be recognized by their endings alone. Many lemmas occur in only one form – usually not enough to infer its morphological class – unless, as is often the case, they contain a derivational morpheme which, together with its inflectional ending, does suffice. For example, a single occurrence of a word ending in *-e* is hardly conclusive, whereas one ending in *-tude* is almost certainly a feminine noun with a plural in *-s*. Nouns without derivational ending cannot be reliably distinguished from adjectives even when they occur in all their forms, i.e both the singular and plural. The table in figure 5 summarizes these data.

We now turn to the question of precision and coverage of rule-extraction of the targeted 27 681 lemmas. We quickly devised a set of 43 rules to extract French nouns (18 rules), verbs (7 rules) and adjectives (18 rules). The verb-rules aimed at *-ir* and *-er* verbs by requiring salient forms for these paradigms, whereas the noun- and adjective rules make heavy use of regularities in derivational morphology to overcome the problems of overlapping forms. Two typical example groups are given below:

```
regexp NOTi = char* (char-"i") ;

paradigm Ver [regard:NOTi]
  = regard+"er"
    {regard+"e" &
     (regard+"é"   | regard+"ée"  |
      regard+"ez"  | regard+"ont" |
      regard+"ons" | regard+"a" )} ;

paradigm Aif
  = sport+"if"
    {sport+"if"  | sport+"ifs" |
     sport+"ive" | sport+"ives"} ;
```

The results of the extraction are shown in figure 6. If possible, one would like to know where one's false positives come from – sloppy rules or noisy data?

	Extr. All	Extr. Non-Junk
False Positives	2031	664
Correctly Indentified	17264	17264
	19295	17928
Precision	89.5%	96.3%

Figure 6. Extraction results on raw text vs. text with junk removed first.

At least one would like to know roughly what to expect. Since we have already annotated this corpus we can give some indicative quantitative data. To assess the impact of misspellings and foreign words – the two main sources for spurious extractions – we show the results of the same extraction performed on the corpus *with all junk removed beforehand.* As expected, false positives increase when junk is added. To be more precise, we get a lot of spurious verbs from English words and proper names in *-er* (e.g farmer, worchester) as well as many nouns, whose identification requires only one form, from misspellings (e.g qestion). Non-junk-related cases of confusion worth mentioning are nouns in *-ment* – the same ending as adverbs – and verbs which have spelling changes (manger-mangeait, appeler-appelle etc).

The rule productiveness, i.e a rule on average catches $17264/43 \approx 401$, must be considered very high. As for coverage, we can see that our rules catch the lions share of the available lemmas, 17264 out of 27 681 (again, not all of which occur in enough forms to predict their morphological class), in the corpus. This is relevant because even if we can always find more raw text cheaply, we want our rules to make maximal use of whatever is available and more raw data is of little help unless we can actually extract a lot of them with reasonable effort. It is also relevant because a precision figure without a recall figure means nothing. It would be easy to tailor 43 rules to perfect precision, perhaps catching one lemma per rule, so what we show is that precision and rule productiveness can be simuntaneously high. In general it is of course up to the user how much of the raw-data lemmas to sacrifice for precision and rule-writing effort, which are usually more important objectives.

5 Related Work

The most important work dealing with the very same problem addressed here, i.e extracting a morphological lexicon given a morphological description, is the study of the acquisition of French verbs and adjectives in (Clément et al. 2004). Likewise, they start from an existing inflection engine and exploit the fact that a new lemma can be inferred with high probability if it occurs in raw text in predictable morphological form(s). Their algorithm ranks hypothetical lemmas based on the frequency of occurrence of its (hypothetical) forms as well as part-of-speech information signalled from surrounding closed-class words. They do not make use of human-written rules but reserve an unclear, yet crucial, role

for the human to hand-validate parts of output and then let the algorithm reiterate. Given the many differences, the results cannot be compared directly to ours but rather illustrate a complementary technique.

Tested on Russian and Croat, (Oliver and Tadić 2004, Oliver 2004:Ch. 3) describe a lexicon extraction strategy very similar to ours. In contrast to human-made rules, they have rules extracted from an existing (part of) a morphological lexicon and use the number of inflected forms found to heuristically choose between multiple lemma-generating rules (additionally also querying the Internet for existence of forms). The resulting rules appear not at all as sharp as hand-made rules with built-in human knowledge of the paradigms involved and their respective frequency (the latter being crucial for recall). Also, in comparison, our search engine is much more powerful and allows for greater flexibility and user convenience.

For the low-density language Assamese, (Sharma et al. 2002) report an experiment to induce both morphology and a morphological lexicon at the same time. Their method is based on segmentation and alignment using string counts only – involving no human annotation or intervention inside the algorithm. It is difficult to assess the strength of their acquired lexicon as it is intertwined with induction of the morphology itself. We feel that inducing morphology and extracting a morphological lexicon should be performed and evaluated separately.

There is a body of work on inducing verb subcategorization information from raw or tagged text (see Kermanidis et al. (2004), Faure and Nédellec (1998), Gamallo et al. (2003) and references therein). However, the parallel between subcategorization frame and morphological class is only lax. The latter is a simple mapping from word forms to a paradigm membership, whereas in verb subcategorization one also has the onus discerning which parts of a sentence are relevant to a certain verb. Moreover, it is far from clear that verb subcategorization comes in well-defined paradigms – instead the goal may be to reduce the amount of parse trees in a parser that uses the extracted subcategorization constraints.

6 Conclusions and Further Work

We have shown that building a morphological lexicon requires relatively little human work. Given a morphological description, typically an inflection engine and a description of the closed word classes, such as pronouns and prepositions, and access to raw text data, a human with knowledge of the language can use a simple but versatile tool that exploits word forms alone. It remains to be seen to what extent syntactic information, e.g part-of-speech information, can further enhance the performance. A more open question is whether the suggested approach can be generalized to collect linguistic information of other kinds than morphology, such as e.g verb subcategorization frames.

References

Clément, L., Sagot, B., and Lang, B. (2004). Morphology based automatic acquisition of large-coverage lexica. In *Proceedings of LREC'04*, pages 1841–1844, Lisboa, Portugal.

Creutz, M. and Lagus, K. (2005). Inducing the morphological lexicon of a natural language from unannotated text. In *Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR '05), 15-17 June, Espoo, Finland*, pages 106–113. Espoo.

Faure, D. and Nédellec, C. (1998). Asium: Learning subcategorization frames and restrictions of selection. In Kodratoff, Y., editor, *10th Conference on Machine Learning (ECML 98) – Workshop on Text Mining, Chemnitz, Germany, Avril 1998*. Springer-Verlag, Berlin.

Forsberg, M. and Ranta, A. (2004). Functional morphology. In *Proceedings of the Ninth ACM SIGPLAN International Conference of Functional Programming, Snowbird, Utah*, pages 213–223.

Gamallo, P., Agustini, A., and Lopes, G. P. (2003). Learning subcategorisation information to model a grammar with "co-restrictions". *Traitement Automatique des Langues*, 44(1):93–177.

Garey, M. R. and Johnson, D. S. (1979). *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman, New York.

Germann, U. (2003). Corpus of hansards of the 36th parliament of canada. Provided by the Natural Language Group of the Universtity of Southern California Information Sciences Institute. Downloadable at `http://www.isi.edu/natural-language/download/hansard/`, accessed 1 Nov 2005. 15 million words.

Hopcroft, J. and Ullman, J. (2001). *Introduction to Automata Theory, Languages, and Computation*. Addison-Wesley, 2 edition.

Kermanidis, K. L., Fakotakis, N., and Kokkinakis, G. (2004). Automatic acquisition of verb subcategorization information by exploiting minimal linguistic resources. *International Journal of Corpus Linguistics*, 9(1):1–28.

Oliver, A. (2004). *Adquisició d'informació lèxica i morfosintàctica a partir de corpus sense anotar: aplicació al rus i al croat*. PhD thesis, Universitat de Barcelona.

Oliver, A. and Tadić, M. (2004). Enlarging the croatian morphological lexicon by automatic lexical acquisition from raw corpora. In *Proceedings of LREC'04*, pages 1259–1262, Lisboa, Portugal.

Sharma, U., Kalita, J., and Das, R. (2002). Unsupervised learning of morphology for building lexicon for a highly inflectional language. In *Proceedings of the 6th Workshop of the ACL Special Interest Group in Computational Phonology (SIGPHON), Philadelphia, July 2002*, pages 1–10. Association for Computational Linguistics.

Chapter IV

Automatic Annotation of Bibliographical References with Target Language

Hammarström, H. (2008). Automatic annotation of bibliographical references with target language. In *Proceedings of MMIES-2: Wokshop on Multi-source, Multilingual Information Extraction and Summarization*, pages 57–64. ACL.

Automatic Annotation of Bibliographical References with Target Language

Harald Hammarström
Department of Computer Science and Engineering
Chalmers University of Technology
and University of Gothenburg
SE-412 96 Göteborg, Sweden
harald2@chalmers.se

Abstract

In a large-scale project to list bibliographical references to all of the ca 7 000 languages of the world, the need arises to automatically annotate the bibliographical entries with ISO-639-3 language identifiers. The task can be seen as a special case of a more general Information Extraction problem: to classify short text snippets in various languages into a large number of classes. We will explore supervised and unsupervised approaches motivated by distributional characterists of the specific domain and availability of data sets. In all cases, we make use of a database with language names and identifiers. The suggested methods are rigorously evaluated on a fresh representative data set.

1 Introduction

There are about 7 000 languages in the world (Hammarström 2008) and there is a quite accurate database of which they are (Gordon 2005). Language description, i.e., producing a phonological description, grammatical description, wordlist, dictionary, text collection or the like, of these 7 000 languages has been on-going on a larger scale since about 200 years. This process is fully de-centralized, and at present there is no database over which languages of the world have been described, which have not, and which have partial descriptions already produced (Hammarström 2007b). We are conducting a large-scale project of listing all published descriptive work on the languages of the world, especially lesser-known languages. In this project, the following problem naturally arises:

Given: A database of the world's languages (consisting minimally of <unique-id, language-name>-pairs)

Input: A bibliographical reference to a work with descriptive language data of (at least one of) the language(s) in the database

Desired output: The identification of which language(s) is described in the bibliographical reference

We would like to achieve this with as little human labour as possible. In particular, this means that thresholds that are to be set by humans are to be avoided. However, we will allow (and do make use of – see below) supervision in the form of databases of language references annotated with target language as long as they are *freely available.*

As an example, say that we are given a bibliographical reference to a descriptive work as follows:

> Dammann, Ernst 1957 *Studien zum Kwangali: Grammatik, Texte, Glossar*, Hamburg: Cram, de Gruyter & Co. [Abhandlungen aus dem Gebiet der Auslandskunde / Reihe B, Völkerkunde, Kulturgeschichte und Sprachen 35]

This reference happens to describe a Namibian-Angolan language called Kwangali [kwn]. The task is to automatically infer this, for an arbitrary bibliographical entry in an arbitrary language, using the database of the world's languages and/or databases of annotated entries, but without humanly tuned thresholds. (We will assume that the bibliographical entry comes segmented into fields, at least as to the title, though this does not matter much.)

Unfortunately, the problem is not simply that of a clean database lookup. As shall be seen, the distributional characteristics of the world language database and input data give rise to a special case of a more general Information Extraction (IE) problem. To be more precise, an abstract IE problem may be defined as follows:

- There is a set of natural language objects O
- There is a fixed set of categories C
- Each object in O belong to zero or more categories, i.e., there is a function $C : O \rightarrow Powerset(C)$
- The task is to find classification function f that mimics C.

The special case we are considering here is such that:

- Each object in O contains a small amount of text, on the order of 100 words
- The language of objects in O varies across objects, i.e., not all objects are written in the same language
- $|C|$ is large, i.e., there are many classes (about 7 000 in our case)

- $|C(o)|$ is small for most objects $o \in O$, i.e., most objects belong to very few categories (typically exactly one category)
- Most objects $o \in O$ contain a few tokens that near-uniquely identifies $C(o)$, i.e., there are some words that are very informative as to category, while the majority of tokens are very little informative. (This characteristic excludes the logical possibility that each token is fairly informative, and that the tokens *together*, on an equal footing, serve to pinpoint category.)

We will explore and compare ways to exploit these skewed distributional properties for more informed database lookups, applied and evaluated on the outlined reference-annotation problem.

2 Data and Specifics

The exact nature of the data at hand is felt to be quite important for design choices in our proposed algorithm, and is assumed to be unfamiliar to most readers, wherefore we go through it in some detail here.

2.1 World Language Database

The Ethnologue (Gordon 2005) is a database that aims to catalogue all the known living languages of the world.[1] As far as language inventory goes, the database is near perfect and language/dialect divisions are generally accurate, though this issue is thornier (Hammarström 2005).

Each language is given a unique three-letter identifier, a canonical name and a set of variant and/or dialect names.[2] The three-letter codes are draft ISO-639-3 standard. This database is freely downloadable[3]. For example, the entry for Kwangali [kwn] contains the following information:

> Canonical name: Kwangali
> ISO 639-3: kwn
> Alternative names[4]: {Kwangali, Shisambyu, Cuangar, Sambio, Kwangari, Kwangare, Sambyu, Sikwangali, Sambiu, Kwangali, Rukwangali}.

The database contains 7 299 languages (thus 7 299 unique id:s) and a total of 42 768 name tokens. Below are some important characteristics of these collections:

[1] It also contains some sign languages and some extinct attested languages, but it does not aim or claim to be complete for extinct and signed languages.

[2] Further information is also given, such as number of speakers and existence of a bible translation is also given, but is of no concern for the present purposes.

[3] From `http://www.sil.org/iso639-3/download.asp` accessed 20 Oct 2007.

[4] The database actually makes a difference between dialect names and other variant names. In this case Sikwangali, Rukwangali, Kwangari, Kwangare are alternname names denoting Kwangali, while Sambyu is the name of a specific dialect and Shisambyu, Sambiu, Sambio are variants of Sambyu. We will not make use of the distinction between a dialect name and some other alternative name.

- Neither the canonical names nor the alternative names are guaranteed to be unique (to one language). There are 39 419 unique name strings (but 42 768 name tokens in the database!). Thus the average number of different languages (= unique id:s) a name denotes is 1.08, the median is 1 and the maximum is 14 (for Miao).

- The average number of names (including the canonical name) of a language is 5.86, the median is 4, and the maximum is 77 (for Armenian [hye]).

- It is not yet well-understood how complete database of alternative names is. In the preparation of the test set (see Section 2.4) an attempt to estimate this was made, yielding the following results. 100 randomly chosen bibliographical entries contained 104 language names in the title. 43 of these names (41.3%) existed in the database as written. 66 (63.5%) existed in the database allowing for variation in spelling (cf. Section 1). A more interesting test, which could not be carried out for practical reasons, would be to look at a language and gather *all* publications relating to that language, and collect the names occurring in titles of these. (To collect the full range of names denoting languages used in the bodies of such publications is probably not a well-defined task.) The Ethnologue itself does not systematically contain bibliographical references, so it is not possible to deduce from where/how the database of alternative names was constructed.

- A rough indication of the ratio between spelling variants versus alternative roots among alternative names is as follows. For each of the 7299 sets of alternative names, we conflate the names which have an edit distance[5] of $\leq i$ for $i = 0, \ldots, 4$. The mean, median and max number of names after conflating is shown below. What this means is that languages in the database have about 3 names on average and another 3 spelling variants on average.

i	Mean	Median	Max	Entry
0	5.86	4	77	'hye'
1	4.80	3	65	'hye'
2	4.07	3	56	'eng'
3	3.41	2	54	'eng'
4	2.70	2	47	'eng'

2.2 Bibliographical Data

Descriptive data on the languages of the world are found in books, PhD/MA theses, journal articles, conference articles, articles in collections and manuscripts. If only a small number of languages is covered in one publication, the title usually carries sufficient information for an experienced human to deduce which language(s) is covered. On the other hand, if a larger number of languages is

[5] Penalty weights set to 1 for deletion, insertion and substitution alike.

targeted, the title usually only contains approximate information as to the covered languages, e.g., *Talen en dialecten van Nederlands Nieuw-Guinea* or *West African Language Data Sheets.* The (meta-)language [as opposed to target language] of descriptive works varies (cf. Section 2.4).

2.3 Free Annotated Databases

Training of a classifier ('language annotator') in a supervised framework, requires a set of annotated entries with a distribution similar to the set of entries to be annotated. We know of only two such databases which can be freely accessed[6]; WALS and the library catalogue of MPI/EVA in Leipzig.

WALS: The bibliography for the *World Atlas of Language Structures* book can now be accessed online (`http://www.wals.info/`). This database contains 5633 entries annotated to 2053 different languages.

MPI/EVA: The library catalogue for the library of the Max Planck Institute for Evolution Anthropology (`http://biblio.eva.mpg.de/`) is queryable online. In May 2006 it contained 7266 entries annotated to 2246 different languages.

Neither database is free from errors, imprecisions and inconsistencies (impressionistically 5% of the entries contain such errors). Nevertheless, for training and development, we used both databases put together. The two databases put together, duplicates removed, contain 8584 entries annotated to 2799 different languages.

2.4 Test Data

In a large-scale on-going project, we are trying to collect all references to descriptive work for lesser-known languages. This is done by tediously going through handbooks, overviews and bibliographical for all parts of the world alike. In this bibliography, the (meta-)language of descriptive data is be English, German, French, Spanish, Portuguese, Russian, Dutch, Italian, Chinese, Indonesian, Thai, Turkish, Persian, Arabic, Urdu, Nepali, Hindi, Georgian, Japanese, Swedish, Norwegian, Danish, Finnish and Bulgarian (in decreasing order of incidence)[7]. Currently it contains 11788 entries. It is this database that needs to be annotated as to target language. The overlap with the joint WALS-MPI/EVA

[6] For example, the very wide coverage database worldcat (`http://www.worldcat.org/`) does not index individual articles and has insufficient language annotation; sometimes no annotation or useless categories such as 'other' or 'Papuan'. The SIL Bibliography (`http://www.ethnologue.com/bibliography.asp`) is well-annotated but contains only work produced by the SIL. (SIL has, however, worked on very many languages, but not all publications of the de-centralized SIL organization are listed in the so-called SIL Bibliography.)

[7] Those entries which are natively written with a different alphabet always also have a transliteration or translation (or both) into ascii characters.

database is 3984 entries.[8] Thus $11788 - 3984 = 7804$ entries remain to be annotated. From these 7 804 entries, 100 were randomly selected and humanly annotated to form a test set. This test set was not used in the development at all, and was kept totally fresh for the final tests.

3 Experiments

We conducted experiments with three different methods, plus the enhancement of spelling variation on top of each one.

Naive Lookup: Each word in the title is looked up as a possible language name in the world language database and the output is the union of all answers to the look-ups.

Term Weight Lookup: Each word is given a weight according to the number of unique-id:s it is associated with in the training data. Based on these weights, the words of the title are split into two groups; informative and non-informative words. The output is the union of the look-up:s of the informative words in the world language database.

Term Weight Lookup with Group Disambiguation: As above, except that names of genealogical (sub-)groups and country names that occur in the title are used for narrowing down the result.

Following a subsection on terminology and definitions, these will be presented in increasing order of sophistication.

3.1 Terminology and Definitions

- C: The set of 7 299 unique three-letter language id:s
- N: The set of 39 419 language name strings in the Ethnologue (as above)
- $C(c)$: The set of names $\subseteq N$ associated with the code $c \in C$ in the Ethnologue database (as above)
- $LN(w) = \{id | w \in C(id), id \in C\}$: The set of id:s $\subseteq C$ that have w as one of its names
- $C_S(c) = \cup_{winC(c)} Spellings(w)$: The set of variant spellings of the set of names $\subseteq N$ associated with the code $c \in C$ in the Ethnologye database. For reference, the $Spelling(w)$-function is defined in detail in Table 1.

[8] This overlap at first appears surprisingly low. Part of the discrepancy is due to the fact that many references in the WALS database are in fact to secondary sources, which are not intended to be covered at all in the on-going project of listing. Another reason for the discrepancy is due to a de-prioritization of better-known languages as well as dictionaries (as opposed to grammars) in the on-going project. Eventually, all unique references will of course be merged.

- $LN_S(w) = \{id|w \in C_S(id), id \in C\}$: The set of id:s $\subseteq C$ that have w as a possible spelling of one of its names
- WE: The set of entries in the joint WALS-MPI/EVA database (as above). Each entry e has a title e_t and a set e_c of language id:s $\subseteq C$
- $Words(e_t)$: The set of words, everything lowercased and interpunctation removed, in the title e_t
- $LWEN(w) = \{id|e \in WE, w \in e_t, id \in e_c\}$: The set of codes associated with the entries whose titles contain the word w
- $TD(w) = LN(w) \cup LWEN(w)$: The set of codes tied to the word w either as a language name or as a word that occurs in a title of an code-tagged entry (in fact, an Ethnologue entry can be seen as a special kind of bibliographical entry, with a title consisting of alternative names annotated with exactly one category)
- $TD_S(w) = LN_S(w) \cup LWEN(w)$: The set of codes tied to the word w either as a (variant spelling of a) language name or as a word that occurs in a title of an code-tagged entry
- $WC(w) = |TD(w)|$: The number of different codes associated with the word w
- $WI(w) = |\{e_t|w \in Words(e_t), e_t \in WE\}|$: The number of different bibliographical entries for which the word w occurs in the title
- A: The set of entries in the test set (as above). Each entry e has a title e_t and a set e_c of language id:s $\subseteq C$
- $PA_A(X) = \frac{|\{e|X(e)==e_c, e \in A\}|}{|A|}$: The perfect accuracy of a classifier function X on test set A is the number of entries in A which are classified correctly (the sets of categories have to be fully equal)
- $SA_A(X) = \sum_{e \in A} \frac{|\{X(e) \cap e_c\}|}{|e_c \cup X(e)|}$: The sum accuracy of a classifier function X on a test set A is the sum of the (possibly imperfect) accuracy of the entries of A (individual entries match with score between 0 and 1)

3.2 Naive Union Lookup

As a baseline to beat, we define a naive lookup classifier. Given an entry e, we define naive union lookup (NUL) as:

$$NUL(e) = \cup_{w \in Words(e_t)} TD(w)$$

For example, consider the following entry e:

#	Substition Reg. Exp.	Replacement	Comment
1.	`\'\‘\^\~\"`	`''`	diacritics truncated
2.	`[qk](?=[ei])`	`qu`	k-sound before soft vowel to qu
3.	`k(?=[aou]|$)|q(?=[ao])`	`c`	k-sound before hard vowel to c
4.	`oo|ou|oe`	`u`	oo, ou, oe to u
5.	`[hgo]?u(?=[aouei]|$)`	`w`	hu-sound before hard vowel to w
6.	`((?:[^aouei]*[aouei][^aouei]*)+?)(?:an$\|ana$|ano$\|o$)`	`\1a`	an? to a
7.	`eca$`	`ec`	eca to ec
8.	`tsch|tx|tj`	`ch`	tsch, tx to ch
9.	`dsch|dj`	`j`	dsch, dj to j
10.	`x(?=i)`	`sh`	x before i to sh
11.	`i(?=[aouei])`	`y`	i before a vowel to y
12.	`ern$\|i?sche?$`	`''`	final sche, ern removed
13.	`([a-z])\1`	`\1`	remove doublets
14.	`[bdgv]`	`b/p,d/t,g/k,v/f`	devoice b, d, g, v
15.	`[oe]`	`o/u,e/i`	lower vowels

Table 1. Given a language name w, its normalized spelling variants are enumerate according to the following (ordered) list of substitution rules. The set of spelling variants $Spelling(w)$ should be understood as the strings $\{w/action_{1-i} | i \leq 15\}$, where $w/action_{1-i}$ is the string with substitutions 1 thru i carried out. This normalization scheme is based on extensive experience with language name searching by the present author.

$Words(e_t)$	$LN(Words(e_t))$	$Words(e_t)$	$LN(Words(e_t))$
etude	{}	cameroun	{}
du	{*dux*}	du	{*dux*}
samba	{*ndi*, *ccg*, *smx*}	nord	{}
leko	{*ndi*, *lse*, *lec*}	famille	{}
parler	{}	adamawa	{}
d'allani	{}		

Table 2. The calculation of NUL for an example entry

Anne Gwenaïélle Fabre 2002 *Étude du Samba Leko, parler d'Allani (Cameroun du Nord, Famille Adamawa)*, PhD Thesis, Université de Paris III – Sorbonne Nouvelle

The steps in its NUL-classification is as follows are given in Table 2.

Finally, $NUL(e) = \{ndi, lse, smx, dux, lec, ccg\}$, but, simply enough, $e_c = \{ndi\}$.

The resulting accuracies for the test set are $PA_{NUL}(A) \approx 0.15$ and $SA_{NUL}(A) \approx 0.21$. NUL performs even worse with spelling variants enabled. Not surprisingly, NUL overclassifies a lot, i.e., it consistently guesses more languages than is the case. This is because guessing that a title word indicates a target language just because there is one language with such a name, is not a sound practice. In fact, common words like *du* [dux], *in* [irr], *the* [thx], *to* [toz], and *la* [wbm, lic, tdd] happen to be names of languages (!).

3.3 Term Weight Lookup

We learn from the Naive Union Lookup experiment that we cannot guess blindly which word(s) in the title indicate the target language. Something has to be done to individuate the informativeness of each word. Domain knowledge tells us two relevant things. Firstly, a title of a publication in language description typically contains one or few words with very precise information on the target language(s), namely the name of the language(s), and in addition a number of words which recur throughout many titles, such as 'a', 'grammar', etc. Secondly, most of the language of the world are poorly described, there are only a few, if any, publications with original descriptive data. Inspired by the tf-idf measure in Information Retrieval (Baeza-Yates and Ribeiro-Neto 1997), we claim that informativeness of a word w, given annotated training data, can be assessed as $WC(w)$, i.e., the number of distinct codes associated with w in the training data or Ethnologue database. The idea is that a uniquitous word like 'the' will be associated with many codes, while a fairly unique language name will be associated with only one or a few codes. For example, consider the following entry:

W. M. Rule 1977 *A Comparative Study of the Foe, Huli and Pole Languages of Papua New Guinea*, University of Sydney, Australia [Oceania Linguistic Monographs 20]

foe	pole	huli	papua	guinea	comparative	new	study	languages	and	a	the	of
1	2	3	57	106	110	145	176	418	1001	1101	1169	1482
1.0	2.0	1.5	19.0	1.86	1.04	1.32	1.21	2.38	2.39	1.10	1.06	1.27

Table 3. The values of $WC(w)$ for w taken from an example entry (mid row). The bottom row shows the *relative increase* of the sequence of values in the mid-row, i.e., each value divided by the previous value (with the first set to 1.0).

Table 3 shows the title words and their associated number of codes associated (sorted in ascending order).

So far so good, we now have an informativeness value for each word, but at which point (above which value?) do the scores mean that word is a near-unique language name rather than a relatively ubiquitous non-informative word? Luckily, we are assuming that there are only those two kinds of words, and that at least one near-unique language will appear. This means that if we cluster the values into two clusters, the two categories are likely to emerge nicely. The simplest kind of clustering of scalar values into two clusters is to sort the values and put the border where the relative increase is the highest. Typically, in titles where there is exactly one near-unique language name, the border will almost always isolate that name. In the example above, where we actually have three near-unique identifiers, this procedure correctly puts the border so that Foe, Pole and Huli are near-unique and the rest are non-informative.

Now, that we have a method to isolate the group of most informative words in a title e_t (denoted $SIG_{WC}(e_t)$), we can restrict lookup only to them. TWL is thus defined as follows:

$$TWL(e) = \cup_{w \in SIG_{WC}(e_t)} TD(w)$$

In the example above, $TWL(e_t)$ is $\{fli, kjy, foi, hui\}$ which is almost correct, containing only a spurious [fli] because Huli is also an alternative name for Fali in Cameroon, nowhere near Papua New Guinea. This is a complication that we will return to in the next section.

The resulting accuracies jump up to $PA_{TWL}(A) \approx 0.57$ and $SA_{TWL}(A) \approx 0.73$.

Given that we "know" which words in the title are the supposed near-unique language names, we can afford, i.e., not risk too much overgeneration, to allow for spelling variants. Define TWL_S ("with spelling variants") as:

$$TWL_S(e) = \cup_{w \in SIG_{WC}(e_t)} TD_S(w)$$

We get slight improvements in accuracy $PA_{TWL_S}(A) \approx 0.61$ and $SA_{TWL_S}(A) \approx 0.74$.

The $WC(w)$-counts make use of the annotated entries in the training data. An intriguing modification is to estimate $WC(w)$ without this annotation. It turns out that $WC(w)$ can be sharply estimated with $WI(w)$, i.e., the raw number of entries in the training set in which w occurs in the title. This identity breaks down to the extent that a word w occurs in many entries, all of them

pointing to one and the same language id. From domain knowledge, we know that this is unlikely if w is a near-unique language name, because most languages do not have many descriptive works about them. The TWL-classifier is now unsupervised in the sense that it does not have to have annotated training entries, but it still needs raw entries which have a realistic distribution. (The test set, or the set of entries to be annotated, can of course itself serve as such a set.)

Modeling Term Weight Lookup with WI in place of WC, call it TWI, yields slight accuracy drops $PA_{TWI}(A) \approx 0.55$ and $SA_{TWI}(A) \approx 0.70$, and with spelling variants $PA_{TWI_S}(A) \approx 0.59$ and $SA_{TWI_S}(A) \approx 0.71$. Since, we do in fact have access to annotated data, we will use the supervised classifier in the future, but it is important to know that the unsupervised variant is nearly as strong.

4 Term Weight Lookup with Group Disambiguation

Again, from our domain knowledge, we know that a large number of entries contain a "group name", i.e., the name of a country, region of genealogical (sub-)group in addition to a near-unique language name. Since group names will naturally tend to be associated with many codes, they will be sorted into the non-informative camp with the TWL-method, and thus ignored. This is unfortunate, because such group names can serve to disambiguate inherent small ambivalences among near-unique language names, as in the case of Huli above. Group names are not like language names. They are much fewer, they are typically longer (often multi-word), and they exhibit less spelling variation.

Fortunately, the Ethnologue database also contains information on language classification and the country (or countries) where each language is spoken. Therefore, it was a simple task to build a database of group names with genealogical groups and sub-groups as well as countries. All group names are unique[9] as group names (but some group names of small genetic groups are the same as that of a prominent language in that group). In total, this database contained 3 202 groups. This database is relatively complete for English names of (sub-)families and countries, but should be enlarged with the corresponding names in other languages.

We can add group-based disambiguation to TWL as follows. The non-significant words of a title is searched for matching group names. The set of languages denoted by a group name is denoted $L(g)$ with $L(g) = C$ if g is not a group name found in the database.

$$\begin{aligned} TWG(e) \quad = \quad & (\cup_{w \in SIG_{WC}(e_\iota)} LN(w)) \\ & \cap_{g \in (Words(e_\iota) \setminus SIG_{WC}(e_\iota))} L(g) \end{aligned}$$

[9] In a few cases they were forced unique, e.g., when two families X, Y were listed as having subgroups called Eastern (or the like), the corresponding group names were forced to Eastern-X and Eastern-Y respectively.

	PA	SA
NUL	0.15	0.21
TWL	0.57	0.73
TWL_S	0.61	0.74
TWI	0.55	0.70
TWI_S	0.59	0.71
TWG	0.59	0.74
TWG_S	0.64	0.77

Table 4. Summary of methods and corresponding accuracy scores.

We get slight improvements in accuracy $PA_{TWG}(A) \approx 0.59$ and $SA_{TWG}(A) \approx 0.74$. The corresponding accuracies with spelling variation enabled are $PA_{TWG}(A) \approx 0.64$ and $SA_{TWG}(A) \approx 0.77$.

5 Discussion

A summary of accuracy scores are given in Table 4.

All scores conform to expected intuitions and motivations. The key step beyond naive lookup is the usage of term weighting (and the fact the we were able to do this without a threshold or the like).

In the future, it appears fruitful to look more closely at automatic extraction of groups from annotated data. Initial experiments along this line were unsucessful, because data with evidence for groups is sparse. It also seems worthwhile to take multiword language names seriously (which is more implementational than conceptual work). Given that near-unique language names and group names can be reliably identified, it is easy to generate frames for typical titles of publications with language description data, in many languages. Such frames can be combed over large amounts of raw data to speed up the collection of further relevant references, in the typical manner of contemporary Information Extraction.

6 Related Work

As far as we are aware, the same problem or an isomorphic problem has not previously been discussed in the literature. It seems likely that isomorphic problems exist, perhaps in Information Extraction in the bioinformatics and/or medical domains, but so far we have not found such work.

The problem of language identification, i.e., identify the language of a (written) document given a set of candidate languages and training data for them, is a very different problem – requiring very different techniques (see Hammarström (2007a) for a survey and references).

We have made important use of ideas from Information Retrieval and Data Clustering.

7 Conclusion

We have presented (what is believed to be) the first algorithms for the specific problem of annotating language references with their target language(s). The methods used are tailored closely to the domain and our knowledge of it, but it is likely that there are isomorphic domains with the same problem(s). We have made a proper evaluation and the accuracy achieved is definetely useful.

Acknowledgements

We wish to thank the responsible entities for posting the Ethnologue, WALS, and the MPI/EVA library catalogue online. Without these resources, this study would have been impossible.

References

Baeza-Yates, R. and Ribeiro-Neto, B. (1997). *Modern Information Retrieval.* Addison-Wesley.

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

Hammarström, H. (2005). Review of the Ethnologue, 15th ed., Raymond G. Gordon, Jr. (ed.), SIL international, Dallas, 2005. *LINGUIST LIST*, 16(2637).

Hammarström, H. (2007a). A fine-grained model for language identification. In *Proceedings of iNEWS-07 Workshop at SIGIR 2007, 23-27 July 2007, Amsterdam*, pages 14–20. ACM.

Hammarström, H. (2007b). *Handbook of Descriptive Language Knowledge: A Full-Scale Reference Guide for Typologists*, volume 22 of *LINCOM Handbooks in Linguistics.* München: Lincom.

Hammarström, H. (2008). On the ethnologue and the number of languages in the world. Submitted Manuscript.

Part Two: Languages of the World

Chapter V | Counting Languages in Dialect Continua Using the Criterion of Mutual Intelligibility

Corrected version of
Hammarström, H. (2008). Counting languages in dialect continua using the criterion of mutual intelligibility. *Journal of Quantitative Linguistics*, 15(1):34–45.

Counting Languages in Dialect Continua Using the Criterion of Mutual Intelligibility

Harald Hammarström
Department of Computer Science and Engineering
Chalmers University of Technology
and University of Gothenburg
SE-412 96 Göteborg, Sweden
harald2@chalmers.se

Abstract

This paper shows how it is possible to count languages vs. dialects if, for every pair of varieties, we are given whether they are mutually intelligible or not. The method is to divide the varieties into a *minimum number* of *internally mutually intelligible* groups where each group counts as one language. Expressed in terms of graphs (as in discrete mathematics), the method is even easier understood as: applying graph-colouring to a graph over varieties with the intelligibility interrelationships as edges. Graph colouring is already mathematically well-understood and we can easily prove properties intuitively associated with the concepts language and dialect, and remove any fears that these concepts should lead to inconsistencies. The presentation requires only a minimal acquaintance with sets, combinatorics and graphs.

1 Introduction

In trying to answer the question "how many languages are there in the world?", linguists have had a hard time coming up with a satisfactory answer. Even when explicitly disregarding non-linguistic criteria (such as ethno-socio-economico-politico-cultural ones), they say that defining languages by the mutual intelligibility criterion (MI) is not possible (e.g. Anderson 2005).

Firstly, mutual intelligibility is not a strict yes/no distinction but a matter of degree. Subsumed hereunder are also cases where the degree of intelligibility is not enough to enable communication immediately, but high enough to enable communication after, say, only a few days of exposure, such as among the Mekeo languages (Jones 1998:19). Also, there may exist cases where intelligibility is not symmetric, i.e., A understands B but not vice versa, although I have yet to

see a genuine well-documented example[1].

Secondly, even if it were simplified into being yes/no and symmetric, counting languages by the MI, would lead to contradictions in dialect-chain situations. E.g. if A is MI with B, B is MI with C but A is not MI with C – a completely realistic situation – then setting A and B to be the same language and B and C as the same language is contradictory because A and C are not the same language by the MI criterion.

In this paper we will show that the second objection is premature. There is a perfectly consistent way to count languages using a symmetric strict yes/no mutual intelligibility criterion that preserves intuitive properties about languages and numbers of languages. Linguists seem to have so far failed to appreciate this[2], as the following selection of quotations exemplify:

> Such situation are referred to by linguists as 'dialect chains', and they result in sometimes arbitrary decisions being made as to how many languages are involved. (Lynch and Crowley 2001:2)

> The criterion that "technically ... mutually intelligible forms of speech are known as **dialects**, and [that] the term **language** is used for mutually unintelligible forms of speech" (Lehmann 1973:33), does not apply satisfactorily to such situations as the Chaga continuum from Siha to Usseri. (Polomé 1980:3)

> A common situation is a string of similar varieties, in which the speakers of variety A understand those of C, and so on, but the speakers of A do not understand the variety at the other end of the continuum, or even those part way along. Even if we can define 'understand', where is the divide between language and dialect in this situation? (Heine and Nurse 2000:2)

> In some cases, the intelligibility criterion actually leads to contradictory results, namely when we have a dialect chain, i.e., a string of dialects such that the adjacent dialects are readily mutually intelligible, but dialects from the far ends of the chain are not mutually intelligible. A good illustration of this is the Dutch-German dialect complex. One could start from the far south of the German-speaking area and move to the far west of the Dutch-speaking area without encountering any sharp boundary across which mutual intelligibility is broken; but the two end points of this chain are speech varieties

[1] Of course, it is easy to think of examples where A and B are closely related and speakers of A tend to understand B but not the other way round. For the sake of an example. take Jamaican Creole and Oxford English. But in most (all?) such cases this is because the A speakers have been exposed to B a lot more, and not purely because of their knowledge of A. I see no reason to differentiate this situation from that where A are B aren't closely related, and speakers of A know B as well, but not vice versa.

[2] See however Hockett (1958:321-330) for an embryo to the approach taken in this paper (whose views recollected by e.g., Heine and Köhler 1981:1-3). Note also that the matter is not discussed in the most recent encyclopaedia entry on dialect chains (Heap 2006).

> so different from one another that there is no mutual intelligibility possible. If one takes a simplified dialect chain A - B - C, where A and B are mutually intelligible, as are B and C, but A and C are mutually unintelligible, then one arrives at the contradictory result that A and B are dialects of the same language, B and C are dialects of the same language, but A and C are different languages. There is in fact no way of resolving this contradiction if we maintain the traditional strict difference between language and dialects, and what such examples show is that this is not an all-or-nothing distinction, but rather a continuum. In this sense, it is impossible to answer the question how many languages are spoken in the world. (Comrie 1987:3)

On the latter quote, a few clarifying remarks are in order: Comrie is discussing the definition that runs "two varieties are the same language **if and only if** they are mutually intelligible"[3]. I am not denying that this definition leads to a contradiction – I am saying that other intuitively acceptable definitions based (solely) on the MI have been ignored. In particular, I will present a definition whereby it is still true that "if two varieties are the same language then they are mutually intelligible" but the converse "if two varieties are mutually intelligible then they are of the same language" does not have to be true. The spirit of the latter is instead rendered by a slightly more relaxed requirement that says that the number of languages should not be unnecessarily multiplied. The definition is given full formal treatment below.

2 Counting Languages

The task is to decide, for a finite set X of speech varieties, how many languages there are using only a binary symmetric strict yes/no relation of mutual intelligibility (henceforth MI). For ease of presentation we shall model the situation as there being n speakers each speaking exactly one variety. It will be seen that the method is really indifferent to the distribution of varieties over people, names or any other grouping, so there is no loss of generality. Speakers will be denoted by capital letters, e.g., A, B, C. Thus let $X = \{A, B, C, \ldots\}$ be a any finite set of speakers.

2.1 Definition

Definition 1 *The number of languages in X is the least k such that one can partition X into k blocks such that all members within a block understand each other.*

[3] Comrie affirms in a personal email (9 Sept 2005) that the quoted paragraph concerns only this particular definition, and the statements therein that may look as if they quantify also over other intuitive definitions based on the MI, should not be so interpreted.

A partition of a set X into blocks is simply a division of the members of X into disjoint non-empty groups that exhaust X. So if say, $X = \{A, B, C\}$, we can partition it into:

1. One block: $\{A, B, C\}$
2. Two blocks: there are exactly three possibilities $\{A, B\}, \{C\}$ or $\{A\}, \{B, C\}$ or $\{A, C\}, \{B\}$.
3. Three blocks: $\{A\}, \{B\}, \{C\}$.

Clearly, the number of blocks in a partition ranges between 1 and the number of members of the set (also known as the *cardinality* of the set).

Now let's say $X = \{A, B, C\}$ depicts the classic dialect chain situation where A and B are MI, B and C are MI, but A and C are not MI. The partition into one block does not satisfy the requirement of the definition, since A and C, that are in the same block, do not understand each other. Of the three partitions into $k = 2$ blocks, two of them satisfy the definition: $\{A, B\}, \{C\}$ is ok because A and B are MI; $\{A\}, \{B, C\}$ is ok because B and C are MI. The partition into three blocks also trivially satisfies the condition that no pair within a block should be mutually unintelligible, but $k = 3$ is not minimal. Thus the number of languages in the example is 2, and we can immediately observe a curious feature of the definition: the number of languages k is unique, but there may be several satisfying partitions into k blocks.

2.2 Properties

It should be obvious that the definition is well-behaved in the sense it yields a unique number of languages k (for example, one way to arrive at the number is to just try out all partitions of the given set X). It should also be clear that a partition defines an assignment of speech varieties into languages such that A and B belong to the same language if and only if they belong to the same block. But what about properties of partitions that satisfy the minimal k and the requirement of inside-block intelligibility? Intuitively, if blocks are to be identified with languages, one would expect the following two properties to hold:

Property 1 *All those who speak the same language speak varieties which are mutually intelligible.*

Property 2 *There are no "superfluous" languages, i.e., for any division of varieties into languages satisfying property 1, a person speaking exactly one variety of each of the languages can communicate with everyone, whereas someone speaking less than k varieties cannot communicate with everyone.*

That the first property holds for the given definition is immediate from the definition. Informally, the second property holds because otherwise k would not be minimal, as required. A more detailed proof, which involves a little more work, is given below in section 3.3.

The definition and the properties emanating from it, is not specific to any particular type of language-variety landscape but extends to completely arbitrary constellations of language varieties and MI-interrelationships. It should actually be understood as an even more abstract counting method: given a set of objects and a symmetric, non-reflexive, non-transitive "is-different" relation over them (here: mutual unintelligibility), what is the minimal number of blocks one can partition this set into such that all members within a block are not different to each other?

For pedagogical reasons we shall now continue the presentation in terms of graphs.

3 Further Examples and Properties

Again, the task is to decide, for n speech varieties, how many languages there are using only a binary symmetric strict yes/no relation of mutual intelligibility.

3.1 Definition

Let the n speakers be vertices V of a graph[4] G. Let G have an edge between vertices $A, B \in V$ if and only if A and B do **not** speak mutually intelligible varieties.

Definition 2 *The number of languages is the smallest k such that one can colour the vertices of G with k colours such that no two vertices that share an edge have the same colour.*

This number is usually called the *chromatic number* of a graph G and is denoted $\chi(G)$ (Read 1968).

3.2 Examples

Again, an example of a graph illustrating the most basic dialect-chain situation is shown in figure 1. For G in figure 1 the chromatic number is 2. It is not possible to colour the vertices A, B and C with only one colour because then A and C would get the same colour – violating the condition that vertices which share an edge should not have the same colour. It would be possible to colour the vertices with (exactly) three colours, one each, without violating the shared-edge-different-colour condition, but 3 is not the chromatic number because it is also possible to colour G with less, namely 2, colours. There are in fact two different ways to colour G with 2 colours (say red and green): 1. $\{A, B\}$ red, $\{C\}$ green; 2. $\{A\}$ red, $\{B, C\}$ green.

Another example, a four-member dialect chain, is shown in figure 2. For G in figure 2 the chromatic number is also 2. There is only one 2-colouring: A, B

[4] For readers not familiar with graphs, a graph can be thought of as a set of points ("vertices") in a two-dimensional space and an arbitrary set of lines ("edges") between pairs of points. More information can be found in any introductory book on discrete mathematics.

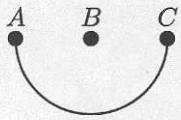


Figure 1. Graph for (A, B), (B, C) are MI but (A, C) are not MI.

green and C, D red. (Clearly, one colour is not sufficient. However, if one tries 2 colours: A must have some colour, then C and D must have a different colour. Nothing prevents C and D from having the same colour, so they get the same colour. Now, only B remains which cannot be coloured by the colour of C and D, but it can have A:s colour. All the choices were forced or colour-conservative so this is the only 2-colour possibility.)

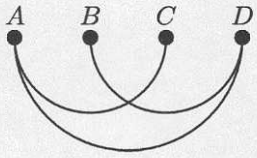


Figure 2. Graph for (A, B), (B, C), (C, D) are MI but no other pairs are MI.

Lastly, a third slightly more complicated example is shown in figure 3. The chromatic number of the graph in figure 3 is 3. There are no less than five different minimal colourings: 1. $\{A\}$ $\{B, D\}$ $\{C, E\}$; 2. $\{B\}$ $\{A, D\}$ $\{C, E\}$; 3. $\{A, D\}$ $\{B, E\}$ $\{C\}$; 4. $\{A, C\}$ $\{B, E\}$ $\{D\}$; 5. $\{A, C\}$ $\{B, D\}$ $\{E\}$. This is perhaps more easily seen if the graph is redrawn (not changed) as a pentagon, i.e by keeping A at the top, but putting D and C at the next level and B and E at the bottom level (as shown in figure 4).

3.3 Properties

As may have been experienced by the reader, it is not trivial to calculate what the chromatic number is, even for a graph of quite moderate size. It should be clear to everyone though, that it is always possible to reach the answer by tediously enumerating and checking all possibilities.

There is a mathematically well-understood systematic method to calculate the chromatic number (and the number of minimal-size colourings), in terms of

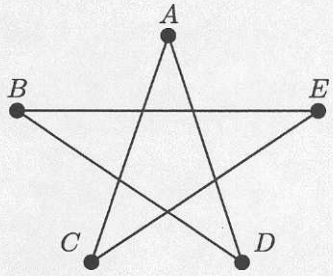


Figure 3. Graph for (A, B), (B, C), (C, D), (D, E), (E, A) are MI but no other pairs are MI.

breaking down any graph into more tractable pieces. The interested reader may consult the excellent introduction by Read (1968).

The bad news is that finding the chromatic number is an NP-complete problem (Garey and Johnson 1979). In layman terms, this implies that there is no known "smarter" method to find the chromatic number than to go through all the possible partitions of the vertices into blocks ("colours"). All known methods rely on the fact that it is "easy" to check whether a given colouring is ok, i.e., just to check if any edges violate the constraint, but still, in the worst case, need to step through essentially all possible partitions. This is the bottleneck because the number of possible partitions of n vertices in blocks is exponential[5] in n. For instance, if $n = 20$ there are 51724158235372 partitions to consider. Such is the problem in its full generality. But of course, for specific cases, there may be symmetries and regularities which makes the solution considerably more digestible.

As has been observed, there may be more than one minimal colouring. For any of one these minimal colourings, we can identify colours as languages. That is, let G again be the graph depicting the situation at hand and let $c_1, c_2, \ldots, c_k \neq \emptyset$ be a minimal colouring, thus satisfying $k = \chi(G)$ (the chromatic number of G), $\bigcup c_i = V$ and $c_i \bigcap c_j \neq \emptyset$ for $i \neq j$. Identifying languages as colours simply means that $c_1, \ldots, c_k$ form the k languages. Languages so defined have the following two crucial properties:

Property 1 *All those who speak the same language speak varieties which are mutually intelligible.*

Property 2 *There are no "superfluous" languages, i.e., for any division of varieties into languages satisfying property 1, a person speaking exactly one variety of each of the languages can communicate with everyone, whereas someone*

[5] In fact, it is $S(n, 1) + S(n, 2) + \ldots + S(n, n)$, where S denotes the Stirling numbers of the second kind. See e.g. (Stanley 1997:33) for more information.

speaking less than k varieties cannot communicate with everyone.

Proof of property 1: Assume there were a pair of non-MI varieties of the same language, i.e., assigned the same colour. Since they were not MI they would share an edge in the graph – contradicting that colouring was legal in the first place.

Proof of Property 2: Assume that one speaks $k' < k$ varieties $Z_1, Z_2, \ldots, Z_{k'}$. Form the k' groups one could communicate with using the respective variety: $c'_i = \{Y \in V | Y \text{ is MI with } Z_i\}$ for $1 \leq i \leq k'$. (If $c'_i \bigcap c'_j \neq \emptyset$ for some $i < j$ then remove the intersecting elements from (say) c'_i). If one could communicate with everyone using the Z_i varieties, then $\bigcup c'_i = V$. If $c'_1, c'_2, \ldots, c'_{k'}$ satisfies property 1, i.e., within each c'_i there are no edges between the members, then $c'_1, c'_2, \ldots, c'_{k'}$ would yield a legal colouring of G – contradicting the minimality of k.

Restating, any minimal colouring of a (graph of) a language/dialect situation has the above two properties. Any count of the number languages other than "as many as the blocks of a minimal colouring" would in some way fail to satisfy one of the two properties about the languages counted. And clearly, the two stated properties must be part of the intuitive understanding of what it means to be a language.

At this point, however, we still cannot give a finished count of the number languages of the world because:

- We still do not have an answer to *when* mutual intelligibility holds given two languages (cf. the first point in the introduction).

- Even if we did have a good (or arbitrary) method to decide when two varieties are mutually intelligible, we do not have complete knowledge of the speech varieties of the world. The best one-piece source on this matter is the Ethnologue (Gordon 2005) but it does not provide (nor does it aim to) systematic detailed information on (any kind of) intelligibility between varieties.

- As alluded to above, even if we did have complete knowledge etc, the resulting graph for the world would have on the order of 6900 vertices, which might be intractable. Since most language varieties, unquestionably, aren't intelligible to each other, this graph would turn out quite easy-handled, but it remains to see just how complex it does get.

Acknowledgements

I wish to thank Richard Sproat for pointing out a serious error in the formulation of property 2.

References

Anderson, S. R. (2005). How many languages are there in the world? Answer to a FAQ by the Linguistic Society of America, Washington, D.C. Published on the web as `http://www.lsadc.org/pdf_files/howmany.pdf`. Accessed 1 September 2005.

Comrie, B., editor (1987). *The World's Major Languages.* London: Croom Helm.

Garey, M. R. and Johnson, D. S. (1979). *Computers and Intractability: A Guide to the Theory of NP-Completeness.* W. H. Freeman, New York.

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

Heap, D. (2006). Dialect chains. In Brown, K., editor, *Encyclopedia of Language and Linguistics*, volume 3, pages 528–530. Amsterdam: Elsevier, 2 edition.

Heine, B. and Köhler, O. (1981). *Linguistik - Ostafrika (Kenya, Uganda, Tanzania): Gliederung der Sprachen und Dialekte.* Afrika-Kartenwerk E 10. Berlin: Gebrüder Borntraeger.

Heine, B. and Nurse, D. (2000). Introduction. In Heine, B. and Nurse, D., editors, *African Languages: An Introduction*, pages 1–10. Cambridge University Press.

Hockett, C. F. (1958). *A Course in Modern Linguistics.* Toronto: MacMillan.

Jones, A. A. (1998). *Towards a Lexicogrammar of Mekeo (An Austronesian Language of West Central Papua)*, volume 138 of *Pacific Linguistics: Series C.* Canberra: Research School of Pacific and Asian Studies, Australian National University.

Lynch, J. and Crowley, T. (2001). *Languages of Vanuatu: A New Survey and Bibliography*, volume 517 of *Pacific Linguistics.* Canberra: Research School of Pacific and Asian Studies, Australian National University.

Polomé, E. C. (1980). The languages of tanzania. In Polomé, E. C. and Hill, C. P., editors, *Language in Tanzania*, Ford Foundation Language Surveys, pages 1–25. Oxford University Press.

Read, R. C. (1968). An introduction to chromatic polynomials. *Journal of Combinatorial Theory*, 4:52–71.

Stanley, R. P. (1997). *Enumerative Combinatorics: Volume I*, volume 49 of *Cambridge Studies in Advanced Mathematics.* Cambridge University Press.

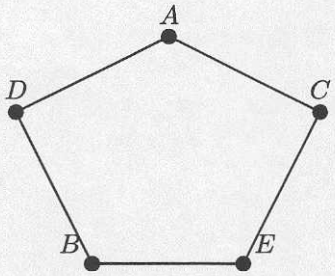


Figure 4. Graph for (A, B), (B, C), (C, D), (D, E), (E, A) are MI but no other pairs are MI.

Appendix: Failing Approaches to Defining Languages Uniquely

As seen in the previous section, the method for counting languages does not always uniquely determine *which* the languages are, it just says *how many* they are and gives a number of alternatives to which they are. It is tempting to look at ways to synthesize the alternatives into a unique language/dialect definition.

For example, it is readily seen in figure 1 that in the two minimal colourings, A and C are never in the same colour, whereas B is once in the company of A and once in C. It is then tempting to define the languages as "the maximum-size set L of vertices that are never in the same colour in any minimal colouring". The rest of the vertices can then be thought of as dialects of the varieties to which they are MI. (A dialect can then be the dialect of several (distinct) languages, and even be a dialect of a dialect). In the example of figure 1 this would beautifully synthesize the two minimal colourings to say that A and C are separate languages and B is a dialect of A as well as a dialect of C.

Unfortunately there are cases where this approach does not "work". In the graph of figure 2, L would not be unique; either of $L = \{A, C\}$, $\{A, D\}$, $\{B, C\}$, $\{B, D\}$ has the maximum-size 2. (Here, of course, since there is only one minimal colouring, we can satisfactorily take AB and CD as the two languages and call A a dialect of AB, B another dialect of AB, and so on).

But, more seriously, in the case of figure 4 (depicting the same situation as in figure 1), there are five different maximum-size $L = \{A, C\}$, $\{A, D\}$, $\{B, D\}$, $\{B, E\}$, $\{C, E\}$, so $|L| \neq k$ which invalidates the idea. The graph of figure 4 is the smallest graph where $|L| \neq k$. I don't believe there is a sensible way of uniquely defining languages in such graphs (that is, graphs which have odd-size circles but whose chromatic number is lower than the number of members of the circle). In the graph in question, we are told that there are exactly 3 languages but all the vertices are symmetric so there seems to be no way to

single out three of them or divide five nodes into three equal-size groups. If we wish to select 3 of the five to be languages and the other 2 dialects, there is no un-arbitrary way to decide which go as languages since all five vertices are structurally indistinguishable. If we wish to divide the 5 into three groups, they would not all be of equal size and, again, there is no basis for putting one or the other vertex in the bigger (or smaller) group.

This might not just be a purely theoretical problem. It is conceivable that such a "ring" could be the correct state of affairs somewhere in the world, say, if a dialect continuum settled around a mountain and the two extremes of the chain meet and influence each other (for a couple of centuries) so that they become intelligible dialects.

Chapter VI | Whence the Kanum base-6 numeral system?

Hammarström, H. (2009). Whence the Kanum base-6 numeral system? *Linguistic Typology*, 13(2):305–319.

Whence the Kanum base-6 numeral system?

Harald Hammarström
Department of Computer Science and Engineering
Chalmers University of Technology
and University of Gothenburg
SE-412 96 Göteborg, Sweden
harald2@chalmers.se

Abstract

Base-6-36 numeral systems, a typological rarity, are found in Kanum languages of New Guinea as testified by Donohue (2008). We look at the probable relatives of the Kanum languages and show that the base-6 system must have emerged in the Tonda group specifically. Since there is no evidence of body-part terms in the base-6 forms attested, we speculate that these systems have a different origin. Specifically, we suggest that the base-6 systems arose for counting yams. The ethnographic data for Kanum and other relevant languages are in concord with such a scenario. Whether there is a historical connection with base-6 systems of the Kolopom languages, near, but not adjacent, to the west, remains an open question. If there is a connection, it is areal rather than genetic, but sufficient evidence for a pre-historic areal connection remains to be amassed. Equally, if not more, puzzling would be the conclusion that there is no historical connection, given the rarity of base-6 in the world as a whole.

Keywords: Numerals, Numeral Systems, New Guinea, Papuan Languages, Areal Diffusion, Kolopom, Frederik-Hendrik-Eiland, Kanum, Nambu, Tonda, Yei, Trans New Guinea, Typological Rarities, Morehead and Upper Maro

Acknowledgements

I wish to thank Mark Donohue for comments and discussion (the usual disclaimers apply), Randy Lebold and Mark Donohue for access to unpublished data, Raoul Zamponi for help with access to Annual Reports for Papua New Guinea and three anonymous reviewers for helpful commentary.

1 Background on Kanum and Related Languages

This commentary attempts to trace the origins of the base-6 numeral system of Kanum, as far as available data permits.[1]

Conventionally, the Kanum languages are classified as belonging to the Morehead and Upper Maro family, consisting of Tonda (ca 10 languages, including Kanum, around the Bensbach [aka Torassi] and Morehead rivers), Nambu (about 6 languages, immediately east of the Tonda languages, east of the Morehead) and Yei (roughly two languages (Sohn 2006:20), immediately north of the Tonda area).

It is not yet certain whether Tonda, Nambu and Yei are genetically related, but there are interesting numeral form resemblances in Tonda-Nambu-Yei for 1 and 2 (see below), so if it is indeed a valid genetic family, these are probably cognates. As far as we can tell, no other languages in the region show suggestive form resemblances in the numeral domain.

2 Numerals in Kanum and other Relevant Languages

It is clear from early and late data that the Yei language(s) originally had a restricted numeral system, i.e., monomorphemic numerals for 1 and 2, above which ad hoc repetitions or inexact words would be used.[2] A selection of the published forms for 1 and 2 are shown in Table 1.

There are some 20 attestations of numerals from varieties of the Nambu group, most of which are collected and analyzed in Lean (1986:52-59). Almost all vocabularies agree on the roots for 1 and 2. These roots resemble the #-marked[3] forms shown in Table 1. In contrast, above 2, the vocabularies wholly diverge. Some vocabularies show a restricted system with 3=2+1 and 4=2+2, whereas others have a monomorphemic 3, 4=2+2 and evidence of base 5 above that. In other words, proto-Nambu must have had a restricted system whereas some modern Nambu varieties show base-5, or incipient base-5 systems.

The base-6 system of Kanum, along with the parallel simple and moderate systems, as recorded by Donohue (2008) are repoduced in Table 3. All other available data on Tonda group languages are collected in Table 2, but, as will

[1] The following relevant unpublished data was inaccessible to the author: Data on Trans-Fly languages collected by Wurm in 1966 and 1970, data on Trans-Fly languages collected by Capell for his survey of the South Pacific 1950s and on, data on Trans-Fly languages collected by Nicholas Evans in the 2000s, A. P. Lyons's journals held at National Cultural Council at Port Moresby, the archives of Francis Edgar Williams, held at the National Archives of Papua New Guinea. On the other hand, the archives of Sidney H. Ray, who published a lot of missionary and patrol mss vocabularies of Papuan languages, at SOAS Library in London (consulted August 2008) did not contain any further data on Trans-Fly languages than was published.

[2] An occasional variety attests an incipient base-5 system, i.e., 1,2,2+1,2+2,5,5+1,5+2,5+2+1,5+2+2 (Nevermann 1942).

[3] We use # in contrast to * to mark hypothetical forms that stem from phonologically inadequate transcriptions and lack the rigour normally associated with the *-symbol.

Table 1. Numerals in Yei and Nambu. Komadeau, Doŋgeab, Kwěl are from Nevermann (1942) while Poo is cited from Sohn (2000) and Jéi [from the Upper Maro] is from Drabbe (1954). The Yei varieties are separated into two languages as suggested by (Sohn 2006:20). The Nambu resemblant forms are based on the data in Lean (1986:52-59). The three-letter codes in brackets are the iso-639-3 codes for the languages in question.

	Tanas-Bupul [jei]	Erambu-Poo [jei]				
	Komadeau	Doŋgeab	Kwěl	Poo	Jéi	Nambu
1	*nâ̌mpŭ̌ē*	*nâ̌mpě*	*nâ̌mpa͡i*	*na:mpaj*	*nāmpèi*	*#ambiro*
2	*jɪ̄tấpē*	*hētấpē*	*ɪ̄tấpa͡i*	*jɛ̀ta:paj*	*jètapaé*	*#tumbi*

be seen, many of the vocabularies do not include numerals beyond 6. The first publication to unambiguously attest a full base-6 system is an anthropological work, frequently overlooked by linguists. Williams (1936:225-227) comprehensively describes the use of a base-6 system used for counting taitu (a smaller variety of yams) with the groups he labels Keraki. The words for 1, 2, 36, 216 and 1296, all monomorphemic, can be found in the text. Now, the peculiarity is that the Keraki speak a Nambu group language and would normally use a base-5 counting system, only borrowing the base-6 system in question from a Tonda group for the said counting activity. Williams himself testifies that for most of the Nambu groups (including Keraki) "The numerals run up to 5, and there is no elaborate system of counting except in so far as it has been learned from the west [i.e., Tonda languages – HH]" (Williams 1936:38), that for most of the Tonda varieties "numerals run up to six" (Williams 1936:35), and finally that "The custom of counting the taitu accurately is apparently a more or less recent fashion among the Keraki, having been imported from beyond the Morehead. It is still necessary, indeed, to have the assistance of experts before the count can be attempted in a Keraki village, for the enumeration is that of the Gambadi and Semariji [= dialects of Kunja, in modern language listings – HH] languages, viz. a 1-6 system, not that of the Keraki group, which employs a 1-5 system." There can be no question that Tonda group languages are what Williams meant by "from the west" and that the final quote witnesses this statement, because Tonda group languages are the only languages to the west considered by Williams.[4]

We may now turn to the analysis of the above Tonda, Nambu and Yei data.

The base-6 system is clearly at home in the Tonda group as it is unattested in Yei and Nambu, apart from the one case in Nambu where it is explicitly recognized as a Tonda group borrowing. The forms for 1-2 in Tonda-Nambu-Yei show some promising resemblance, but, as the transciption is dubious for most of the data, it is hardly meaningful to elaborate on these potential correspondences. No form correspondence between any Nambu and Tonda numeral above

[4] Save for possible Marind incursions, which, in any case, could not be the source for elaborate counting, since elaborate counting is not attested in Marind (Wirz 1922).

Table 2. Numerals in Tonda group languages. Sources and non-trivial identifications into the language division of Gordon (2005) are as follows. "Bangu" (MacGregor 1897) is a dialect of Kunja (Ayres 1983:11). "Keraki" (Williams 1936:225-227) borrowed the listed numerals from Kunja. "Mani" is not from native speakers but from Ngkâlmpu Kanum speakers who knew Mani (Nevermann 1939:58), and while the Mani village was abandoned in the 50s (Hitchcock 2004:388), it would have been a dialect of Kunja in the time of Nevermann. Tokwasa (Lyons 1914) is a dialect of Kunja (Ayres 1983:141). "Peremka" (Riley and Ray 1931:849-850) is a dialect of Wára. Wandatokwe (Lyons 1914) is a dialect of Wára (Ayres 1983:139). Ngkâlmpu 1 is from Drabbe (1954:37). Ngkâlmpu 2 is from Nevermann (1939). "N'gowugar" (Nevermann 1939) is a dialect of Blafe. "Tonda" (Lean 1986:48-49) is a dialect of Blafe. Sota Kanum is from Donohue (no date). The three-letter codes in brackets are the iso-639-3 codes for the languages in question.

	Kunja [pop]				Wára [tci]		Kanum, Ngkâlmpw [kcd]		Blafe [bfh]		Sota [krz]
	"Bangu"	"Keraki"	Mani	Tokwasa	"Peremka"	Wandatokwe	Ngkâlmpu 1	Ngkâlmpu 2	N'gowugar	"Tonda"	Sota
1	*nambi/ gnambi*	*nyambi*	*námbe*	*nambi*	*ncambi*	*niambi*	*nyambi/namper*	*náməto*	*námbi*	*nabi*	*æmpi*
2	*yethombi/ kethembi*	*yenta*	*jéwembi*	*enda*	*yendar*	*jenda*	*yenta/jempokà*	*jemuká*	*jenbádu*	*yalmbe*	*yənampɛ*
3	*yetho*		*djéro*	*e-tho*		*i-edo*	*juaw*	*juáw*	*jenbádɔ-námbido*	*yala*	*yɛlɯ*
4	*asár*		*ásar*	*asa*		*asa*	*esèr*	*aetsár*	*jíno*	*hasar*	*æsɛr*
5	*tambothoi/ tambothui*		*tâûbero*	*tabruja*	*ncambi*	*tombodoi*	*tàmpùi*	*tžoábəlu*	*n'gor*	*tambui*	*pɒplu*
6	*nimbo/nimb*		*tâûbero-námbe*				*tarawò*			*trawa*	*mæwɛ*
7							*pesmèri empi*				
8							*pesmèri jalmpö*				
9							*pesmèri jelà*				
10							*pesmèri esèr*				
11							*pesmèri tàmpùi*				
12							*pesmeri tarawò/ jempokà tarawò*			*yalmbe trawa*	
18							*juaw tarawò*			*yala trawa*	
24							*esèr tarawò*				
30							*tàmpùi tarawò*				
36		*peta*					*peta/nimpè*				
72							*jalmpö nimpè/ jempokà nimpè*				
216		*tarumba*					*tarumba*				
1296		*dameno*					*dameno*				

Table 3. The base-6 system of Kanum, along with the parallel simple and moderate systems Donohue (2008). The forms are from Yanggandur which belong to Ngkâlmpw Kanum [kcd] in the division of Gordon (2005).

	Kanum, Ngkâlmpw [kcd] Yanggandur		
	simple	moderate	complex
1	*naempr*	*aempy*	*aempy*
2	*yempoka*	*ynaoaempy*	*ynaoaempy*
3	*ywaw*	*ylla*	*ylla*
4	*eser*	*eser*	*eser*
5	*swabra*	*tampwy*	*tamp*
6	*'swy*	*traowao*	*ptae*
7		*psymery aempy*	*aempy ptae*
8		*psymery ynaoaempy*	*ynaoaempy ptae*
9		*psymery ylla*	*ylla ptae*
10		*psymery eser*	*eser ptae*
11		*psymery tampwy*	*tamp ptae*
12		*psymery traowao* or *yempoka traowao*	*tarwmpao*
13			*aempy tarwmpao*
14			*ynaoaempy tarwmpao*
15			*ylla tarwmpao*
16			*eser tarwmpao*
17			*tamp tarwmpao*
18			*ntamnao*
19			*aempy ntamnao*
20			*ynaoaemy ntamnao*
24			*wramaekr*
25			*aempy wramaekr*
30			*ptae wramaekr*
31			*aempy ptae wramaekr*
36			*(ntaop) ptae*
37			*aempy (ntaop) ptae*
50			*ynaoaempy tarwmpao (ntaop) ptae*
100			*eser wramaekr ptae ynaoaempy*
216			*tarwmpao*
1296			*(ntaop) ntamnao*
7776			*(ntaop) wramaekr*

2 has been noted. Consequently, if Tonda-Nambu-Yei are related, the simplest hypothesis is that proto-Tonda-Nambu-Yei had a restricted system, and that the Tonda base-6 system post-dates the breakup of the hypothesized family. The alternative hypothesis, that a Tonda-Tambu-Yei proto-language had a base-6 system, is less favourable because is requires two traceless evaporations of the base-6 system, once in Tonda and once in Yei.

To pinpoint the origin of the base-6 system within the Tonda group is not easy with the data at hand and the fact that, in Kanum as described by Donohue (2008) – the case where we do have reliable data – several systems are at play at the same time (cf. Table 3). If we posit the base-6 system to proto-Tonda all vocabularies are reasonably well-explained, except Mani of Nevermann which then could be deemed erroneous (recall that it does not come from native speakers). The divergence in forms for 5 and 6 could be accounted for by contrasts between simple/moderate systems and the forms for 3 as divergences after the break-up of proto-Tonda. On the other hand, the possibility that the base-6 was borrowed between Tonda varieties after the break-up of proto-Tonda cannot be ruled out; this hypothesis has the advantage of explaining Mani of Nevermann as a dialect that simply did not borrow it, and would also readily explain why the higher numerals 36, 216 and 1296 turn up in identical forms.

Be the phylogeny of the base-6 system within Tonda as it may, we must ask what provoked the appearance of this system, for it is a major typological rarity. In a survey that covers languages from *every* language family in the world (including isolates), and in each family covers most of the languages of the family, we find general purpose base-6-36 systems on only two places in the world; in Tonda as above, and on Kolopom Island (see below) (Hammarström pear).[5]

Usually, restricted numeral systems, as they develop normed expressions for higher exact quantities, take the path via hands and feet to make 5-10-20 systems. This is usually apparent in the etymologies for the forms in questions. What happened in Tonda is different – the base-6 system does not connect with fingers, hands, feets or any other body part counting. If the Tonda base-6 system does not come from counting on the body, where does it come from? As we shall see, there is an intriguing connection with the counting of yams!

The hypothesis can be formulated as follows. As a culture switches from hunting and gathering to a more tuber-cultivating subsistence mode, which requires storing and planting, there is more incentive for exact counting, thus more pressure for a speech community to develop normed expressions for higher exact quantities than 2 or 3 (= number of objects whose number one immediately recognizes, without grouping or counting). Almost always, tallying on the hands (fingers) and feet (toes) bootstraps the emergence of such normed expressions, yielding a 5-10-20 structure. If, as in the case of Tonda, the human body is not the source of these expressions, we should get some other structure than the 5-10-20 with hand-feet-man etymologies.

[5] Leont'ev (1974:68-69)'s claim that also Kati is base-6 is erroneous, as Kati, like other Ok languages, has a body-tally system (Galis 1955).

This hypothesis explains a number of facts. The first part, which says that tuber-agricultural languages should have non-restricted systems[6], explains why the base-6 arose in Tonda and was borrowed into Nambu where yams are staple food (Ayres 1983, Williams 1936, Hitchcock 2004), rather than in Yei, who are essentially hunter-gatherers (van Baal 1982). This is further strengthened by the fact that both Williams (1936:225-227) and Lean (1986:48) independently (Lean was unaware of Williams) adduce that the base-6 was intrinsically connected with counting yams[7]. The second part, which says that that if the system is not 5-10-20 then the etymologies of the forms should not involve body parts[8] explains why we fail to find 'hand' etc. etymologies in the base-6 systems in Tonda. Now, of course, there is nothing we know about yams that predicts 6 – it could have been 4, 7 or some other number – the point is that it is in sharp contrast with the otherwise ubiquitous 5-10-20 systems.

To make the epistemology clear, let us summarize the kind of evidence – so far absent – that would refute the above hypothesis about the emergence of the Tonda base-6 systems.

- Etymologies of crucial Tonda forms involving body-parts (or anything disconnected from tuber cultivation).
- Evidence of tuber cultivation, either in the past or present, in nearby varieties which do not show base-6 numeral systems.
- Presence of base-6 systems, either in the past or present, in nearby varieties without a subsistence type with similar importance of tuber cultivation.
- Very many examples around the world of tuber cultivating societies with restricted numeral systems.
- Ethnographic data about counting and specific cultural objects that would render the comments by Lean and Williams without discriminatory power. For example, if counting some other objects than yam was prior to, or more common, in Tonda and surrounding varieties, this would weaken the causal interpretation of the yams connection. Likewise, if ethnographic comments connecting the actual numeral system and some specific thing counted can be found in the non-base-6 languages "everywhere" then this evidence does not "select" the Tonda languages.

There is one more point to be made as to the Tonda base-6 numeral systems, namely the base-6 system(s) attested in the Kolopom languages. The Kolopom

[6] This claim, on a worldwide scale, has good explanatory power (though not exceptionless) but is beyond the scope of this commentary (Hammarström 2008).

[7] Lean (1986:48) discloses that it is "uncertain whether the system shown .. is the standard one used for everyday purposes or a special one used for counting yams which are grouped in sextets for this purpose".

[8] The prediction is not entirely vacuous; e.g., Meek (1931) attests hand gesture counting which uses hands and eyes – a closed fist covering one eye making 6.

languages are Kimaama, Riantana and Ndom, occupying most of Kolopom island (formerly Frederik-Hendrik-Eiland) in southeast Indonesian Papua. Kimaama and Riantana are almost certainly genetically related, since the 1st and 2nd person pronouns correspond in singular as well as plural, and have lexicostatistical agreement in the 20-40% range (Drabbe 1949, Voorhoeve 1975, Menanti and Susanto 2001). Ndom is likely to be genetically related to Kimaama-Riantana as well, but there is a little room for doubt. Ndom also shares the 1st person singular and plural pronoun form but lexicostatistical figures between Ndom and Kimaama/Riantana villages may drop below 10% (Menanti and Susanto 2001). Table 4 reproduces all available Kolopom numeral data.[9]

Considering that the data by Drabbe is the most reliable, we can interpret the vocabularies as follows. Drabbe's data attests base-6 for Kimaama below 20, and this is corroborated by "Teri-Kalwasch" of Geurtjens (up to 10) and by "Täri-Kalwa" of Nevermann, though Nevermann presumably made an error at 13: Since 13 (cf. 13 and 19 of Riantana) can be either *nĩ* or *nĩ növere* (Drabbe 1949:8), Nevermann must have got the impression that the former was 13 and the latter 14, resulting in a wordlist with a 6-13(!) system. This is understandable given the difficult circumstances under which he collected this data (Nevermann 1935b:56-59). At 20 (19?), Kimaama of Drabbe turns into base-20, and in "Klader", "Kimaam" of Nevermann as well as "Kaladdarsch" of Geurtjens, the systems lapse into base-5 already at 7. Note that both the *#ibuda* and *#turwa* roots adapt the meaning 5 rather than 6 in these cases! Clearly, the base-6 system in the Kimaama area was a competition with a base-5 system of the commonplace type – as expected *ketsja nda kawé/kitjanta kuwe* literally means "two hands" (Kluge 1938:148), and *tjĩ* means "man" (Drabbe 1949:8). Not unexpectedly, Donohue who collected Kimaama data more recently, found only a base-5 system (p.c. July 2008). Riantana is base 6 up to 24 which seems to have been the limit of counting ("en verder schijnt men daar niet te tellen") (Drabbe 1949:8). Also, *tarö* is the Riantana word (both) for 'many' and 'all' (Drabbe 1949:24). Finally, Drabbe's Ndom informant gave a consistent 6-36 system up to 180, with multiplications, a feature which is lacking in the Kimaam and Riantana data.

As for resemblances in form, it will be seen that the Kimaama and Riantana 3-6 seem to correspond, and not impossibly 1-2 as well. However, the Teri-Kalwa Kimaama of Geurtjens constructs 6-10 with the #me-morpheme that Drabbe ascribes to Riantana rather than Kimaama. Thus we are faced with borrowing, parallel systems or both – in a way we cannot hope to unravel at this point. The bottom line is that we now have a total of three base-6 systems with independent sets of forms, Kimaama-Riantana, Ndom and Tonda.

We have found no suggestive etymologies for any of the morphemes involved (except 10 and 20 of some Kimaam varieties, as above). Kluge (1938:148) did propose a link between Kimaam *durua* 'upper arm'[10] and *turua* '6', but there are good reasons for considering it a minimal pair rather than an etymological

[9] SIL Indonesia surveyed Kimaam district in August 2001 but the vocabularies from this survey had not been typed up as of writing this (p.c. Randy Lebold Feb 2008).

[10] In fact, unknown to Kluge, the word *durua* also means 'leg' (Drabbe 1949:15).

link[11]. Firstly, Drabbe fails to mention any such connection, which one suspects he would have caught if it was real. Secondly, the three independent attestations distinguish the 'upper arm/leg'-word with *d* from the '6'-word with *t* as *durua* ~ *turua* (Nevermann), *durò* ~ *turò* (Drabbe) and *doerwa* ~ *toeroea* (Geurtjens) showing that we are not dealing with allophonic variation for initial d/t. Also, two likely Austronesian loans with an unvoiced initial dental stop, *tùano* 'owner' and *tamoekoe* 'tobacco', are rendered as such, i.e., with t rather than d (Drabbe 1949).

The three systems have independent morphemes but are similar in structure. The fact that this structure, base-6, is extremely rare in the world, merits an investigation of a possible historical connection between the three. It is extremely unlikely that a genetic connection is the reason for the shared base-6 systems, since we cannot find likely lexical cognates bridging Tonda to any Kolopom language, and, as we have seen, the base-6 system does not go back to the most likely relatives of Tonda. Ndom is adjacent to Kimaama-Riantana (and only to Kimaama-Riantana) so a historical connection here is almost certain, in spite of the differing morphemes.

However, between Kolopom and Tonda we find intervening Komolom, Yelmek-Maklew, Morori and Marind languages which show no traces of base-6. The Komolom languages have base-5. The Yelmek-Maklew and Marind languages have restricted systems (Galis 1955), as did Morori originally (Nevermann 1939:69). It is relevant here to note that, as expected, the Kolopom (Serpenti 1965)[12] and Komolom (Nevermann 1935a) are tuber-agriculturalists while Yelmek-Maklew (Walker and Mansoben 1990, Aubaile-Sallenave and Bahuchet 1994), Morori (Nevermann 1939:37) and Marind speakers rely more on sago, hunting and gathering (van Baal 1966, Wirz 1922).

Figure 1 has a map of all the languages involved in the discussion.

If there is a historical connection between the Kolopom and Tonda languages it should thus be along the following scenario. There was once a stretch of tuber-cultivators connecting the Kolopom and Tonda languages geographically; this stretch was broken by invading Marind family speakers who rather depend on sago, hunting and gathering. While this scenario has in fact been painted before, based on other similarities than base-6 (Nevermann 1939:6), it remains very speculative. Should one wish to entertain this speculation further, there are, in fact, two potentially cognate forms in key terms in the base-6 systems, namely #*turua* (Kimaama-Riantana) ~ #traowao (Tonda) for '6' and #*teroamä* (Kimaama-Riantana) ~ #tarwmpao (Tonda) for '12'.

The opposite hypothesis, that the only two veritable base-6 systems appeared in a tiny area in South West New Guinea by chance, is no stronger than accounting for phonemic click languages in Eastern Africa and Southern Africa by independent innovation!

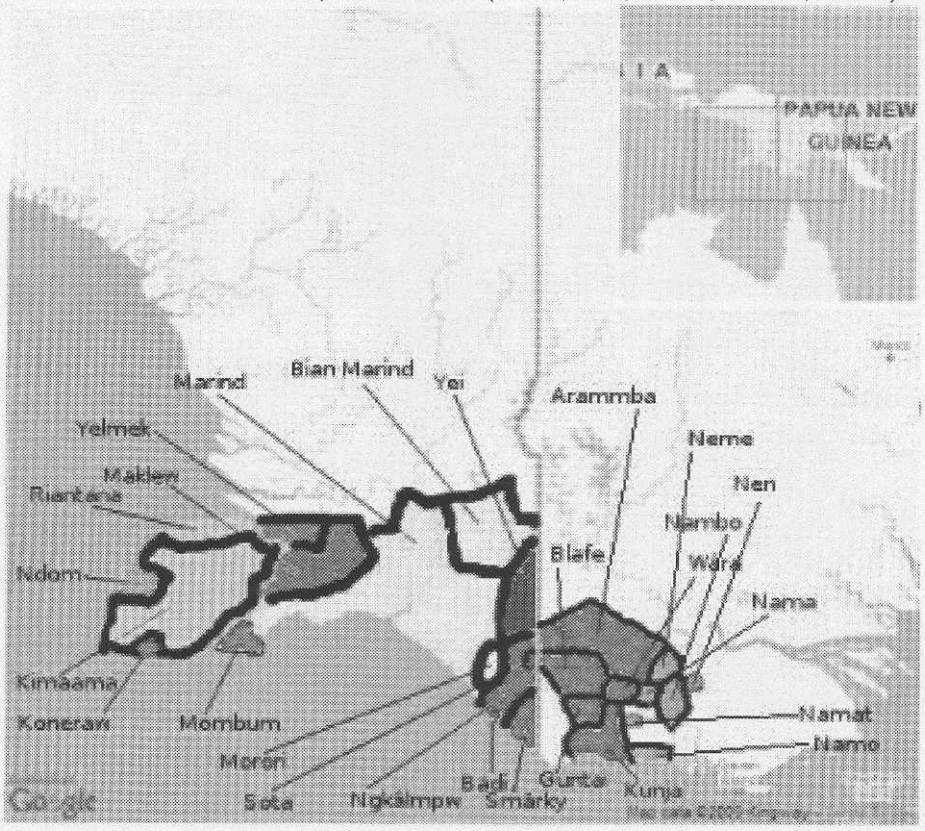
[11] Nevermind that an etymology for '6' as 'upper arm' or 'leg' would be unparallelled anywhere else in the world, since nearly the same is true for base-6 systems.

[12] I searched hard for any evidence of a similar yam counting connection for the base-6 systems at Kolopom in missionary manuscripts at KITLV archives, but found nothing, either pro or contra.

Table 4. Published data on numerals in the Kolopom languages. Sources and non-trivial identifications into the language division of Gordon (2005) are as follows. Ndom, Riantana, and Kimaghama are from Drabbe (1949:6-8). "Kaladdarsch" and "Teri-Kalwasch" (Geurtjens 1933) are Kimaama villages (Gordon 2005). "Klader", "Kimaam" and "Täri-Kalwa" in Kluge (1938:148), originally from manuscript vocabularies taken up by Nevermann in 1933-1934, are all all village names within the boundaries of Kimaama. The three-letter codes in brackets are the iso-639-3 codes for the languages in question.

	Ndom [nqm]	Riantana [ran]	Kimaama [kig]					
	Ndom	Riantana	Kimaghama	Kaladdarsch	Teri-Kalwasch	Klader	Kimaam	Täri-Kalwa
1	*sas*	*mebö*	*növere*	*nuwwéda*	*dubbadnè*	*nüwoda*	*mübärä*	*newere*
2	*thef*	*enàva*	*kàvé*	*kawé*	*kabi*	*kawä*	*kawe*	*gawir*
3	*ithin*	*pendö*	*péndji*	*pendsji*	*péndji*	*pändji*	*pinlji*	*pendji*
4	*thonith*	*wendö*	*jàndo*	*jándo*	*jándoe*	*jandoa*	*jaemdo*	*jamdo*
5	*merègh*	*mata*	*màdo*	*madoa*	*mando*	*madoa*	*mado*	*mado*
6	*mer*	*törwa*	*turò*	*toeroea*	*tulwa*	*turua*	*turo*	*terva*
7	*(mer) abo sas*	*mebö be*	*iburo növere*	*turrewahabé*	*dubbad-nènè*	*twuruaghawä*	*turoaidje*	*newänemä*
8	*(mer) abo thef*	*enàva me*	*iburo kàvé*	*turrawendsji*	*kabiàmà*	*ibudapändji*	*turobindje*	*gawirmä*
9	*(mer) abo ithin*	*pendö me*	*iburo péndji*	*turra éndoan*	*pandjimè*	*ibudajandoa*	-	*pendjimä*
10	*(mer) abo thonith*	*wendö me*	*iburo jàndo*	*ketsja nda kawé*	*jándoemè*	*ibudamadoa*	*kiljanta kuwe*	*jamdomä*
11	*(mer) abo merègh*	*mata me*	*iburo màdo*					*madomä*
12	*mer an thef*	*törwa me*	*iburo turò*					*teroamä*
13	*mer an thef abo sas*	*ni/ni mebö*	*nī/nī növere*					*ni*
14	*mer an thef abo thef*	*ni enàva*	*nī kàvé*					*ni-newere*
15	*mer an thef abo ithin*	*ni pendö*	*nī péndji*					*ningkawir*
16	*mer an thef abo thonith*	*ni wendö*	*nī jàndo*					*nimpendji*
17	*mer an thef abo merègh*	*ni mata*	*nī màdo*					*ninijamdo*
18	*löndör*	*ni törwa*	*nī turò*					*nimado*
19	*löndör abo sas*	*tarö/tarö mebö*						
20	*löndör abo thef*	*tarö enàva*	*tjī növere*					
21	*löndör abo ithin*	*tarö pendö*						
22	*löndör abo thonith*	*tarö wendö*						
23	*löndör abo merègh*	*tarö mata*						
24	*löndör abo mer*	*tarö törwa*						
36	*nif*							
40			*tjī kàvé*					
60			*tjī péndji*					
72	*nif thef*							
80			*tjī jàndo*					
100			*tjī màdo*					
108	*nif ithin*							
144	*nif thonith*							
180	*nif merègh*							
200			*tjī iburo jàndo*					

Figure 1. Map drawn based in information in Menanti and Susanto (2001) and Gordon (2005:801-880). Colours are according to obvious genetic groupings, as follows. Ndom, Kimaama-Riantana, Bulaka River (Yelmek and Maklew), Komolom (Koneraw and Mombum), Marind (Marind, Bian Marind and other languages not shown on map), Yei (Tanas-Bupul and Erambu-Poo), Moraori (Morori), Tonda (Sota, Smärky, Ngkâlmpw and Bädi Kanum, Guntai, Kunja, Blafe, Arammba and Wára) and Nambu (Neme, Nen, Nama, Namat, Namo).



3 Conclusion

We have traced the Kanum base-6 numeral system to the Tonda group specifically. We suggest that it arose for counting yams and then "climbed in" as a general purpose counting system. We provide glimpses of evidence for such stages – evidence which is hard to account for otherwise. Whether there is a historical connection with base-6 systems of the Kolopom languages, near, but not adjacent, to the west, remains an open question. If there is a connection, it is areal rather than genetic, but sufficient evidence for a pre-historic areal connection remains to be amassed. Equally, if not more, puzzling would be the conclusion that there is no historical connection, as base-6 counting systems are typologically very rare in the world as a whole. At the present state of data availability, we cannot go beyond such general directions, and, since there is no pre-20th century data, we may never be able to reach very deep into history.

References

Aubaile-Sallenave, Françoise & Serge Bahuchet. 1994. Nouvelle Guinée. In Serge Bahuchet (ed.), *Situation des populations indigènes des forêts denses et humides*, 353-386. Luxembourg: Office des publications officielles des Communautés européennes.

Ayres, Mary Clifton. 1983. This side, that side: locality and exogamous group definition in Morehead area, Southwestern Papua. University of Chicago doctoral dissertation.

Donohue, Mark. 2008. Complexities with restricted numeral systems. *Linguistic Typology* 12(3). 423–429.

Donohue, Mark. (no date). The Languages of Wasur National Park, Irian Jaya. Unpublished Manuscript, Sydney University, Australia.

Drabbe, Peter. 1949. Bijzonderheden uit de Talen van Frederik-Hendrik-Eiland: Kimaghama, Ndom en Riantana. *Bijdragen tot Taal-, Land- en Volkenkunde* 105. 1–24.

Drabbe, Peter. 1954. *Talen en dialecten van zuid-west Nieuw-Guinea* (Micro-biblioteca Anthropos 11). Posieux/Fribourg: Instituut Anthropos.

Galis, Klaas Wilhelm. 1955. Talen en dialecten van Nederlands Nieuw-Guinea. *Tijdschrift Nieuw-Guinea* 16. 109–118, 134–145, 161–178.

Geurtjens, Hendrik. 1933. Woordenlijsten der talen die het Marindineesche taalgebied begrenzen. In *Marindineesch-Nederlandsch Woordenboek* (Verhandelingen van het Koninklijk Bataviaasch Genootschap van Kunsten en Wetenschappen 71:5), 397-429. Bandoeng: Nix.

Gordon, Raymond G. Jr. (ed.). 2005. *Ethnologue: Languages of the World.* 15th edn. Dallas: SIL International.

Hammarström, Harald. 2008. Small Numeral Systems and the Hunter-Gatherer Connection. Paper Presented at the International Conference on Language, Communication and Cognition (LCC), Brighton, UK, August 4-7 2008.

Hammarström, Harald. 2009 [to appear]. Rarities in Numeral Systems. In Jan Wohlgemuth & Michael Cysouw (eds.), *Rara & Rarissima: Collecting and interpreting unusual characteristics of human languages* (Empirical Approaches to Language Typology). Mouton de Gruyter.

Hitchcock, Garrick. 2004. Wildlife is our Gold: Political Ecology of the Torassi River Borderland, Southwest Papua New Guinea. University of Queensland doctoral dissertation.

Kluge, Theodor. 1938. *Die Zahlbegriffe der Australier, Papua und Bantuneger nebst einer Einleitung ueber die Zahl; ein Beitrag zur Geistesgeschichte des Menschen.* Berlin-Steglitz.

Lean, Glendon A. 1986. *Western Province* (Counting Systems of Papua New Guinea 12). Port Moresby: Papua New Guinea University of Technology. Draft Edition.

Leont'ev, Aleksej Alekseevič. 1974. *Papuasskie Jazyki.* Moscow: Akademia Nauk SSSR.

Lyons, A. P. 1914. Appendix III g): Vocabularies of the Languages Spoken Between the Wassï Küssa and the Dutch Boundary. *Annual Report of British New Guinea* 1913-14. 193–194.

MacGregor, William. 1897. Appendix BB: British New Guinea. Native Dialects. *Annual Report of British New Guinea* 1895-1896. 100–120.

Meek, Charles K. 1931. *Tribal Studies in Northern Nigeria* volume 2. London: Kegan Paul, Trench, Trübner.

Menanti, Jacqueline & Yunita Susanto. 2001. Draft Report on the Kimaam District Survey in Papua, Indonesia. To appear in the SIL Electronic Survey Reports.

Nevermann, Hans. 1935a. Die Insel Komolóm. In *Bei Sumpfmenschen und Kopfjägern: Reisen durch die unerforschte Inselwelt und die Südküste von Niederländisch-Neuguinea*, 146-159. Stuttgart.

Nevermann, Hans. 1935b. Nach Frederik-Hendrik-Eiland. In *Bei Sumpfmenschen und Kopfjägern: Reisen durch die unerforschte Inselwelt und die Südküste von Niederländisch-Neuguinea*, 51-66. Stuttgart.

Nevermann, Hans. 1939. Die Kanum-Irebe und ihre Nachbarn. *Zeitschrift für Ethnologie* 71. 1–70.

Nevermann, Hans. 1942. Die Je-Nan. *Baessler-Archiv: Beiträge zur Völkerkunde* 24. 87–221.

Riley, Baxter E. & Sidney H. Ray. 1930, 1930, 1931. Sixteen Vocabularies from the Fly River, Papua. *Anthropos* XXV, XXV, XXVI. 173–193, 831–850, 171–192.

Serpenti, Laurentius M. 1965. *Cultivators in the Swamps: social structure and horticulture in a New Guinean society* (Samenlevingen buiten Europa 5). Assen: Van Gorcum.

Sohn, Myo-Sook. 2000. Report on the Merauke Subdistrict Survey Papua, Indonesia 11-21 October 2000. To appear in the SIL Electronic Survey Reports.

Sohn, Myo-Sook. 2006. Report on the Muting District Survey. SIL International, Dallas. SIL Electronic Survey Reports 2007-017 `http://www.sil.org/silesr/abstract.asp?ref=2006-006`.

van Baal, Jan. 1966. *Dema: description and analysis of Marind-Anim culture (South New Guinea)* (Translation series / Koninklijk instituut voor taal-, land- en volkenkunde 9). The Hague: Martinus Nijhoff.

van Baal, Jan. 1982. *Jan Verschueren's Description of Yéi-Nan Culture* (Verhandelingen van het Koninglijk Instituut voor Taal-, Land- en Volkenkunde 99). The Hague: Martinus Nijhoff.

Voorhoeve, C. L. 1975. The Central and Western Areas of the Trans-New Guinea Phylum: Central and Western Trans-New Guinea Phylum Languages. In Stephen A. Wurm (ed.), *New Guinea Area Languages and Language Study Vol 1: Papuan Languages and the New Guinea linguistic scene* (Pacific Linguistics: Series C 38), 345-460. Canberra: Research School of Pacific and Asian Studies, Australian National University.

Walker, Malcolm & Johszua Mansoben. 1990. Irian Jaya Cultures: An Overview. *Irian* XVIII. 1–16.

Williams, Francis Edgar. 1936. *Papuans of the Trans-Fly.* Oxford: Clarendon Press.

Wirz, Paul. 1922. *Die Marind-anim von Holländisch-Süd-Neu-Guinea* (Abhandlungen aus dem Gebiet der Auslandskunde: Reihe B, Völkerkunde, Kulturgeschichte und Sprachen 6). Hamburg: Friedrichsen. 2 vols.

Chapter VII | Rarities in Numeral Systems

Hammarström, H. (2009 [to appear]). Rarities in numeral systems. In Wohlgemuth, J. and Cysouw, M., editors, *Rara & Rarissima: Collecting and interpreting unusual characteristics of human languages*, Empirical Approaches to Language Typology, pages 7–55. Mouton de Gruyter.

Rarities in Numeral Systems

Harald Hammarström
Department of Computer Science and Engineering
Chalmers University of Technology
and University of Gothenburg
SE-412 96 Göteborg, Sweden
harald2@chalmers.se

Abstract

We present an extensive survey of rare structural properties in numeral systems in the world's languages, foremostly the question of rare number bases. The survey emphasizes comprehensiveness and status of evidence.

1 Introduction

The paper surveys rarities in numeral systems across the world. Space permits us only to look at the most conspicuous kinds of rarities that are featured in the vast set of languages in the world. The study aims at a high level of preciseness as to what counts as a numeral and what counts as rare, and doubtful cases will be treated pre-emptively in footnotes.

2 Numerals

2.1 What are Numerals?

In this paper, we define numerals as:

1. *spoken*
2. *normed expressions* that are used to denote the
3. *exact number* of objects for an
4. *open class of objects* in an
5. *open class of social situations* with
6. *the whole speech community* in question

With the first point we mean to disregard symbol combination systems, e.g., Roman numerals, that are confined to written communication (but, of course, essentially all of our primary data come from written representations of the spoken language).

The second point serves to exclude expressions that also denote exact numbers, but are not the normal or neutral way to say those numbers, e.g., 'eight-times-nine-and-another-two' for the normal 'seventy-four', but also to demarcate the area where the numeral system ends, which is, when there aren't any normed expressions.

As for the third point, languages usually have a rich set of expressions for inexact quantities, 'a lot', 'few', 'really many', 'about fifty' (but hardly *'about fifty-one') that have relatively high frequency in discourse. These are interesting in themselves but will not be included here because of their different fuzzy nature compared to exact number expressions.

Concerning the fourth point, some languages have special counting systems for a restricted class of objects (e.g., in Wuvulu (Hafford 1999:37-39) for counting coconuts). These can be quite idiosyncratic and since all languages which have exact enumeration must have a means for counting an open class of objects, it is preferable to study that, as it corresponds to a general kind of communicative need of a society.

The reason for the fifth point, the requirement on social situations, is to take a stand on so-called body-tally systems (cf. Lean 1992:2.4-2.6). A body-tally-system may be defined as follows. Assume a sequence of body parts beginning with the fingers of one hand continuing with some points along the lower and upper arm, reaching one or more points of the head, then ending with the corresponding body-parts on the opposite arm and finally hand. A number n is then denoted by the nth body-part-term in the sequence, e.g., 'nose' or 'elbow on the other side'. There are features that distinguish body-tally systems from other counting systems with etymologies from body parts. Non-body-tally systems use only fingers, toes, hands, occasionally eye and head, whereas body-tally systems always use some intermediate points, such as elbow, shoulder or nose, and let them form a sequential order from one side of the body to the other. Typically, body-tally systems are only used in special circumstances, such as bridal price negotiations, and in other cases you would use a different numeral system or not use exact enumeration at all. The information on the social status of the body-tally numeral systems is very incomplete; We can say that for the vast majority we do not have such information, but for those in which we do, the social situation restriction applies. Body-tallying has to be done on a physically present person and to understand what number is referred to the process must be watched, so, for instance, body-tallying numerals would be infelicitous when it is dark. For instance, de Vries (1998) found that body-tally numerals in a Bible translation could not be understood, i.e., were often mis-translated back to Indonesian by bilingual persons. Of course, there could be some other language(s), unknown to us at present, where body-tally numerals can be used in a fully open class of social situations; such a body-tally system would accordingly be included in the study. Body tally systems are attested

in abundance in Papua New Guinea and Indonesian Papua, in a geographically continuous area centered at the Ok family and, even if in decline, are still used today. Although many writers have neglected to mention it, there are also indisputable attestations of long extinct body-tally systems from Kulin (Pama-Nyungan, Australia) varieties in southeast Australia (Howitt 1889:317-318) (Howitt 1904:697-703)!

Finally, regarding the sixth point, we are not interested in numeral systems which are particular to some small subsets of the speakers of the language in question (e.g., professional mathematicians) because such systems might not respond to the conditions and needs of the majority of a society.

Numerals provide a good testing bed for patterns across languages given their comparatively clear semantics and modularity. As to numeral semantics, languages may differ as to which quantificational meanings they express/lexicalize, notably in approximate numeration and whether a counted set of objects constitute a group or not, but these matters are minor compared to differences languages show, e.g., in verbal tense/aspect. Likewise, although not universally, numerals tend to have uniform, clearly identifiable, syntactic behaviour within a language. Also, if two languages have exact numeration for a certain range of numbers, one expects the two to give a similar functional load to these expressions, excluding possibilities such as numbers also being used for, say, colours or as metaphors significantly wider in one language or the other. This appears sound also in the light of the only corpus study of numeral frequencies in a language with a restricted numeral system – McGregor (2004:204) – which shows that 'one' and 'two' in Gooniyandi (Bunaban, Australia) occur with comparable frequency to 'one' and 'two' in English.

2.2 Rareness

In this paper we present cases that are rare, either in that (a) they are present in few languages or in that (b) they are present in few geographical spheres. Most cases are of the (a)-kind, but for example, base-12 systems in northern Nigeria are present in relatively many languages, from several different families, but are confined to just this geographical sphere, so they are counted as rare in the sense of (b) only. Geographically separate instances are likely to be independent, and the bottom line is that we are interested in rare independent innovations – whether or not they have grown genetically or areally onto many languages.

2.3 Survey

Lots of data is available in one form or another for numerals. It seems that numerals together with pronouns, kinship terms, body part terms, and other basic vocabulary (sun, water, etc), and perhaps "sketchy" phonological inventory, are the parts of language where there exists empirical data for a really large subset of the world's known languages. One may legitimately ask just how large this subset is when it comes to numerals – for how many languages do we have data

on numerals? Let's say we count about 7 000 attested native spoken languages for the world. A definite lower bound is 3 880, since we can produce a list of references to numeral data from 3 880 definitely distinct languages. An upper bound is harder to give. We entertain the rather time-consuming methodology of trying to obtain every first-hand descriptive data reference found in any handbook or relevant publication whatsoever. The survey in the present paper is based on the data we have collected so far. We currently have about 13 500 references, some describing numeral systems of many languages in the same publication, and, with 7 000 languages in the world, many different publications describe the same language. (The fact that often there is more than one independent source for one and the same language helps us to determine the accuracy.) It is impossible at this point to say how many languages the sources account for since they attest dialectal varieties, varieties from the same location but different centuries, partial data, data of varying quality, duplicated data, etc. However, at least one language from every attested language family or isolate is included in the survey (if numeral data is at all attested for the family in question).

In addition to first hand sources, we have also drawn inspiration from the rich existing literature on numerals in general. The subject, in fact, goes back more than 200 years in time – the first major work being the remarkable *Aritmetica Delle Nazioni* by Hervás y Panduro (1786). Since then, our bibliography counts some 20 PhD:s, over 100 further monographs and 700+ articles to have appeared. These range from purely descriptive accounts to areal, comparative-historical, typological, and deep syntactic studies – solely devoted to spoken language numerals as defined above. (The literature on written symbol systems for mathematics is even more voluminous.) However, since most of the literature just re-hashes the same data, the recourse to first-hand sources is essential in order to understand the true diversity in numerals in the world's languages.

3 Rarities

3.1 Rare Bases

Perhaps the most salient single characteristic of a numeral system is its base, or more correctly speaking, its set of bases. The *set of bases* of a natural language numeral system may be defined as follows.

> the number n is a base iff
>
> 1. the next higher base (or the end of the normed expressions) is a multiple of n; and
> 2. a proper majority of the expressions for numbers between n and the next higher base are formed by (a single) addition or subtraction of n or a multiple of n with expressions for numbers smaller than n.

This assumes that for any expression the linguist can unambiguously analyze each numeral expression into its constituent parts (or analyze it as consisting of only one part). As an example, for Swedish we would begin by finding the biggest part of the highest normed expression, which according to our own knowledge is *miljard* (10^9). Thereafter we can find the next lower base by trying divisors x of 10^9 to see if the numbers between x and 10^9 are expressed in the required form. For example, $x = 5 \cdot 10^8$ is not, because we do not say **en-halv-miljard plus ett* (*half-a-billion plus one) or the like for $5 \cdot 10^8 + 1$ or any, let alone a majority, of the numbers between $5 \cdot 10^8$ and 10^9. However, 'miljon' (10^6) fulfils the requirements, and, continuing with the same analysis for lower and lower numbers, we arrive at the conclusion that Swedish has $\{10, 10^2, 10^3, 10^6, 10^9\}$ as its set of bases.

The definition of base as stated gives unambiguous decisions for formations which are sometimes (and sometimes not) called base by other authors; systematic subtractions, special lexemes for base-multiples, or isolated cases of addition, e.g., only 7=6+1 but otherwise no additions involving 6. Examples of such cases and their systematic resolution with our definition are given in Table 1. It is important here to note that there doesn't have to be a monomorphemic word for something that is a base. In the case of Kare, at least if we assume that the numbers above 20 are formed parallel to 30, then 20 is a base. Further, 10 or 15 are not bases even though the words for them monomorphemic – the definition interprets them as special words for multiples of 5, just like some base-10 systems have monomorphemic words for 20, 30, ..., 90.

The expression 'base-x system' will be used to mean that 'x is in the set of bases' for the numeral system in question. Similarly, 'base-x_1-...-x_n' system will mean that all of x_i is in the set of bases, without any commitment that the $x_1, \ldots, x_n$ exhaust the set of bases.

No Base

There are a number of languages for which there is an explicit statement in the descriptive literature that they lack (exact) numerals above one:

Nadëb (Nadahup, Brazil): According to Weir (1984:103-104), the words for 2 and 3 are inexact. The vocabulary of a closely related variety lists completely different words for 1-3 (Schultz 1959) and the study by Münzel (1972) lacks information on numerals (cf. Epps 2006:263). We have not seen the wordlist collected by Natterer (Koch-Grünberg 1906:881), though this might not include numerals anyway.

Pre-contact Jarawara (Arawán, Brazil): According to Dixon (2004:559) and indeed the only other published wordlists for Jarawara (and closely related varieties) show some overlap between forms for 2, 3, 'few' and 'many' (Anonby and Anonby 2007:25).

Pre-contact Yuqui (Tupi-Guaraní/Tupí, Bolivia): According to Villafañe (2003:68). As far as we are aware, there are no other published descriptions of this language that include the numerals.

	Lutuami Klamath-Modoc, USA (Dixon and Kroeber 1907:673)		**Nyokon** Bantoid/Atlantic-Congo, Cameroon (Richardson 1957:30)		**Kare** Bantu/Atlantic-Congo, Sudan (Dijkmans 1974:147)		**Ainu** Isolate, Japan (Refsing 1986:110)	
	Analysis	Expression	Analysis	Expression	Analysis	Expression	Analysis	Expression
1	1	nas	1	ámɔ̀	1	emotí	1	sine
2	2	lap	2	àfɔ́ɔ̀	2	ibili	2	tu
3	3	ndan	3	átár	3	etotu	3	re
4	4	umit	4	ɟnnǹ̩s	4	biu	4	ine
5	5	tunip	5	ɟtɔ́ɔ̀r	5	etano	5	asikne
6	5+1	nas-ksapt	6	átʃʃín	5+1	etano na emoti	10-4	iwan
7	5+2	lap-ksapt	6+1	ɟtʃʃín námɔ̀	5+2	etano na ibili	10-3	arwan
8	5+3	ndan-ksapt	?	ɟyáá nǹ̩ màn	5+3	etano na etotu	10-2	tupesan
9	10-1	nas-xept	8+1	ɟyáá nǹ̩ màn námɔ̀	5+4	etano na bînu	10-1	sinepesan
10	10	te-unip	10	àwát	10	la-ato	10	wan
11	10+1	taunep-anta nas	10+1	àwát ámɔ̀	10+1	laäto na emoti	10+1	sine ikasma wan
...	...	...	...	...	...	...	...	...
15	...	...	...	...	15	sanga		
16	...	...	...	...	15+1	sanga-na-emoti		
...	...	...	...	...	...	...		
20	2x10	lap-eni taunep	20	nǹ̩tʃʃín	2x10	atumbili	20	hot
21	2x10+1	lap-eni taunep-anta nas	20+1	nǹ̩tʃʃín ámɔ̀	...	...	20+1	sine ikasma hot
...	...	...	...	...	...	...	...	...
30	3x10	nda-ni taunep	3x10	àwát átár	2x10+10	atumbili na laato	20+10	wan e tu hot
...	...	...	...	...	...	...	...	...
40	...	...	...	...	...	...	2x20	tu hot
Base	**5-10**		**10**		**5-20**		**5-10-20**	

Table 1. Examples of formation types and outcomes of the definition of base (see text).

Canela-Krahô (Jê/Jê-Jabutí, Brazil): According to Green (1997:181). As far as we are aware, there are no other published descriptions of this variety that include the numerals.

Krenák (Aimoré, Brazil): According to a synthesis of earlier data by Loukotka (1955:125-126) which follows observations such as Renault (1903:1111). Even if there were no normed oral expressions, small numbers could be communicated using fingers on the hand (Ehrenreich 1887:41-46).

Parintintin (Tupí-Guaraní/Tupí, Brazil): According to Nimuendajú (1924:240-241). Indeed, the larger dictionary by Betts (1981) agrees that the word frequently glossed as 'two' (cf. Sampaio 1997:57-58) actually has an inexact meaning.

Wari' (Chapacura-Wanham, Brazil): According to one vocabulary collected by Hanke (1956). A later, more extensive, description of a variety in the same dialect cluster does show a word for 'two' albeit glossed literally as 'facing each other' (Everett and Kern 1997:452-459). An attempt at documentation of the most closely related language, the moribund Oro Win, failed to uncover any number words (Popky 1999:38).

Chiquitano (Isolate, Bolivia): According to Adam and Henry (1880:19) which is corroborated by d'Orbigny (1839:163) and Clark (1937:118-119,138) and several later attestations of Chiquitano dialects show Spanish (Nordenskiöld 1911:232) (Nordenskiöld nd) (Tormo 1993:15,108) or Portuguese (Santana 2005:94) loans for 'two' and above. However, there are also dialects where a native term for 'two' is attested (Montaño Aragón 1989:335-400).

"All" Campa and Machigenga groups (Pre-Andine/Arawak, Peru): According to Wise and Riggle (1979:88). As far as we are aware, published vocabularies (too many to list) show little indication that the words given for 'two' (and sometimes above) are in reality inexact. However, Wise and Riggle (1979) did work with basic mathematics education among these groups and therefore their judgement is arguably deeper.

Culina (Arawán, Peru): According to Wise and Riggle (1979:88). Unfortunately, we have not had access other materials on either Brazilian or Peruvian Kulina to double check the claim.

Arabela (Zaparoan, Peru): According to Wise and Riggle (1979:88), though the later, quite extensive dictionary of Rich (1999) does show distinct expressions for 'two' and 'three'. Possibly, Wise and Riggle (1979) who did work with basic mathematics education looked at these expressions and their meaning more closely.

Achuar (Jivaroan, Ecuador): According to Wise and Riggle (1979:88), though later more extensive descriptions show expressions for 'two' and higher numerals (Fast and Fast 1981:58-59) (Fast et al. 1996). It is possible that expressions for 'two' and higher numerals crystallized as a result of increased

contact with a counting culture (Gnerre 1986) or even reflects normative rather than descriptive usage. Therefore, Wise and Riggle (1979) who did work with basic mathematics, could very well be descriptively more accurate for the traditional state of the language.

Fuyuge (Goilalan, Papua New Guinea): One early description of Fuyuge says that the 'two' word is also used for a small number (Ray 1912:313-314). However, there is a word listed as 'three' but no explicit statement to the fact that this, like 'two', also has an inexact meaning. A very small vocabulary, probably collected by the same person lists 1,2,2+1 and no further comments (Fastre 1920:116), and the later, more modern description by Bradshaw (2007:45) attests a native 1,2,2+1,2+2, ... system.

Viid (Border, Indonesia): In one wordlist (a.2) of Viid from Senggi (Smits and Voorhoeve 1994:211-212), 'tambla' is listed both with the meaning 2 and 3, but this is not borne out in other early wordlists (Smits and Voorhoeve 1994:211-212) or the more recent (Menanti 2005), which have 3=2+1.

Gedaged (Oceanic/Austronesian, Papua New Guinea): Nikolaj von Miklucho-Maclay, a pioneer researcher on the Rai-coast of Papua New Guinea, reports that (von der Gabelentz and Meyer 1882:503):

> Sehr viele Papuas kennen die Zahlwörter ihres eigenen Dialektes nicht. In Mitebog [a village speaking a dialect of Gedaged – HH] fragte ich fünf oder sechs Eingeborene, aber die Angaben waren widersprechend und jedenfalls unrichtig, nur olam (eins) konnte ich als sicher notiren [sic!].

One interpretation of this statement is that there was no normed expression for numerals above one in the lect of Mitebog. A later, longer description of a different dialect shows monomorphemic numerals 1-5 inherited from Austronesian (Dempwolff nd:36-37).

To lack numerals above one means that the normed expressions for the quantities above one are inexact. We may call such systems 1-few-many for the time being. In these languages, it may be possible to communicate a higher exact quantity successfully, perhaps using gestures, context, one-to-one pairings, repetition or a specialized lexical item e.g., 'twin' for a certain kind of exact quantity. However, in these languages, the normed expressions are still 'one', 'a few', 'many', ... when these quantities occur in discourse. In no case does it appear to be possible, or normed, to say few+1, 1+1 or few+few to designate an exact number, so there is no base.

From the above cases, one certainly gets the impression that there is a thin line between 1-few-many systems and 1-2-many systems. In some cases, different observers on the same language variety differ as to whether the 'two'-word is approximate or exact in meaning. In other cases, the speech community seems to have acquired norms for number expressions over time. One may then conjecture that many more 1-few-many systems would have been found if more

languages had been documented in detail before extensive contact with modern society.[1] It is also apparent that questions on this level of granularity are

[1] Further cases may include the following. 1. Aikhenvald and Dixon (1999:358) conjecture that Djeoromitxí (Jabutí/Jê-Jabutí, Brazil) "originally had no numbers" since the word *je-bo* for 'two' given by Pires (1992:66) is from a root with the meaning to 'be equal'. However, an etymology for 'two', even if correct (cf. van der Voort 2004:212 and van der Voort 2007:162) does not automatically mean that there was no original word for 'two', nor that a present meaning of 'two' (Ribeiro 2008:42) is somehow subordinate to the etymological meaning. Also, early attestations of yawo yawo (2+2) for 4 in (Loukotka 1963:50) speaks against an inexact meaning for 'two'. 2. Barriga Puente (1998:132,263) reports that Esmeraldeño (Isolate, Ecuador) has a limit of counting at one, based on a misreading of Lehmann (1920:37). There is only one vocabulary of the now extinct Esmeraldeño which has been reprinted a number of times (Adelaar 2004:155-161). However, the earliest of these publications (Wolf 1892:528) is clear that the lack of native Esmeraldeño words above one could be due to the memory of the last speaker(s). So we are not in a position to assert that Esmeraldeño ever lacked numerals above one. 3. Koch-Grünberg (1928:316) describes numerals in Sapé (Isolate, Venezuela) as 1 'mẹyakán' and 2 'mẹ̊yakán' and remarks "Die Kaliána haben nur ein Zahlwort und gebrauchen stets denselben Ausdruck, in dem sie dabei an den Fingern und dann an den Zehen weiterzählen". It's not clear what to make of this, but, in any case, the only other two published vocabularies do show distinct words for lower numerals (de Matallana and de Armellada 1943, Migliazza 1978) and Koch-Grünbergs vocabularies contain other cases of puzzling numeral elicitations (Zerries and Schuster 1974:56). 4. It is hard to know whether the Guayakí variety recorded from two youths by Vogt extended to a whole community of speakers (Vogt 1903:861) and another attestation from roughly the same time appears to give forms for exact 1 and 2 (Mayntzhusen 1920:20), though it may be that these forms are etymologizable (Vellard 1935). 5. On the grounds that the present-day numerals can be etymologized to 'that', 'pair/couple', 'few' and 'another', proto-Tupi (Schleicher 1998:12-13) may be argued to lack numerals. 6. A vocabulary of Ofaye has 1 hœhá, 2 ñokoádi, 3 ñokoádn 4 ñokoádi (Hanke 1964:29), i.e., 2 is the same as 4. A good guess, following more recent documentation (das Dores de Oliveira 2006:109-110), is that the 4 in this earlier vocabulary is simply an error of some kind. 7. Bernatzik (1942) claims that Yumbri lacked numerals above one. There is no further material on this variety but the closely related Minor Mlabri (Rischel 1995) has numerals up to three. Bernatzik's account has a sweeping and condescending flavour, and also has other doubtful claims of the same kind, e.g., lack of fiction which does not hold for Minor Mlabri either (cf. Velder 1963:15). Another complicating factor is that he is able to discuss twin births at length with the people he says cannot comprehend any more distinctions than 'one' and 'many'. 8. The oft-repeated claim (Parker 1909:85) of lack of numerals in Vedda (Unclassified, Sri Lanka) appears, on closer scrutiny of the underlying sources, to be hearsay (Seligmann and Seligmann 1911:33,412). The only thing we can say is that no native term above two is could be collected from the memories of the descendants, which does not necessarily mean that none existed. 9. The first record of the language of Utanata (Asmat-Kamoro/Trans New Guinea, Indonesia) indicates counting inability on the part of the inhabitants (Earl 1837). However, lower numerals are attested in all subsequent descriptions – especially the most extensive piece (Drabbe 1953) – and have cognates in other Asmat-Kamoro languages (Galis 1955). Therefore, the counting inability reported probably reflects some kind of misunderstanding in the midst of the very difficult communication circumstances. 10. Grondona (1998:91) conjectures that pre-contact Mocoví (Guaicuruan, Argentina) lacked numerals above one as 2 and above are Spanish loans ("It seems that Mocoví lacked numeral forms, and has borrowed all its numerals from Spanish"). While it is true that Mocoví borrowed 2 and above from Spanish (cf. Gualdieri (1998:211-212) and for the related Pilagá (Vidal 2001:129)), it does not necessarily follow that Mocoví lacked 2 and above, before the borrowing. Older sources do, in fact, consistently attest a specific form for 2, see Koch-Grünberg (1903:112-124) as well as Lafone Quevedo (1893:244) and Lafone Quevedo (1892:410). 11. Paiconeca (Bolivia-Parana/Arawakan, Bolivia) is a poorly attested extinct language of presumed Arawakan affiliation (Montaño Aragón 1989:161-

almost beyond the scope of classical forms of language documentation. Of languages potentially showing 1-few-many systems or 1-2-many systems only two, Mundurukú (Mundurukú/Tupí, Brazil Pica et al. 2004) and Pirahã (see below), have been subject to investigations approaching standards of experimental psychology.

There are two further languages in the Amazon, Pirahã (Mura-Pirahã, Brazil) and Xilixana (Yanomama, Brazil) that stand apart from the above 1-few-many systems in that they are argued to lack all exact numerals, i.e., there is no normed way to denote an exact quantity even for 'one'.

In Pirahã, there are two words which prototypically mean 'one' and 'a couple' respectively, but it has been checked fairly extensively that their meanings are fuzzy 'one' and 'two' rather than discrete quantities (Everett 2005, 2004, Frank et al. 2008). It is not possible to combine or repeat them to denote higher (inexact?) quantities either (Gordon 2004). The Pirahã have the same cognitive capabilities as other humans and they are able to perform tasks which require discerning exact numeration up to the subitizing limit, i.e. about 3 (Gordon 2004). They just do not have normed expressions even for low quantities, and live their life happily without paying much attention to exact numbers. It does not appear to be possible to express an exact quantity simply by repeating an expression the appropriate number of times, like one can and often does in, e.g., Sanuma (Yanomama, Brazil) for 2 and 3 (Borgman 1990:152). If one says "I'll be back after it gets dark and it gets dark again" this might just as well be interpreted as two days or as three days (p.c. Daniel L. Everett 2005). It seems relevant to note that Pirahã grammar lacks singular-plural distinctions of any kind, even in pronouns (p.c. Daniel L. Everett 2008). A wordlist of the only known relative of Pirahã, the extinct Mura language[2], features words glossed 'one' and 'two' (Nimuendajú 1932, Nimuendajú and do Valle Bentes 1923). The 'one'-word is an obvious cognate to the Pirahã fuzzy one, and the 'two'-word is an obvious loan from some Tupi language.

Xilixana is the language of a group which has been on the Mucujai river at least for the past century. In modern divisions, it is sorted as a dialect of Ninam, also known as Yanam or Central Ninam (superseding Southern Ninam in older

173). The naturalist d'Orbigny (1839:191) travelled through the area in the 19th century and is the only source for numerals in the language. Since this is the only source, we can neither confirm or deny his report of lack of numerals:

> Il n'y a, dans cette langue, aucun système de numération, qu'y remplacent à peine quelques termes de comparaison, eux-mêmes, très-bornés.

. 12. In all descriptive publications, Khoedam (Khoe/Khoe-Kwadi, Namibia) |úí and |áḿ are glossed as 'one' and 'two' respectively, but closer inspection reveals that these are really meanings accustomed to linguistic elicitation, and 'singleness' and 'dualness' are more appropriate glossings. There is a subtle difference between 'dualness' and 'twoness' in that dualness implies an association between the items in question. So |áḿ children would mean twins rather than two children. If this difference is deemed significant, then there was no word for 'two' in traditional Khoedam (Brenzinger 2009).

[2] We regret that we have not been able to access two relevant-looking publications on the Mura language (Hanke 1950, 1952).

terminology) (Migliazza 1972). Swain (2000)[3] describes Xilixana numerals as not even having an exact 'one':

'one'	mõli	Note: Means one or a few.
'two'	kup; yalukup	Note: Means two or a few.
'three'	pək	Note: Can refer to any number more than two or a few.

John Peters, the first missionary to live among the same group, also describes the same expressions as having inexact value and adduces that "exact numbers were not important" (Peters 1998:52). The closest other Yanomami variety for which there is a grammar is the dialect Shiriana, of the Uraricoera, to the north (Gómez 1990). This describes the numerals 'one' and 'two' as exact, but the author only spent 14 weeks in the field. Also Migliazza (1972:117-118,422), who spent many years in all of the Yanomama territory, describes Shiriana lower numerals as exact in the numerals section of his thesis and, in fact, all other description of Yanomama languages we have been able to consult describe 'one' and 'two' as exact (Ramirez 1994b,a, Zerries and Schuster 1974, Becher 1960, Knobloch 1967, Vinci 1956, Wilbert 1962, de Matallana and de Armellada 1943, Koch-Grünberg 1928, Mattei-Müller 2007). Also, most Yanomama varieties have singular, dual and plural but we do not know the precise status of Xilixana, and if so, if they are inexact as well. However, on one page (Migliazza 1972:38) the #moli word is glossed as 'one, few' (in contrast to pages 117-118 and 422). This is significant because language descriptions rarely claim 'one' and 'few' overlap in meaning, and now three independent observers do it for the same or nearly the same language. Carole Swain was a UFM/MEVA missionary who lived with the Xilixana for very long periods of time in the 70-90s and therefore she is certainly not a superficial observer. The Xilixana were monolingual (except for an occasional captured Dekwana) and uncontacted by modern society up to at least 1957 (Early and Peters 2000).

Base-3

Base-3 appears to be rarer than base-4. We have found only a few cases[4], some of them somewhat sporadic within their respective dialect cluster:

Ambulas of Wingei (Ndu, Papua New Guinea): An Ambulas dialect survey (Wilson 1976:57) mentions that the variety of Wingei counts in units of three, and the actual forms can be found in Wilson (1989a:16-17). The forms are reproduced in Table 2. Presumably, this is the same case that Laycock (1970) refers to when speaking (without forms given) of base-3-6-24 system(s) in the Ndu family, citing personal communication from

[3] Carole Swain has also submitted the same information for the Ninam entry for the Numeral Systems of the World's Languages website at `http://lingweb.eva.mpg.de/numeral/Ninam.htm`, accessed 1 July 2009.

[4] Ross and Paul (1978:60) give expressions for 1-8 in Waskia (Adelbert Range/Trans New Guinea, Papua New Guinea) with the structure 1,2,2+1,2+2,2+2+1,(2+1)+(2+1),(2+1)+(2+1)+1,(2+1)+(2+1)+2, that is, 6-8 are formed with additions based on (2+1)+(2+1) for 6. This comes close, but does not count as base-3 according to the definition used in this paper.

Anthony Forge. The etymology of the forms reveal that the system is much like a commonplace 5-10-20 or 5-20 system except that the hand is seen as having six features! At the time of elicitation only older people knew the indigenous system, whereas the young used Tok Pisin or English for higher numerals. Other, better described, varieties of Ambulas (Wilson 1976, 1980) show no base-3 and comparative evidence shows that the original Ambulas (1-3) and Ndu (1-2) system were restricted (Aikhenvald 2008:595, Laycock 1965:173-174).

Waimirí of Atroarí (North Amazonian Carib/Cariban, Brazil): Base-3 counting could be used up to about 9 according to Green (1997:6-7) , who cites personal communication with Ana Carla de Bruno Santos. However, the more recent grammar by Bruno (2003:140-142) states that Portuguese loans are used above 3 and is silent about a possible base-3 alternative.

Som (Finisterre-Huon/Trans New Guinea, Papua New Guinea): According to Smith (1988:29) base-3 counting can be used up to about 9. We know of no other description of this variety.

Bine (Eastern Trans-Fly, Papua New Guinea): In at least on vocabulary repoduced in Wolfers (1972:218) and Wolfers (1971:79), a variety of Bine is base-3 and reaches up to 9. However, all other attestations of Bine show only a restricted system and/or a body-tally system (Lean 1986d), including the lengthiest description (Fleischmann and Turpeinen 1975:16). The base-3 vocabulary must therefore be considered somewhat dubious.

Bukiyip (Arapeshan[5], Papua New Guinea): Fortune (1942:58-60) describes the Rohwim dialect of Mountain Arapesh to have a base-3 system for counting some objects and a base-4 system for counting other objects, which seems to have reached up to 24. A later description of an inland Bukiyip (Conrad and Wogiga 1991:73-76) variety shows a conflation of the two systems (with no indication of them being used for different objects). Robert Conrad submitted the base-3 system for the entry on Bukiyip (dialect not indicated) on Numeral Systems of the World's Languages website[6]. Available data on other Arapeshan languages, such as Abu' Arapesh (Nekitel 1985:82-84) and Mufian (Conrad et al. 1978:104), show base-5, at least from 7 and up.

Base-4

Base-4 systems are attested on four continents:

North America: Some extinct Chumash languages (Chumashan, USA) show original base-4 systems, running up to 32 (Mamet 2005:113-115) (Beeler 1967, 1963, Hughes 1974). Base-4-8 is also documented with the older generation in the now extinct Yuki (Isolate, USA). For Yuki, Kroeber (1925)

[6] Shown at `http://lingweb.eva.mpg.de/numeral/Bukiyip.htm`, accessed 1 July 2009.

	Wingei	Maprik	Wosera-Mamu	Wosera-Kamu-K
1	nawurak	nakurak	vétik	vétik
2	vétik	vétik	vétik	vétik
3	kupuk	kupuk	kupuk	kupuk
4	kupukiva	nakwasa/wan vétik wan vétik	vétik vétik	vétik vétik
5	kupuk'etik	naktaba	taambak	taambak
6	taabak	naktaba sékét naktaba nakurak		
7	taabak kaayek	naktaba sékét naktaba vétik		
8	taabak kaayek vétik	naktaba sékét naktaba kupuk		
9	taabak kaayek kupik	naktaba sékét naktaba kupuk wan vétik wan vétik		
10	vétik taaba vétik	taaba vétik		
11	nawurak taaba vétik	taaba vétik sékérék maan-ba kayék nakurak		
12	taaba vétik	taaba vétik sékérék maan-ba kayék vétik		
20		maan vétik taava vétik	nakurak mi	nakurak dumi/maan vétik taaba vétik
24	nawura mi			

Table 2. Numerals in Wingei Ambulas (Wilson 1989a:16-17), Maprik Ambulas (Wilson 1980), Wosera-Mamu Ambulas from around Serangwandu (Wilson 1989b:15) and Wosera-Kamu-K from around Kunjingini (Wilson 1990:15). Etymologies of roots are as follows #maa is 'foot, leg', #taaba is 'hand, arm', #mi is 'tree' and #du is 'man'. Apparently, in Wingei counting, the hand is seen to have six features. The etymology of the expresssion *nawura mi/nakurak mi/nakurak dumi* is not clear but it may have to do with either tree (typologically unusual but matches *mi*) or man (typologically very common, but resembles only *dumi*).

describes how base-4 is related to hand-counting by considering the spaces between the fingers (cf. Hinton 1994)[7]. The Chumashan languages and Yuki are both in California but quite distantly apart, with Yuki in the north and Chumashan in the south, and other language families intervening.

South America: The extinct Lule (Isolate, Argentina) of Clark (1937:102) and Machoni de Cerdeña (1732:84-86) as well as the poorly attested extinct Charrúa (Charruan, Uruguay) reported in (Ibarra Grasso 1939b:202) appear to have had base-4 up to 10, at which point the system turns into a commonplace 5-10-20 system with hands and feet. It cannot be inferred from the data hand that there was ever true base-4 system here, beyond 10. A couple of descriptions of a Guaraní variety in Paraguay (Tupí-Guaraní/Tupi, Paraguay) show base-4 up to 10, but the expressions for numbers above 10 are not shown (Ibarra Grasso 1938:278) (Ibarra Grasso 1939a:590). Other old and new descriptions of any varieties of Guaraní (too many to list) do not show any traces of base-4. Isolated vocabularies of Mocovi and Toba (Guaicuruan, Argentina) show base-4 up to 8 and 10 respectively (Koch-Grünberg 1903:114-124), but the vast majority of vocabularies for these languages (too many to list) show no trace of this. The extinct Payaguá (Isolate[8], Paraguay) has one attestation with alternative base-4 forms up to 20 (Koch-Grünberg 1903:114-124). All these cases occur within a relatively small area of South America, but there is otherwise little evidence for an areal connection.

Indonesia/Papua New Guinea: An indeterminate number of languages in the highlands have a variations of a base-4 system (Lean 1986a,c, 1992:13-86,15-59,Ch. 5), where at least one, Kakoli (Hagen/Trans New Guinea, Papua New Guinea) is attested with as base 4-24 (Bowers and Lepi 1975). Kewa (Engan/Trans New Guinea, Papua New Guinea) has several parallel numeral system, one of them being base-4 (Franklin and Franklin 1962) and goes at least up to 20, and beyond that it may be combined with a body-tally system to form higher numbers in units of four (Pumuge 1975). The word for '4' is 'hand', i.e., four fingers constitute one hand and the thumb is separate. The traditional counting system of Mbowamb (Hagen/Trans New Guinea, Papua New Guinea) near Mt. Hagen has been been described with a fair amount of detail. It is clearly a 2-4-8 system, for which Vicedom and Tischner (1948:268-270) gives expressions up to 24, and says the system can be used up to about 80. Another description seems to indicate that after 20, counting can be done in units of 20 (Strauss 1962:315-318), cf. also Lancy and Strathern (1981). As in Kewa, the base-4 is connected with counting the fingers of one hand, the thumb counted separately. The origin of the highland base-4 system(s) has not been

[7] I wish to thank Peter Bakker for highlighting this reference to me.

[8] Payaguá, though poorly attested, is often counted as related to (at least) the Guaicuruan languages (Viegas Barros 2004) but we do not think the evidence is conclusive.

systematically investigated, but given the geographical proximity and the fact that the Engan and Hagen languages are not closely related, an areal connection seems likely even if this is not directly observable in the forms in question.

On the north coast, around the border between Indonesian Papua and Papua New Guinea, base-4 is also present variously in the Sko languages (most of the data is collected in Lean 1986b), – see Donohue (2008) for a good attestation of 4-12-24 in Skou – as well as 4-24 in Tobati (Sarmi-Jayapura Bay/Austronesian) for which the best attestion is Moolenburgh (1904). Given the proximity of the languages and the fact that they are genetically unrelated, there is almost certainly an areal connection between base-4 in Skou and the Sarmi-Jayapura Bay Oceanic languages.

Africa: An indeterminate number of languages in the northeastern Democratic Republic of the Congo (DRC) have (traces of) a base-4 system. The first attestation appears to be a Nyali (Bantu, DRC) variety for which Stuhlmann (1894:624) notes that 8=2*4, 9=2*4+1, 13=12+1, 14=12+2, 16=2*8, 17=2*8+1 but 20=2*10. Later reports of related Bantu varieties show that there was/is a fully systematic 4-24 or 4-32 underlying these forms (van Geluwe 1960, Kalunga Mwela-Ubi 1999, Bokula and Ngandi 1985). Furthermore, thanks to Kutsch Lojenga (1994:353-357), we have a full attestation of almost obsolete Ngiti (Lendu/Central Sudanic, DRC) and Lendu (Lendu/Central Sudanic, DRC) 4-32 systems (p.c. Constance Kutsch Lojenga 2007). Various wordlists attest traces of the same base-4 systems in decay or amalgamation with base-10 and base-20 in closely related Bantu and Central Sudanic languages (Johnston 1922b, Struck 1910, Johnston 1904, Bokula 1970, Harries 1959, Lojenga 1994, Schebesta 1966, 1934, Asangama 1983, Czekanowski 1924, Stuhlmann 1917) and unpublished SIL survey lists.

In addition, there are a number of languages which have been claimed to be base-4 in the literature but which are not base-4 according to the definition used in this paper. We will mention a few of the most important ones here. The language called Āfúdu (Unassigned[9], West Africa) by Koelle (1854) uses some additions with 4 in the numbers below 10 but is decimal in the range 10-20. Bodo and Deuri (Bodo-Garo/Sino-Tibetan, India) have vestiges of base-4 counting extending higher than 20 and Bai (Bai/Sino-Tibetan, China) is documented with a base-4-16-80 system for shell money in medieval times (Mazaudon 2007). Yiwom (West Chadic A/Afro-Asiatic, Nigeria) has 7-9 as 4+3,4+4,4+5 but no other forms are based on 4 (Ibriszimow 1988). de Castelnau (1851a:10-13) reports base-4 (actually base-2-4) in Apinayé (Jê/Jê-Jabutí, Brazil) but no actual forms are given (de Castelnau 1851b:270-274) and is likely to be spurious in the absense of corroborating data in this rather well-documented language (too many references to list). Base-4 for counting special objects is widely

[9] This language has not yet been identified with any modern variety (p.c. Jouni Filip Maho 2004, p.c. Roger Blench 2009).

attested in the Oceanic languages of Melanesia (Kolia 1975, Friederici 1912, Parkinson 1907).

Base-6

Base-6 systems are attested on Kolopom island (formerly Frederik-Hendrik-Eiland) in southwest Indonesian Papua, as well as in the Kanum and Nambu languages in southern New Guinea around the Indonesian-Papua New Guinea border. Their origins have been discussed extensively (Donohue 2008, Evans 2009, Hammarström 2009, Plank 2009) and need not be repeated here.

In addition, there are a number of languages which have been claimed to be base-6 in the literature but which are not base-6 according to the definition used in this paper (cf. Plank 2009, Gamble 1980, Beeler 1961, Ibarra Grasso 1939b). A few require comment. One early attestation of Balanta (Northern Atlantic/Atlantic-Congo, Senegal/Guinea Bissau) has additions of 6 for the numbers 7-12 (Koelle 1854). But since we do not know the continuation beyond 12, it is unsure whether the 6:s generalize (cf. Wilson 1961a). Also, later attestations give different, non-base-6, forms (Wilson 1961b, Quintina 1961, Fudeman 1999). Similarly, Less Traditional Tiwi (Isolate, Australia) may have formed some numbers in the range 7-10 with 6 (Lee 1987:96-100), but not further.

Base-8

Northern Pame (Otopamean/Otomanguean, Mexico), the sole case of a base-8 language (attested up to 32) which does not have 4 as a sub-base is presented and discussed in Avelino (2006), though 5-8 have etymologies which involve 5.

Base-12

Dhivehi (Indo-Aryan/Indo-European, Maldives) has an early attested (Gray 1878) but long extinct base-12 which is attested up to 96 thanks to the efforts of Fritz (2002:107-123).[10] Apart from that case, there are base-12 systems in the Plateau area of northern Nigeria. The first known attestations of such systems[11] come from the famous *Polyglotta Africana* by Koelle (1854) which includes numerals 1-20 in a number of West African languages and the first proclamation of duodecimality as a system appears to be Schubert (1888). As

[10] With some speculative etymologizing, Chepang (Mahakiranti/Sino-Tibetan, Nepal) may have had 12 atoms and duodecimal counting up to 50, for a counting system associated with hunting (Caughley 1988, 1972, Hale 1973). One synopsis of Brúnkajk (Talamancan/Chibchan, Costa Rica) says that "también se cuenta por medio de docenas" (Arroyo Soto 1972:32), but it is not clear on what this statement is based. It is not corroborated by a ten or so other descriptions of Brúnkajk, and it was not normed anyway, so it does not count as a base-12 system. In a modern description of Kinikinau (Bolivia-Parana/Arawakan, Brazil) higher numbers may be expressed using (dúzias) dozens (de Carvalho Couto 2005:51), but this does not appear to be normed for exact enumeration of quantities that are not exact multiplies of twelve.

[11] However, vocabularies including monomorphemic 1-12 are listed for Hyam (there called 'Java') a few years earlier (de Castelnau 1851c:59).

shown in Table 3, we have tried to collect all independent attestations that have been published, or, unpublished but available on the internet.[12] However, not all of them are necessarily independent as this information is not always deducible from the text. It is likely that there are a few more attestations in publications that we do not have access to. For many, if not all, other sources on the same varieties attest base-10 rather than base-12, which means that the base-12 systems are currently under pressure.

[12] We wish to thank Roger Blench for help with sorting out various Plateau language identifications and classification questions.

Table 3. Published attestation of base-12 systems in the Plateau area. 12-144 means that the attestation gives forms ≤ 12, forms 12+x, multiples of 12, and a word for 144; 12+ means forms ≤ 12 and forms 12+x or multiples of 12; ≤ 12 means forms ≤ 12; "12" means that the source simply states that there was a "duodecimal system" but gives no forms; Cont.-10 means an attested 10-system contaminated by forms following a "duodecimal system" and Spec.-12 means that some duodecimal connection is speculated. Further half-attestations are as follows. Arago (base-10 in Judd 1923), Kagoma and Agatu were judged "uncertain" by Thomas (1920a). Gwara, a Margi variety (Biu-Mandara A/Afro-Asiatic, Nigeria) has monomorphemic 1-10 and forms 11-12 with formations that may include 1 and 2 – a bit like Germanic – but there is otherwise no reason to suspect base-12 counting (Wolff 1975).

Language	Source	Type	Family	Comment
Ake	Blench 2006a	≤ 12	Plateau	
Afo	Bouquiaux 1962	"12"	Plateau	
Afo (Apho)	Schubert 1888	"12"	Plateau	
Afo (extinct Afu)	Thomas 1920a	"12"	Plateau	
Afo	Meek 1925:142-143	12+	Plateau	
Afo (Eloyi)	Mackay 1964, Armstrong 1983	12+	Plateau	
Aten	Blench 2006d	≤ 12	Plateau	
Aten (Ganawuri)	Bouquiaux 1964, 1962	12-144	Plateau	
Aten (Ganawuri)	Meek 1925:142-143	12+	Plateau	
Birom	Bouquiaux 1970	12-144	Plateau	
Birom	Thomas 1920b	"12"	Plateau	
Birom (Tahoss)	Blench 2006g	≤ 12	Plateau	
Che (Rukuba)	Gerhardt 1987	Spec.-12	Plateau	Cites BCCWL.
Che (Rukuba)	Blench et al. 2006	≤ 12	Plateau	
Eggon	Blench and Hepburn 2006	≤ 12	Plateau	
Eggon	Gerhardt 1983:47	"12"	Plateau	
Eggon	Gerhardt 1987	"12"	Plateau	Cites Gospel 1935 + Lukas 1952 fieldnotes
Eggon	Shimizu 1975	"12"	Plateau	
Hyam	de Castelnau 1851c:59	≤ 12	Plateau	
Hyam (Jaba-Kwoi)	Meek 1931:123	12-144	Plateau	Also base-10 forms
Hyam (Jaba)	Bouquiaux 1962	"12"	Plateau	
Hyam	Thomas 1920b	≤ 12	Plateau	
Hyam	Blench 2006f	≤ 12	Plateau	
Ikulu	Seitz 1993:37-38	Spec.-12	Plateau	
Izere (Fobur)	Blench and Kaze 2006	≤ 12	Plateau	
Izere (Ganang)	Blench 2006c	≤ 12	Plateau	
Izere (Zarek-Gana)	Gerhardt 1987	"12"	Plateau	Citing BCCWL
Kaningkom	Gerhardt 1987	"12"	Plateau	
Koro	Thomas 1920b	12+	Plateau	
Koro	Williamson 1973:453	12+	Plateau	
Koro (Idū)	Blench 2009a	12+	Plateau	
Koro (Nyankpa)	Thomas 1920b, Gerhardt 2005, Blench 2009b	12+	Plateau	
Koro (Tinɔr)	Gerhardt 7273	"12"	Plateau	
Koro (Tinɔr)	Blench 2009c	≤ 12	Plateau	
Lungu	Gerhardt 1987	"12"	Plateau	
Mada	Blench and Kato 2006	≤ 12	Plateau	
Mada	Thomas 1920a	"12"	Plateau	
Mada (S. Mada)	Mathews 1917	12-144	Plateau	
Ninkyop	Blench 2006e	≤ 12	Plateau	
Ninzam	Mathews 1917	12-144	Plateau	
Ninzam	Thomas 1920a	"12"	Plateau	
Nungu	Mathews 1917	12-144	Plateau	
Nungu	Thomas 1920a	"12"	Plateau	
Rigwe	Bouquiaux 1962	"12"	Plateau	
Rigwe	Gerhardt 1987	"12"	Plateau	
Rigwe	Gerhardt 1969:125ff	≤ 12	Plateau	
Teria (Cara)	Blench 2006b	≤ 12	Plateau	
Teria/Fachara	Meek 1925:142-143	12+	Plateau	
Tesu	Blench 2006f,h	≤ 12	Plateau	
Tyap (Gworok)	Adwiraah 1989	"12"	Plateau	
Tyap (Gworok)	Gerhardt 1987	≤ 12	Plateau	Not confirmed in Gerhardt 1968
Amo	Luzio 1973	Cont.-10	E. Kainji	
Gure	Meek 1931:203	≤ 12	E. Kainji	
Iguta	Shimizu 1979	12+	E. Kainji	
Janji	Meek 1931:185-187	≤ 12	E. Kainji	

Continued on next page

Language	Source	Type	Family	Comment
Janji	Shimizu 1979	≤ 12	E. Kainji	
Janji	Bouquiaux 1962	"12"	E. Kainji	
Jere	Shimizu 1982	≤ 12	E. Kainji	
Jere (Boze, Akwęrę clan)	Nengel 2004, 1999	≤ 12	E. Kainji	
Kahugu	Meek 1931:212	≤ 12	E. Kainji	
Lemoro	Shimizu 1979	≤ 12	E. Kainji	Not Cokobo
Piti	Meek 1931:139	12+	E. Kainji	Switched to base-10
Piti	Matsushita 1998	"12"	E. Kainji	
Rop	Meek 1925:142-143	12+	E. Kainji	
Sanga	Shimizu 1979	≤ 12	E. Kainji	
Dyarim	Blench 2007	Spec.-12	W. Chadic	Etymological Connection
Gwandara	Shimizu 1975	"12"	W. Chadic	Citing P. Newman p.c
Gwandara (Nimbia)	Matsushita 1998	12-144	W. Chadic	
Mwaghvul	Jungraithmayr 1963	12+	W. Chadic	
Ron von Daffo	Seibert 1998	12+	W. Chadic	Not confirmed in Jungraithmayr 1970
Mumuye	Matsushita 1998	"12"	Adamawa	Not Zing Mumuye pace Shimizu 1983
Mama (Kantana)	Gerhardt 1987	"12"	Jarawan Bantu	
Mama	Thomas 1927	≤ 12	Jarawan Bantu	
Mama	Mathews 1917	12-144	Jarawan Bantu	
Mama	Thomas 1920a	"12"	Jarawan Bantu	

The base-12 systems occur only in languages in the area of Jos plateau of Nigeria, but which belong to different (sub-)families, namely Plateau (Atlantic-Congo), East Kainji (Atlantic-Congo), West Chadic (Afro-Asiatic), Adamawa (Atlantic-Congo) and Jarawan Bantu (Atlantic-Congo). A root resembling #sok for 12, with plausible sound correspondences (Gerhardt reconstructs *suak), is widespread in Plateau, wherefore it is very likely that base-12 is old in Plateau. The same root occurs in Jarawan Bantu and Ron of Daffo, both of which are isolated instance of this root, or indeed base-12, in their respective families, so borrowing from (proto-southwest) Plateau is highly likely (if not certain, as concluded by Maddieson and Williamson 1975:136 and Gerhardt 1997:140-141 for Jarawan Bantu). In East Kainji and the Beromic subgroup of Plateau, a root #kuri occurs for 12, which makes a borrowing in either direction likely. Furthermore, #piri is 12 in Gure and Kahugu (East Kainji) and #zowa is 12 in Ake and Koro (Plateau) and yet other roots for 12 appear in the remaining West Chadic cases. Since base-12 is so rare in the languages of the world, the variety of non-ancient roots suggest that a base-12 *system* may be borrowed even without key morphemes. The root for 12 in the alleged Mumuye variety with base-12 allegation is not known.

There are no obvious clues as to the unusual choice of 12 as a base. A few of the base-12 languages in Meek (1931) have hand gestures that often are used accompanying the spoken expression. A combination of fingers and eyes make up 12 in at least one of these cases, but no traces of words meaning eye, hand or finger can be found in the corresponding spoken expressions. On the other hand, although not a base, 12 bears a special position in several modern European languages too, with a special word like 'dozen' and an elevated frequency (Dehaene and Mehler 1992). The reason(s) for this is not well-understood either.

Base-15

There appears to be only one case of a language attested as base-15, at least for a number of decades, namely Huli (East New Guinea Highlands/Trans New Guinea, Papua New Guinea) of the southern highland fringes. It is clearly an original body-tally system with a cycle of 29 – midway/centerpoint is thus

15 – which under influence from a Tok Pisin base-system turned into base-15 (Cheetham 1978, Lomas 1988).

Rare Second Bases

Some rarities in the next higher bases after 5, 10 or 20 are as follows:

10-40: Pech (Paya/Chibchan, Honduras) as of Conzemius (1928:264-265) and Hawaiian (Oceanic/Austronesian, USA) until it restructured to 10-100 under foreign pressure (von Chamisso 1837, Dwight 1848, Hughes 1982).

5-20-40: Southwestern Pomo (Pomoan, USA) in one attestation (Closs 1986:35-41).

10-60: Attested (Drabbe 1952) in Ekagi (Paniai Lakes/Trans New Guinea, Indonesia) and Ntomba (Bantu/Atlantic-Congo, DRC) until it restructured to 10-100 under foreign pressure (Gilliard 1928, 1924).

5-10-20-60: Famously known from the long extinct Sumerian (Isolate, Iraq), see, e.g. Powell (1972).

(5-)10-20-(60/)80: Attested in Mande (Monteil 1905, Dombrowski and Dombrowski 1991, Delafosse 1928, Hartner 1943), Dogon (Calame-Griaule 1968), Gur (Carlson 1994, Welmers 1950:167-169) and Bangi Me (Blench 2005) languages in a relatively small area in West Africa, wherefore an areal connection is almost certain. In the Mande attestations, the systems vary between 60 and 80 as per a certain root that sometimes means 60 and sometimes 80.

5-25: Gumatj (Anindilyakwa, Australia) is described, with ample examples, to be 5-25 (upto 625). However, one would not usually use exact numbers for counting this high in this language and there is a certain likelihood that the system was extended this high only at the time of elicitation with one single speaker (Harris 1982, Sobek 2005), especially since an earlier attestation, if anything, shows a commonplace vigesimal count (Tindale 1928:128-129). At least one speaker of Biwat (Yuat River, Papua New Guinea) appears to have made the same 5-25 innovation (McElvenny 2006), as two other earlier attestations rather show a commonplace 5-20 system (Haberland and Seyfarth 1974, Mead 1932)[13]. It is remarkable that there is no[14] incontestable attestation of a 5-25 system that extends to a whole speech community. The contrast with 5-20 systems, which are ubiquitous, reveals much as to the evolution of normed number expression within a community.

[13] I wish to thank James McElvenny for access to archival material on Biwat.

[14] The extinct Saraveka has 'five hands' attested for 25 but no numerals 20-24 nor above 25 are recorded (de Créqui-Montfort and Rivet 1913). The 5-25-50 counting system in Kikongo (Bantu/Atlantic-Congo, DRC) referred to in Schmidl (1915:181) was for counting pearls only (Laman 1968, 1912, 1936).

1	ngū	6	šū̱	11	tǫ́
2	žú	7	žàadù	12	rxá
3	nīé	8	šɻ́	13	šé
4	ňū̱ú̱	9	nīà	14	rxǫ̀
5	žú̱	10	tè	15	rxǫ̀?

Table 4. The monomorphemic numerals up to 15 in Chocho of Santa Catarina Ocotlán. 15-19 are formed as 15+1 etc and 20 is a base (Veerman-Leichsenring 2000:33-34), cf. also Mock (1977:153-154).

Last Notes

At least two cases of alleged base-11 exist, both of which appear to be mistaken. Pañgwa (Bantu/Atlantic-Congo, Tanzania) is presented with a base-11 vocabulary (Johnston 1922a:477), but this cannot be corroborated in other attestations (Stirnimann 1983) so it is presumably an error. A fairly early discussion of Maori (Balbi 1826:256-257) likewise claims undecimality, but this was refuted already in the same century (Conant 1896:122-123). One alleged case of counting in 30:s is in Klingenheben (1927:43) but this too has failed to be corroborated later.

3.2 Other Rarities

Other than base, there are a few very interesting rarities which we mention below.

Streak of unanalyzable forms

Several, but not all, of the base-12 languages have monomorphemic words for all of 1-12 as does, e.g., Chalchihuitán Tzotzil (Mayan, Mexico) (Hopkins 1967:16). However, the record streak appears to be 15[15], as evidenced in Chocho of Santa Catarina Ocotlán (Popolocan/Oto-Manguean, Mexico) in Table 4. A claim of monomorphemic 1-20 in Munda (subfamily of Austroasiatic, India) appears, on closer scrutiny, to be artificial or unsubstantiated[16].

[15] I wish to thank Thomas Hanke for bringing this case to my attention.

[16] Sharma (2003:63) claims that

> We may say Munda speakers are the earliest known people who practised this system of counting which had monomorphemic units of counting upto twenty.

but gives no source and no forms. Monomorphemic 1-20 forms cannot be found in the monograph on Munda numerals by Zide (1978) nor in any published description of Kharia or any other Munda language we have been able to consult. Nevertheless, a recent unpublished description of Kharia (Peterson 2006:138-139), a set of monomorphemic 11-19 are recorded as alternative forms alongside a set of composite forms. Peterson notes, however, that the monomorphemic forms were given to him by youths who all confirmed that they had be taught them in school (and themselves used Sadani loans for the numbers in question). Further inquiries by Peterson with experienced local teachers also point towards an "artificial" origin of the 11-19 forms (p.c. John Peterson 2008).

Order of Additive Units

As we have seen, all languages which have numerals above 20 form the higher numbers using addition and multiplication of integers (and occasionally subtraction as well multiplication with fractions). Both addition and subtraction are commutative operations so languages are free to change the order of the operands. Not surprisingly, the order or multiplier and multiplicand is usually the same the order of numeral and noun in the language in question. For additive units the situation is more interesting. For expressions where the sum is less than, say, a 100, we find both smaller-precedes-larger and larger-precedes-smaller in the languages of the world. A lot of languages have one order for the teens and the opposite order for higher sums. For sums above 100, the situation is quite different. Almost all languages, and a multitude of the cases must be independent, show larger-precedes-smaller order. At least three ancient languages[17] – Classicial Attic Greek, Classical Arabic, Sanskrit (as well as Vedic) – are attested with both orders possible. The only modern languages with invariable smaller-bigger order between additive units in numeral expressions ≥ 100 appear to be (certain dialects of) Malagasy (Barito/Austronesian, Madagascar), Chuj (Mayan, Guatemala) and Tzotzil (Mayan, Mexico)[18], see Daval-Markussen et al. (2008) for references.

Cardinal Dominance?

In natural languages, it appears that cardinal numerals hold a primary position over other kinds of numerals, e.g., distributive numerals, and exact number marking in general, in the sense that the non-cardinals are morphosyntactically derived from the cardinals and that the cardinals run higher. The dominance appears to be exceptionless for all languages which have numerals above 3, but we will review two intereresting challenges below.

One description of a Great Andamanese variety explicitly says that there are more ordinals than cardinals (Man 1883a:100), or to be more specific, that there are only 2 cardinals but 6 ordinals. But a closer inspection of the forms reveals that the six "ordinals" are not true ordinals. 3-6 do not mean third-sixth but in the middle, the next one, last and so on. They only acquire the fixed ordinal meaning in the context of a game or the like when the number of patricipants is known (Man 1883b:413).

One description (Mathews 1904) of Wuddyāwūrru (West Victoria/Pama-Nyungan, Australia) says that there are more grammatical numbers (singular, plural, *trial*, and plural) than cardinals (one, two). This is not contradicted by other sources on the same or related languages (too many too list). However, there is no linguistic data in this case to ascertain that the trial was a true trial (rather than a paucal) and Mathews has described many other Australian langages as having trials where this is questionable (p.c. Barry Blake 2005). We

[17] A modern example may be the recently innovated Palikúr (North Arawak/Arawak, Brazil-Guyana) numeral system, but it is not fully clear what the norms are (Green 1994, Launey 2003).

[18] I wish to thank Aymeric Daval Rasmussen for bringing the Mayan cases to my attention.

will never know for sure whether this language had a true trial or not, since the language is extinct.

4 Conclusion

This paper has surveyed rarities for a number of structural properties of numeral systems. We have given full primacy to data presentation rather than interpretation to make the factual status of the data maximally clear. With this, we hope to have set the stage for future generalizations and interpretations of rareness with a high level of empirical validity.

Acknowledgements

I wish to thank the following libraries for granting access and services: Centralbiblioteket (Gothenburg), Institutionen för orientaliska och afrikanska språk (Gothenburg), Etnografiska Muséet (Göteborg), LAI (Göteborg), Carolina Rediviva (Uppsala), NAI (Uppsala), Karin Boye (Uppsala), KB (Stockholm), SUB (Stockholm), LAI (Stockholm), Universiteitsbibliotheek (Leiden), KITLV (Leiden), Universiteitsbibliotheek (Amsterdam), Institute for Asian and African Studies (Helsinki), MPI-EVA (Leipzig), Universitätsbibliothek (Leipzig), Butler/Columbia University (New York City), IfA (Cologne), BNF (Paris), SOAS (London), ILPGA (Paris), ZAS (Zürich) and Völkerkundliche Bibliothek (Frankfurt). I am also indebted to (in no particular order) Hein van der Voort, Lincoln Almir Amarante Ribeiro, Eduardo Rivail Ribeiro, Michael Cysouw, Nathan Hill, Jesús Mario Girón, Karsten Legère, Helene Fatima Idris, Bernard Comrie, Pedro Viegas Barros, Lionel M. Bender, John Kalespi, Hilário de Sousa, Frank Seidel, Tom Güldemann, Lourens de Vries, Ian Tupper, Johanna Fenton, Randy Lebold, Willem Adelaar, Lyle Campbell, Norbert Cyffer, Maarten Mous, Thilo Schadeberg, Raoul Zamponi, Paul Whitehouse, Swintha Danielsen, Lauren Campbell, Dmitry Idiatov, Nick Evans, Matthew Dryer, Mark Donohue, Roger Blench and Peter Bakker for help with access to data. The bibliographies by Alain Fabre (for South America) and Jouni Filip Maho (for Africa) have been very helpful in bibliographical searching leading up to this study.

References

Adam, Lucien & V. Henry. 1880. *Arte y Vocabulario de la Lengua Chiquita con algunos textos traducidos y explicados compuestos sobre manuscritos inéditos del XVIII* (Bibliothèque Linguistique Américaine VI). Paris: Librairie-Éditeur J. Maisonneuve.

Adelaar, Willem F. H. 2004. *The Languages of the Andes* (Cambridge Language Surveys). Cambridge University Press.

Adwiraah, Eleonore. 1989. *Grammatik des Gworok (Kagoro): Phonologie, Tonologie, Morphologie und Textanalyse* (Europäische Hochschulschriften: Reihe XXI: Linguistik 71). Frankfurt am Main: Peter Lang.

Aikhenvald, Alexandra & R. M. W. Dixon. 1999. Other Small Families and Isolates. In R. M. W. Dixon & A. Aikhenvald (eds.), *The Amazonian Languages* (Cambridge Language Surveys), 341-383. Cambridge University Press.

Aikhenvald, Alexandra Y. 2008. *The Manambu language of East Sepik, Papua New Guinea.* Oxford University Press.

Anonby, Stan & Sandy Anonby. 2007. A Report on Three Arauan Speech Varieties (Jamamadi, Jarawara, and Banawá) of the Amazon. SIL International, Dallas. SIL Electronic Survey Reports 2007-022 `http://www.sil.org/silesr/abstract.asp?ref=2007-022`.

Armstrong, Robert G. 1983. The Idomoid Languages of the Benue and Cross River Valleys. *Journal of West African Languages* XIII(1). 91–147.

Arroyo Soto, Víctor Manuel. 1972. *Lenguas indígenas costarricenses.* 2nd edn. San José, Costa Rica: .

Asangama, N. 1983. Le Budu: langue bantu du nord-est du Zaïre: Esquisse Phonologique et grammaticale. Paris: Université de la Sorbonne doctoral dissertation.

Avelino, H. 2006. The typology of Pame number systems and the limits of Mesoamerica as a linguistic area. *Linguistic Typology* 10(1). 493–513.

Balbi, Adrien. 1826. *Discours préliminaire et introduction* (Atlas Ethnographique du Globe I). Paris: Rey et Gravier. Also titled "Introduction à l'Atlas Ethnographique du Globe".

Barriga Puente, Francisco. 1998. *Los Sistemas de Numeración Indoamericanos: un enfoque areotipológico* (Colección Lingüística Índigena 7). México: Universidad Nacional Autoónoma de México.

Becher, Hans. 1960. *Die Surára und Pakidái: Zwei Yanonámi-Stämme in Nordwestbrasilien, mit anhang über die Sprache der Surára und Pakidái von Aryon D. Rodrigues* (Mitteilungen aus dem Museum für Völkerkunde in Hamburg XXVI). Hamburg: Kommissionsverlag Cram, De Gruyter & Co.

Beeler, M. S. 1961. Senary Counting in California Penutian. *Anthropological Linguistics* 3(6). 1–8.

Beeler, Madison S. 1963. Ventureño Numerals. In W. Bright (ed.), *Studies in Californian Linguistics* (University of California Publications in Linguistics 34), 13-18. Berkeley and Los Angeles: University of California Press.

Beeler, Madison S. 1967. *The Ventureño Confesario of José Señán, O.F.M.* (University of California Publications in Linguistics 47). Berkeley: University of California Press.

Bernatzik, Hugo Adolf. 1942. *De Gula Bladens Andar: Forskningsresor i Bortre Indien.* Stockholm: Bokförlaget Natur och Kultur.

Betts, LaVera. 1981. *Dicionário parintintín-português português-parintintín.* Brasília: Summer Institute of Linguistics.

Blench, Roger. 2005. Baŋgi Me, a language of unknown affiliation in Northern Mali and its affinities. Draft Manuscript March 18, 2005.

Blench, Roger. 2006a. The Ake language of Central Nigeria and its affinities. Draft Manuscript January 2, 2006.

Blench, Roger. 2006b. The Cara language of Central Nigeria and its affinities. Draft Manuscript January 3, 2006.

Blench, Roger. 2006c. The Ganang Language of Central Nigeria and its Affinities. Draft Manuscript January 3, 2006.

Blench, Roger. 2006d. Iten-English Dictionary. Draft Manuscript January 2, 2006.

Blench, Roger. 2006e. The Ninkyop Language of Central Nigeria and its Affinities. Draft Manuscript January 7, 2006.

Blench, Roger. 2006f. Prospecting Proto-Plateau. Draft Manuscript January 3, 2006.

Blench, Roger. 2006g. The Tahoss Dialect of the Berom Language of Central Nigeria and its Affinities. Draft Manuscript January 3, 2006.

Blench, Roger. 2006h. The Təsu language of Central Nigeria and its affinities. Draft Manuscript September 16, 2006.

Blench, Roger. 2007. The Dyarim language of Central Nigeria and its affinities. In Henry Tourneux (ed.), *Topics in Chadic Linguistics III: Historical Studies: Papers from the 3rd Biennial International Colloquium on Chadic Languages, Villejuif, November 24-25, 2005* (Tschadistik/Linguistique Tchadique 4), 41-59. Köln: Rüdiger Köppe.

Blench, Roger. 2009a. The Idũ language of Central Nigeria: Phonology, wordlist and suggestions for orthography changes. Draft Manuscript May 20, 2009.

Blench, Roger. 2009b. The Nyankpa [= Yeskwa] language of Central Nigeria. Draft Manuscript May 20, 2009.

Blench, Roger. 2009c. The Tinɔr [= Koro Waci] language of Central Nigeria and its affinities. Draft Manuscript May 20, 2009.

Blench, Roger, Ruth Adiwu & Gideon Asukutu. 2006. The Ce [Rukuba] language of Central Nigeria and its affinities. Draft Manuscript January 4, 2006.

Blench, Roger & I. D. Hepburn. 2006. A Dictionary of Eggon. Draft Manuscript January 2, 2006.

Blench, Roger & Barau Kato. 2006. A Dictionary of Mada, a Plateau Language of Central Nigeria based on the Rija Dialect: Mada-English with an English-Mada finderlist. Draft Manuscript January 4, 2006.

Blench, Roger & Bitrus Bulus Kaze. 2006. A Dictionary of the Izere Language of Fobur. Draft Manuscript January 6, 2006.

Bokula, F.-X. 1970. La Langue Bodo: Formes Nominales. *Africana Linguistica* IV. 63–84.

Bokula, M. & L. Ngandi. 1985. Numération Cardinale dans les Langues Bantu du Haut-Zaire. *Annales Æquatoria* 6. 189–196. Also in Annales de l'I.S.P.-Kisangani 13 (July):700-707, 1984.

Borgman, Donald M. 1990. Sanuma. In Desmond C. Derbyshire & Geoffrey K. Pullum (eds.), *Handbook of Amazonian Languages* volume II, 15-248. Mouton de Gruyter.

Bouquiaux, Luc. 1962. A propos de numération: L'emploi du système décimal et du système duodécimal dans la langue Birom (Nigéria septentrional). *Africana Linguistica* I. 7–10.

Bouquiaux, Luc. 1964. A Word List of Aten (Ganawuri). *Journal of West African Languages* I(2). 5–26.

Bouquiaux, Luc. 1970. *La langue Birom (Nigeria Septentrional): phonologie, morphologie, syntaxe* (Bibliothèque de la Faculté de Philosophie et Lettres de l'Université de Liège – Fascicule CLXXXV). Paris: Société d'Édition "les belles lettres".

Bowers, Nancy & Pundia Lepi. 1975. Kaugel Valley Systems of Reckoning. *Journal of the Polynesian Society* 84. 309–324.

Bradshaw, Robert. 2007. *Fuyug grammar sketch* (Data Papers on Papua New Guinea Languages 53). Ukarumpa, Papua New Guinea: SIL-PNG Academic Publications.

Brenzinger, Matthias. 2009. Documenting concepts in contact. Paper presented at the CIPL Conference on the.

Bruno, Ana Carla. 2003. Waimiri Atroari Grammar: Some Phonological, Morphological, and Syntactic Aspects. University of Arizona doctoral dissertation.

Calame-Griaule, G. 1968. *Dictionnaire Dogon Dialecte Tørø: Langue et Civilisation* (Langues et Littératures de l'Afrique Noire IV). Paris: Librairie C. Klincksieck.

Carlson, Robert. 1994. *A Grammar of Supyire* (Mouton Grammar Library 14). Mouton de Gruyter.

Caughley, Ross C. 1972. *A vocabulary of the Chepang language.* Kirtipur: Summer Institute of Linguistics and Institute of Nepal Studies, Tribhuvan University.

Caughley, Ross C. 1988. Chepang: A Sino-Tibetan language with a duodecimal numeral base?. In David Bradley, Eugénie J. A. Henderson & Martine Mazaudon (eds.), *Prosodic analysis and Asian linguistics: To honour R. K. Sprigg* (Pacific Linguistics: Series C 104), 197-199. Canberra: Australian National University.

Cheetham, B. 1978. Counting and Number in Huli. *Papua New Guinea Journal of Education* 14. 16–27.

Clark, Charles U. 1937. Jesuit Letters to Hervás on American Languages and Customs. *Journal de la Société des Américanistes* XXIX. 97–145.

Closs, Michael P. 1986. Native American Number Systems. In Michael P. Closs (ed.), *Native American Mathematics*, 3-44. Austin: University of Texas Press.

Conant, Leonard Levi. 1896. *The Number Concept: its origin and development.* New York: MacMillan.

Conrad, Robert J., Joshua Lukas & John Alungum. 1978. Some Muhiang grammatical notes. In Richard Loving (ed.), *Miscellaneous papers on Dobu and Arapesh* (Workpapers in Papua New Guinea Languages 25), 89-130. Ukarumpa: Summer Institute of Linguistics.

Conrad, Robert J. & Kepas Wogiga. 1991. *An Outline of Bukiyip Grammar* (Pacific Linguistics: Series C 113). Canberra: Research School of Pacific and Asian Studies, Australian National University.

Conzemius, E. 1927, 1928. Los Indios Payas de Honduras Estudio geográfico, histórico, etnográfico y lingüístico. *Journal de la Société des Américanistes* XIX, XX. 245–302, 253–360.

Czekanowski, Jan. 1924. Sprachenaufnahmen. In *Forschungen im Nil-Kongo-Zwischengebiet: Zweiter Band: Ethnographie Uele/Ituri/Nil-Länder* (Wissenschaftliche Ergebnisse der deutschen Zentral-Afrika-Expedition 1907-1908: Ethnographie-Anthropologie VI:2), 575-714. Leipzig: Klinkhardt & Biermann.

Maria das Dores de Oliveira. 2006. Ofayé, a língua do povo do mel: Fonologia e Gramática. Maceió: Universidade Federal de Alagoas doctoral dissertation.

Daval-Markussen, Aymeric, Peter Bakker & Harald Hammarström. 2008. On the Origins of the Malagasy Numeral System. Submitted.

de Carvalho Couto, Valéria Guimarães . 2005. A Língua Kinikinau: Estudo do Vocabulário e conceitos Gramaticais. Três Lagoas: Universidade Federal de Mato Grosso do Sul, Câmpus de Três Lagoas doctoral dissertation.

de Castelnau, Francis. 1851a. *Histoire du Voyage* (Expédition dans les parties central de l'Amérique du Sud III). Paris: P. Bertrand.

de Castelnau, Francis. 1851b. *Histoire du Voyage* (Expédition dans les parties central de l'Amérique du Sud V). Paris: P. Bertrand.

de Castelnau, Francis. 1851c. *Renseignements sur l'Afrique Centrale et sur une Nation d'Hommes à queue qui s'y trouverait.* Paris: P. Bertrand.

de Créqui-Montfort, G. & P. Rivet. 1913. Linguistique Bolivienne: La Langue Saraveka. *Journal de la Société des Américanistes* X. 497–540.

de Matallana, B. & Cesareo de Armellada. 1943. Exploración del Paragua. *Boletín de la Sociedad Venezolana de ciencias naturales* VIII(53). 61–110.

de Vries, Lourens J. 1998. Body part tally counting and Bible translation in Papua New Guinea and Irian Jaya. *The Bible Translator (Practical Papers)* 49(4). 409–415.

Dehaene, Stanislas & Jacques Mehler. 1992. Cross-linguistic Regularities in the Frequency of Number Words. *Cognition* 43. 1–29.

Delafosse, Maurice. 1928. La numération chez les Nègres. *Africa, Journal of the International Institute of African Languages* 1(3). 387–390.

Dempwolff, Otto. n.d.. *A Grammar of the Graged Language.* Narer, Karkar Island: Lutheran Mission.

Dijkmans, Joseph J. M. 1974. *Kare-Taal: Lijst van woorden gangbaar bij het restvolk Kare opgenomen in de jaren 1927-1947.* Sankt Augustin: Anthropos-Institut - Haus Völker und Culturen.

Dixon, Roland B. & A. L. Kroeber. 1907. Numeral Systems of the Languages of California. *American Anthropologist* 9(4). 663–689.

Dixon, R. M. W. 2004. *The Jarawara Language of Southern Amazonia.* Oxford University Press.

Dombrowski, F. A. & Bruno W. W. Dombrowski. 1991. Numerals and numeral systems in the Hamito-Semitic and other language groups. In Alan S. Kaye (ed.), *Semitic Studies in Honor of Wolf Leslau* volume 1, 340-381. Wiesbaden: Otto Harrassowitz.

Donohue, Mark. 2008. Complexities with restricted numeral systems. *Linguistic Typology* 12(3). 423–429.

d'Orbigny, Alcide Dessalines. 1839. *L'homme américain (de l'Amérique Méridionale): considéré sous ses rapports physiologiques et moraux* volume 2. Paris: Pitois-Levrault et C.

Drabbe, Peter. 1952. *Spraakkunst van het Ekagi: Wisselmeren Ned. N. Guinea.* The Hague: Martinus Nijhoff.

Drabbe, Peter. 1953. *Spraakkunst van de Kamoro-Taal.* 'S-Gravenhage: Martinus Nijhoff. Koninklijk Instituut voor Taal-, Land- en Volkenkunde.

Dwight, T. 1848. Sketch of the Polynesian Language. *Transations of the American Ethnological Society* II. 223–234.

Earl, G. W. 1837. Review of Verhaal nan eene Reize naar en langs de zuidwest kust van Nieuw Guinea, gedaan in 1828, door Z. M. Corvet Triton, en Z. M. Coloniale schoener de Iris, door J. Modera, Lieut. ter Zee, van de tweede Klasse by Z. M. Corvet Triton; J. Modera. *Journal of the Royal Geographical Society of London* 7. 383–395.

Early, John D. & John F. Peters. 2000. *The Xilixana Yanomami of the Amazon: history, social structure, and population dynamics.* University Press of Florida.

Ehrenreich, Paul. 1887. Ueber die Botocudos der brasilianischen Provinzen Espiritu Santo und Minas Geraes. *Zeitschrift für Ethnologie* XIX. 1–46, 49–82.

Epps, Patience. 2006. Growing a numeral system: The historical development of numerals in an Amazonian language family. *Diachronica* 23(2). 259–288.

Evans, Nicholas. 2009. Two *pus* one makes thirteen: Senary numerals in the Morehead-Maro region. *Linguistic Typology* 13(2). 321–335.

Everett, Daniel L. 2004. The Absense of Numerals in Pirahã. Presentation at the Workshop on Numerals, MPI, Leipzig, 29-30 of March, 2004.

Everett, Daniel L. 2005. Cultural Constraints on Grammar and Cognition in Pirahã: Another Look at the Design Features of Human Language. *Current Anthropology* 46(4). 89–130.

Everett, Daniel L. & Barbara Kern. 1997. *Wari': the Pacaas Novos language of Western Brazil* (Descriptive Grammars Series). London & New York: Routledge.

Fast, Daniel, Ruby Fast & Gerhard Fast. 1996. *Diccionario Achuar-Shiwiar - Castellano* (Serie Lingüística Peruana 36). Yarinacocha: Ministerio de Educación and Instituto Lingüístico de Verano.

Fast, Gerhard & Ruby Fast. 1981. *Introducción al idioma achuar* (Documento de Trabajo 20). Lima: Ministerio de Educación and Instituto Lingüístico de Verano.

Fastre, P. 1920. Vocabulary: Name of Tribe, Fujuge (Mafulu), Name of Village, Sivu. *Commonwealth of Australia. Papua: Annual Report for the Year* 1918-19. 116–116.

Fleischmann, L. & S. Turpeinen. 1975. Bine Grammar Essentials. Ukarumpa, Papua New Guinea: Unpublished Typescript, The Summer Institute of Linguistics.

Fortune, Reo F. 1942. *Arapesh* (Publications of the American Ethnological Society XIX). New York: J. J. Augustin Publisher.

Frank, Michael C., Daniel L. Everett, Evelina Fedorenko & Edward Gibson. 2008. Number as a cognitive technology: Evidence from Pirahã language and cognition. *Cognition* 108. 819–824.

Franklin, Karl & Joyce Franklin. 1962. The Kewa Counting Systems. *Journal of the Polynesian Society* 71(2). 188–192.

Friederici, Georg. 1912. *Wissenschaftliche Ergebnisse: einer amtlichen Forschungsreise nach dem Bismarck-Archipel im Jahre 1908: II Beiträge zur Völker- und Sprachenkunde von Deutsch-Neuguinea* (Ergänzungsheft der Mitteilungen aus den Deutschen Schutzgebieten 5). Berlin: Ernst Siegfried Mittler und Sohn.

Fritz, Sonja. 2002. *The Dhivehi Language* (Beiträge zur Südasienforschung 191). Würzburg: Ergon.

Fudeman, Kirsten Anne. 1999. Topics in the Morphology and Syntax of Balanta, an atlantic language of Senegal. Cornell University doctoral dissertation.

Galis, Klaas Wilhelm. 1955. Talen en dialecten van Nederlands Nieuw-Guinea. *Tijdschrift Nieuw-Guinea* 16. 109–118, 134–145, 161–178.

Gamble, Geoffrey L. 1980. The "old time" Chunut Count. In Kathryn Klar, Margaret Langdon & Shirley Silver (eds.), *American Indian and Indoeuropean Studies: Papers in Honor of Madison S. Beeler*, 51-55. Hague: Mouton de Gruyter.

Gerhardt, Ludwig. 1967/1968. Analytische und Vergleichende Untersuchungen zu einigen zentralnigerianischen Klassensprachen: Teil 1. *Afrika und Übersee* LI. 161–198.

Gerhardt, Ludwig. 1968/1969. Analytische und Vergleichende Untersuchungen zu einigen zentralnigerianischen Klassensprachen (fortsetzung). *Afrika und Übersee* LII. 23–57, 125–143, 207–242.

Gerhardt, Ludwig. 1972/73. Abriß der nominalen Klassen im Koro, North Central State, Nigeria. *Afrika und Übersee* LVI(4). 245–266.

Gerhardt, Ludwig. 1983. The Classification of Eggon: Plateau or Benue Group?. *Journal of West African Languages* XIII(1). 37–50.

Gerhardt, Ludwig. 1987. Some Remarks on the Numerical Systems of the Plateau Languages. *Afrika und Übersee* 70. 19–29.

Gerhardt, L. 1997. Die Jarawan-Bantu-Sprachen und ihr linguistisches und kulturelles Umfeld. *Frankfurter Afrikanistische Blätter* 9. 129–146.

Gerhardt, Ludwig. 2005. Some notes on Yeskwa (North-Western Plateau, Nigeria) with comments on Koelle's Polyglotta Africana. *Hamburger Afrikanistische Arbeitspapiere* 3. 35–52.

Gilliard, L. 1924. La numérotation des Ntomba, riverains du Lac Léopold II. *Congo* 5(II(3)). 374–378.

Gilliard, L. 1928. *Grammaire synthétique de Lontomba: suivie d'un vocabulaire* (Bibliothèque-Congo 20). Bruxelles: Éditions de l'Essorial.

Gnerre, Mauricio Covaz. 1986. Some notes on quantification and numerals in an Amazon Indian language. In Michael P. Closs (ed.), *Native American Mathematics*, 71-92. Austin: University of Texas Press.

Gómez, Gale Goodwin. 1990. The Shiriana Dialect of Yanam (Northern Brazil). Columbia University doctoral dissertation.

Gordon, Peter. 2004. Numeral Cognition Without Words: Evidence from Amazonia. *Science* 306. 496–499.

Gray, Albert. 1878. The Maldive Islands with a vocabulary taken from Farancois Pyrard de Laval, 1602-1607. *Journal of the Royal Asiatic Society of Great Britain and Ireland, N.S.* 10. 173–209.

Green, Diana. 1994. O Sistema Numérico da Língua Palikúr. *Boletim do Museu Paraense Emílio Goeldi, Série Antropologia* 10(2). 261–303.

Green, Diana. 1997. Diferenças entre termos numéricos em algumas línguas indígenas do Brasil. *Boletim do Museu Paraense Emílio Goeldi, Série Antropologia* 13(2). 179–207.

Grondona, Verónica M. 1998. A Grammar of Mocovi. University of Pittsburgh doctoral dissertation.

Gualdieri, C. B. 1998. Mocoví (Guaycurú): Fonologia e Morfossintaxe. Universidade Estadual de Campinas doctoral dissertation.

Haberland, Eike & Siegfried Seyfarth. 1974. *Die Yimar am Oberen Korowori (Neuguinea)* (Studien zur Kulturkunde 36). Wiesbaden: Franz Steiner.

Hafford, James A. 1999. Elements of Wuvulu Grammar. University of Texas at Arlington masters thesis.

Hale, Austin. 1973. *Clause, sentence, and discourse patterns in selected languages of Nepal 4: Word lists* volume 40:4. Summer Institute of Linguistics: Publications in Linguistics. Chepang numerals are displaces one column to the left starting with '39'.

Hammarström, Harald. 2009. Whence the Kanum Base-6 Numeral System?. *Linguistic Typology* 13(2). 305–319.

Hanke, Wanda. 1950. Vocabulário e idioma mura dos índios mura do rio Manicoré. *Arquivos* 12. 3–8. Arquivos is published at Manaus.

Hanke, Wanda. 1952. *O idioma Mura* (Documentação do Amazonas 1). Manaus.

Hanke, Wanda. 1956. Beobachtungen über den Stamm der huari (Rio Corumbiara) Brasilien. *Archiv für Völkerkunde* 11. 67–82.

Hanke, Wanda. 1964. Verstreute Indianerdörfer im Südosten des Mato Grosso. In *Völkerkundliche Forschungen in Südamerika* (Kulturgeschichtliche Forschungen 11), 9-33. Braunschweig: Albert Limbach.

Harries, Lyndon. 1959. Nyali, a Bantoid Language (Belgian Congo). *Kongo-Overzee* XXV. 174–205.

Harris, John. 1982. Facts and Fallacies of Aboriginal Number Systems. In Susanne Hargrave (ed.), *Language and Culture*, 153-182. Darwin: Summer Institute of Linguistics.

Hartner, W. 1943. Zahlen und Zahlensysteme bei Primitiv- und Hochkulturvölkern. *Paideuma* 6/7. 286–326.

Hervás y Panduro, Lorenzo. 1786. *Aritmetica Delle Nazioni e Divisione del fra L'Orientali* (Idea dell'Universo XIX). Cesena: Gregorio Biasini.

Hinton, Leanne. 1994. 10: California Counting. In *Flutes of Fire: Essays on California Indian Languages*, 113-121. Berkeley, California: Heyday Books.

Hopkins, Nicholas A. 1967. A Short Sketch of Chalchihuitán Tzotzil. *Anthropological Linguistics* 9(4). 9–25.

Howitt, Alfred William. 1889. Notes on Australian Message Sticks and Messengers. *Journal of the Royal Anthropological Institute of Great Britain and Ireland* XVIII. 314–332.

Howitt, Alfred William. 1904. *The Native Tribes of South-East Australia.* New York: MacMillan.

Hughes, Barnabas. 1982. Hawaiian Number Systems. *The Mathematics Teacher* 75(3). 253–256.

Hughes, Barnabas B. 1974. The earliest known record of California Indian numbers. *Historia Mathematica* 1. 79–82.

Ibarra Grasso, Dick Edgar. 1938. Las Numeraciones Indígenas Americanas. *Boletín de la Academia Argentina de Letras* VI(23-24). 397–417.

Ibarra Grasso, Dick Edgar. 1939a. Las Numeraciones Cuaternarias. *Boletín de la Academia Argentina de Letras* VII(28). 585–606.

Ibarra Grasso, Dick Edgar. 1939b. Las Numeraciones Senario-decimales en Sudamérica. *Boletín de la Academia Argentina de Letras* VII(25-26). 187–213.

Ibriszimow, D. 1988. Some Remarks on Chadic Numerals. In W. J. G. Möhlig (ed.), *Afrikanistische Beiträge zum XXIV. Deutschen Orientalistentag 26.–30. September 1988* (Afrikanistische Arbeitspapiere: Sondernummer), 67-74. Köln: Universität zu Köln.

Johnston, Harry. 1904. *The Uganda Protectorate* volume 2. 2nd edn. London: Hutchinson.

Johnston, Harry H. 1919, 1922b. *A Comparative Study of the Bantu and Semi-Bantu Languages.* Oxford University Press.

Johnston, Harry H. 1922a. Chapter XIII: The Bantu and Semi-Bantu Numerals. In *A Comparative Study of the Bantu and Semi-Bantu Languages* volume II, 463-482. Oxford University Press.

Judd, A. S. 1923. Notes on the Language of the Arago or Alago Tribe of Nigeria. *Journal of the African Society* XXIII(LXXXIX). 30–38.

Jungraithmayr, Herrmann. 1963. Die Sprache der Sura (Maghavul) in Nordnigerien. *Afrika und Übersee* XLVII. 8–89, 204–220.

Jungraithmayr, Herrmann. 1970. *Die Ron-Sprachen: Taschadohamitische Studien in Nordnigerien* (Afrikanistische Forschungen III). Glückstadt: J. J. Augustin.

Kalunga Mwela-Ubi, Marcel. 1999. Le Numéral en Bantu: Considérations Typologiques. Université de Lubumbashi doctoral dissertation.

Klingenheben, August. 1926-1927. Zu den Zählmethoden in den Berbersprachen. *Zeitschrift für Eingeborenensprachen* XVII. 40–51.

Knobloch, Franz. 1967. *Die Aharaibu-Indianer in Nordwest-Brasilien* (Collectanea Instituti Anthropos 1). St. Augustin bei Bonn: Anthropos Institut.

Koch-Grünberg, Theodor. 1903. Die Guaikurú-Gruppe. *Mitteilungen der Anthropologischen Gesellschaft in Wien* XXXIII. 1–128.

Koch-Grünberg, Theodor. 1906. Makú. *Anthropos* I. 877–906.

Koch-Grünberg, Theodor. 1928. *Sprachen* (Von Roroima zum Orinoco: Ergebnisse einer Reise in Nordbrasilien und Venezuela in den Jahren 1911-13 4). Stuttgart: Strecker und Schröder.

Koelle, Sigismund W. 1854. *Polyglotta Africana or Comparative Vocabulary of Nearly Three Hundred Words and Phrases in more than One Hundred Distinct African Languages.* London: Church Missionary House.

Kolia, J. A. 1975. A Balawaia grammar sketch and vocabulary. In T. E. Dutton (ed.), *Studies in languages of central and south-east Papua* (Pacific Linguistics: Series C 29), 107-226. Canberra: Research School of Pacific and Asian Studies, Australian National University.

Kroeber, A. L. 1925. Yuki: Culture. In *Handbook of the Indians of California* (Bulletin of the Bureau of American Ethnology 78), 169-181. St. Clair Shores, Michigan: Scholarly Press.

Kutsch Lojenga, Constance. 1994. *Ngiti: A Central-Sudanic Language of Zaire* (Nilo-Saharan: Linguistic Analyses and Documentation 9). Köln: Rüdiger Köppe.

Lafone Quevedo, Samuel A. 1890, 1890, 1891, 1891, 1891, 1892, 1893. Notas ó sea principios de gramática mocoví segun ellos se desprenden de los trabajos de Tavolini, Dobrizhoffer, Barcena y otros. *Revista del Museo de La Plata* 1, 1, 2, 2, 2, 3, 4. 111–144, 305–328, 241–272, 289–352, 393–424, 129–167, 257–287.

Lafone Quevedo, Samuel A. 1892. Introducción al "Arte Mocoví" del Padre Tavolini: Estudio de Gramática Comparada. *Revista Museo de la Plata* 4. 369–432.

Laman, Karl Edvard. 1912. *Grammar of the Kongo Language (Kikongo).* New York: The Christian Alliance Publishing Co.

Laman, Karl Edvard. 1936. *Dictionnaire kikongo-français avec une étude phonétique décivant les dialectes les plus importants de la langue dite kikongo* (Mémoires de l'Institut Royal colonial Belge: Section des sciences morales et politiques: Collection in-8o). Bruxelles: Institut Royal colonial Belge.

Laman, Karl Edvard. 1953-1968. *The Kongo* (Studia Ethnographica Upsaliensia). Uppsala. 4 vols.

Lancy, David F. & Andrew J. Strathern. 1981. "Making Twos": Pairing as an Alternative to the Taxonomic Mode of Representation. *American Anthropologist, N.S.* 83(4). 773–795.

Launey, M. 2003. *Awna Parikwaki: Introduction à la langue Palikur de Guyane et de l'Amapá* (Didactiques). Paris: IRD.

Laycock, Donald C. 1965. *The Ndu language family (Sepik District, New Guinea)* (Linguistic Circle of Canberra Publications: Series C, Books 1). Canberra: Australian National University.

Laycock, Donald C. 1970. Eliciting Basic Vocabulary in New Guinea. In Stephen A. Wurm & Donald C. Laycock (eds.), *Pacific linguistic studies in honour of Arthur Capell* (Pacific Linguistics: Series C 13), 1127-1176. Canberra: Research School of Pacific and Asian Studies, Australian National University.

Lean, Glendon A. 1986a. *Enga, Western Highlands, Simbu* (Counting Systems of Papua New Guinea 9). Port Moresby: Papua New Guinea University of Technology. Draft Edition.

Lean, Glendon A. 1986b. *Sandaun Bay Province* (Counting Systems of Papua New Guinea 7). Port Moresby: Papua New Guinea University of Technology. Draft Edition.

Lean, Glendon A. 1986c. *Southern Highlands* (Counting Systems of Papua New Guinea 10). Port Moresby: Papua New Guinea University of Technology. Draft Edition.

Lean, Glendon A. 1986d. *Western Province* (Counting Systems of Papua New Guinea 12). Port Moresby: Papua New Guinea University of Technology. Draft Edition.

Lean, Glendon A. 1992. Counting Systems of Papua New Guinea and Oceania. Papua New Guinea University of Technology doctoral dissertation.

Lee, Jennifer. 1987. *Tiwi Today: A Study of Language Change in a Contact Situation* (Pacific Linguistics: Series C 96). Canberra: Research School of Pacific and Asian Studies, Australian National University.

Lehmann, Walther. 1920. *Zentral-Amerika: Die Sprachen Zentral-Amerikas* volume I. Berlin: Dietrich Reimer.

Lojenga, Constance Kutsch. 1994. Kibudu: A Bantu Language with nine Vowels. *Africana Linguistica* XI. 127–133.

Lomas, Gabe. 1988. The Huli language of Papua New Guinea. Macquarie University doctoral dissertation.

Loukotka, Čestmír. 1955. Les Indiens Botocudo et leur Langue. *Lingua Posnaniensis* V. 112–135.

Loukotka, Čestmír. 1963. Documents et vocabulaires inédits de langues et de dialectes sud-américains. *Journal de la Société des Américanistes* LII. 7–60.

Luzio, Aldo Di. 1972/1973. Preliminary Description of the Amo Language. *Afrika und Übersee* LVI. 3–61.

Machoni de Cerdeña, Antonio. 1877 [1732]. *Arte y vocabulario de la lengua lule o tonocoté.* Buenos Aires: Coni.

Mackay, Hugh D. 1964. A Word List of Eloyi. *Journal of West African Languages* I(1, 2). 5–12, 60.

Maddieson, I. & Kay Williamson. 1975. Jarawan Bantu. *African Languages/Langues Africaines* I. 124–163.

Mamet, Ingo. 2005. *Die Ventureño-Chumash-Sprache (Südkalifornien) in den Aufzeichnungen John Peabody Harringtons* (Europäische Hochschulschriften: Reihe 19, Volkskunde/Ethnologie: Abteilung B: Ethnologie 67). Frankfurt am Main: Peter Lang.

Man, E. H. 1883a. On the Aboriginal Inhabitants of the Andaman Islands (Part I.). *Journal of the Royal Anthropological Institute of Great Britain and Ireland* 12. 69–116.

Man, E. H. 1883b. On the Aboriginal Inhabitants of the Andaman Islands (Part III.). *Journal of the Royal Anthropological Institute of Great Britain and Ireland* 12. 327–434.

Mathews, H. F. 1917. Notes on the Nungu Tribe, Nassawara Province, Northern Nigeria, and the Neighboring Tribes which use the Duodecimal System of Numeration. *Harvard African Studies* I. 83–93.

Mathews, R. H. 1904. Language of the Wuddyāwūrru Tribe, Victoria. *Zeitschrift für Ethnologie* 36. 729–734.

Matsushita, Shuji. 1998. Decimal vs. Duodecimal: An interaction between two Systems of Numeration. Paper presented at the 2nd Meeting of the AFLANG, October 1998, Tokyo.

Mattei-Müller, Marie-Claude. 2007. *Lengua y Cultura Yanomamɨ: Diccionario ilustrado Yanomamɨ-Español/Español-Yanomamɨ.* Caracas: Consejo Nacional de la Cultura.

Mayntzhusen, F. C. 1919-1920. Die Sprache der Guayaki. *Zeitschrift für Eingeborenensprachen* X. 2–22.

Mazaudon, Martine. 2007. Number building in Tibeto-Burman languages. Paper presented at the North-East India Languages Symposium 2 (NEILS-2), Gauhati (Assam, India), 5-9 February 2007.

McElvenny, James. 2006. Draft sketch grammar of Biwat (Mudukumo/Mundugumor), a Yuat language of East Sepik Province, PNG. Available on request from the author `www.pfed.info/james` accessed 13 May 2009.

McGregor, William B. 2004. *The Languages of the Kimberley, Western Australia.* London & New York: Routledge.

Mead, Margaret. 1932. Mundugumor Linguistics. Mss typed up for Donald C. Layock 1973.

Meek, Charles K. 1925. *The Northern Tribes of Nigeria.* Oxford University Press.

Meek, Charles K. 1931. *Tribal Studies in Northern Nigeria* volume 2. London: Kegan Paul, Trench, Trübner.

Menanti, Jackie. 2005. Laporan Sosiolinguistik Bahasa Viid di Kampung Senggi, Kabupaten Keerom, Papua - Indonesia. To appear in the SIL Electronic Survey Reports.

Migliazza, Ernest. 1972. Yanomama grammar and intelligibility. Indiana University doctoral dissertation.

Migliazza, E. C. 1978. Maku, Sape and Uruak Languages: Current Status and Basic Lexicon. *Anthropological Linguistics* XX(3). 133–140.

Mock, Carol. 1977. *Chocho: Santa Catarina Ocotlán, Oaxaca* (Archivo de Lenguas Indígenas de México 4). Mexico: Centro de Investigación para la Integración Social.

Montaño Aragón, Mario. 1989. *Tribus de la Selva II* (Guía etnográfica lingüística de Bolivia). La Paz, Bolivia: Don Bosco.

Monteil, Charles. 1905. Considérations générales sur le nombre et la numération chez les Mande. *L'Anthropologie* 16. 485–502.

Moolenburgh, P. E. 1904. Extract uit een verslag der Noord Nieuw-Guinea expeditie. *Tijdschrift voor Indische Taal-, Land- en Volkenkunde (TBG)* 47. 168–188, 381–385.

Münzel, M. 1969-1972. Notas preliminares sobre os Kabori (Makú entre o río negro e o japura). *Revista de Antropologiía* 17-20. 137–181.

Nekitel, Otto. 1985. Sociolinguistic Aspects of Abu', a Papuan Language of the Sepik Area, Papua New Guinea. Canberra: Australian National University doctoral dissertation.

Nengel, J. G. 1999. *Precolonial African Intergroup Relations in the Kauru and Pengana Polities of Central Nigerian Highlands 1800-1900* (European University Studies: Series III 814). Berlin: Peter Lang.

Nengel, J. G. ca 2004. Endangerment and Survival Prospects of the ɛBoze Language (Northern Jos Plateau of Central Nigeria). Manuscript, University of Jos.

Nimuendajú, Curt. 1924. Os Indios Parintintin do Alto Madeira. *Journal de la Société des Américanistes* XVI. 201–278.

Nimuendajú, Curt. 1932. Wortlisten aus Amazonien. *Journal de la Société des Américanistes* XXIV. 93–119.

Nimuendajú, Curt & E. H. do Valle Bentes. 1923. Documents sur quelques langues peu connues de l'Amazone. *Journal de la Société des Américanistes* XV. 215–222.

Nordenskiöld, Erland. 1911. *Indianer och Hvita i Nordöstra Bolivia.* Stockholm: Bonniers.

Nordenskiöld, Erland. n.d.. Anteckningar om sydamerikanska infödingsspråk: Bolivia- och Chacostammar. Gothenburg: Manuscript B5858, Världskulturmuséet.

Parker, H. 1909. *Ancient Ceylon: An account of the aborigines and of part of the early civilisation.* New Delhi: Marwah Publications.

Parkinson, R. 1907. *Dreißig Jahre in der Südsee.* Stuttgart: Strecker & Schröder.

Peters, John F. 1998. *Life among the Yanomami: the story of change among the Xilixana on the Mucajai River in Brazil.* Peterborough, Ontario: Broadview.

Peterson, John. 2006. *Kharia: A South Munda Language.* Habilitationsschrift, Universität Osnabrück.

Pica, Pierre, Cathy Lemer, Véronique Izard & Stanislas Dehaene. 2004. Exact and Approximate Arithmetic in an Amazonian Indigene Group. *Science* 306. 499–503.

Pires, Nádia N. 1992. Estudo da gramática da língua Jeoromitxi (Jabuti). Universidade Estadual de Campinas masters thesis.

Plank, Frans. 2009. Senary summary so far. *Linguistic Typology* 13(2). 337–345.

Popky, Donna. 1999. Oro Win: A Descriptive and Comparative Outlook of an Endangered Language. University of Pittsburgh masters thesis.

Powell, Marvin A. Jr. 1972. The Origin of the Sexagesimal System: The Interaction of Language and Writing. *Visible Language* 6(1). 5–18.

Pumuge, Hilary Manda. 1975. The Counting System of the Pekai-Alue Tribe of the Topopul Village in the Ialibu Sub-district in the Southern Highlands District, Papua New Guinea. *Science in New Guinea* 3(1). 19–25.

Quintina, Fernando R. 1961. Conhencimento da língua Balanta. *Boletim Cultural da Guiné Portuguesa* XVI(64). 737–768.

Ramirez, H. 1994a. *Iniciação à língua Yanomama: Dialetos do Médio Rio Catrimani e de Xitei: Curso de língua Yanomama.* Brasil: Diocese do Roraima, Boa Vista.

Ramirez, H. 1994b. Le Parler Yanomami des Xamatauteri. Aix-en-Provence: Université de Provence doctoral dissertation.

Ray, Sidney H. 1912. A Grammar of the Fuyuge Language. In Robert W. Williamson (ed.), *The Mafulu: Mountain People of British New Guinea*, 307-331, 336-344. London: MacMillian and Co.

Refsing, Kirsten. 1986. *The Ainu Language: The Morphology and Syntax of the Shizunai Dialect.* Aarhus: Aarhus University Press.

Renault, Pedro Victor. 1903. Exploração dos rios Mucury e Todos os Santos e seus affluentes – feita por ordem do governo da Provincia pelo engenheiro Pedro Victor Renault. [Coleccionada e organizada por Léon Renault]. *Revista do Archivo Publico Mineiro* VIII. 1049–1115.

Ribeiro, Michela Araújo. 2008. Dicionário Djeoromitxi-Português: registro da língua do povo Jabuti. Universidade Federal de Rondônia, Guajará-Mirim masters thesis.

Rich, Rolland. 1999. *Diccionario Arabela-Castellano* (Serie Lingüística Peruana 49). Lima: Instituto Lingüístico de Verano.

Richardson, Irvine. 1957. *Linguistic Survey of the Northern Bantu Borderland* (Linguistic Survey of the Northern Bantu Borderland 2). Oxford University Press.

Rischel, Jørgen. 1995. *Minor Mlabri: A Hunter-Gatherer Language of Northern Indochina.* Museum Tusculanum Press, University of Copenhagen.

Ross, Malcolm & John Natu Paul. 1978. *A Waskia Grammar Sketch and Vocabulary* (Pacific Linguistics: Series B 56). Canberra: Research School of Pacific and Asian Studies, Australian National University.

Sampaio, Wany Bernadete de Araujo. 1997. Estudo comparativo sincrônico entre o Parintintin (Tenharim) e o Uru-eu-uau-uau (Amondava): contribuições para uma revisão na classificação das línguas Tupi-Kawahib. Universidade Estadual de Campinas masters thesis.

Santana, Áurea Cavalcante. 2005. Transnacionalidade lingüística: a língua Chiquitano no Brasil. Goiânia: Universidade Federal de Goiás masters thesis.

Schebesta, Paul. 1934. *Vollblutneger under Halbzwerge: Forschungen unter Waldnegern und Halbpygmäen.* Salzburg-Leipzig: Anton Pustet.

Schebesta, Paul. 1966. Die Süd-Nyali oder bafuaNuma am Albertsee. *Wiener Völkerkundliche Mitteilungen* VIII. 37–54.

Schleicher, C. O. 1998. Comparative and Internal Reconstruction of the Tupi-Guarani Languages Family. Madison: University of Wisconsin doctoral dissertation.

Schmidl, Marianne. 1915. Zahl und Zählen in Afrika. *Mitteilungen der Anthropologischen Gesellschaft in Wien* XXXXV. 165–209.

Schubert, H. 1888. Das Zählen. *Dr. Neumayer's Anleitung zu Wissenschaftlichen Beobachtungen auf Reisen* II. 288–294.

Schultz, Harald. 1959. Ligeiras notas sobre os Maku do paraná Boá-Boá. *Revista do Museu Paulista, N. S.* 11. 109–132.

Seibert, Uwe. 1998. *Das Ron von Daffo (Jos-Plateau, Zentralnigeria): morphologische, syntaktische und textlinguistische Strukturen einer westtschadischen Sprache* (Europäische Hochschulschriften: Reihe XXVII: Asiatische und Afrikanische Studien 66). Frankfurt am Main: Peter Lang.

Seitz, Gitte. 1993. Ikulu – Untersuchengen zu einer zentralnigeriansichen Klassensprache. Universität Hamburg masters thesis.

Seligmann, C. G. & Brenda Z. Seligmann. 1911. *The Veddas.* Cambridge University Press.

Sharma, D. D. 2003. *Munda Sub-Stratum of Tibeto-Himalayan Languages* (Studies in Tibeto-Himalayan Languages VII). New Delhi, India: Mittal Publications.

Shimizu, Kiyoshi. 1975. A Lexicostatical Study of Plateau Languages and Jukun. *Anthropological Linguistics* 17. 413–418.

Shimizu, Kiyoshi. 1979. Five Wordlists with Analyses from the Northern Jos Group of Plateau Languages. *Afrika und Übersee* LXII. 253–271.

Shimizu, Kiyoshi. 1982. Ten More Wordlists with Analyses from the Northern Jos Group of Plateau Languages. *Afrika und Übersee* LXV. 97–134.

Shimizu, Kiyoshi. 1983. *The Zing Dialect of Mumuye: A Descriptive Grammar with a Mumuye-English Dictionary and an English-Mumuye index.* Hamburg: Helmut Buske.

Smith, Geoffrey P. 1988. Morobe Counting Systems. In *Papers in New Guinea Linguistics 26* (Pacific Linguistics: Series A 76), 1-132. Canberra: Research School of Pacific and Asian Studies, Australian National University.

Smits, L. & C. L. Voorhoeve. 1994. *The J. C. Anceaux collection of wordlists of Irian Jaya languages B: Non-Austronesian (Papuan) languages (Part I)* (Irian Jaya Source Material No. 9 Series B 3). Leiden-Jakarta: DSALCUL/IRIS.

Sobek, Vivienne. 2005. Response to Question about the Gumatj base-5 Numeral System. Personal Email Received 27 June 2005.

Stirnimann, Hans. 1983. *Praktische Grammatik der Pangwa-Sprache (SW-Tansania)/Indaki cha luchovo lwa vaPANGWA.* Schweiz: Universitätsverlag Freiburg.

Strauss, Hermann. 1962. *Die Mi-Kultur der Hagenberg-Stämme im Östlichen Zentral-Neuguinea* (Monographien zur Völkerkunde / Hamburgisches Museum für Völkerkunde III). Kommissionsverlag Cram, de Gruyter & Co., Hamburg.

Struck, Bernhard. 1910. Vokabularien der Bakondjo-, Baamba-, Bambuba-, Babira-, Balega-, Lendu- und Banyarisprachen aus dem linguistischen Nachlaß Emin-Paschas. *Mittheilungen des Seminars für Orientalische Sprachen* XIII. 133–165.

Stuhlmann, Franz. 1894. *Mit Emin Pascha ins Herz von Afrika.* Berlin: Dietrich Reimer.

Stuhlmann, F. 1916-1917. Wortlisten zentralafrikanscher Stämme. *Zeitschrift für Kolonialsprachen* VII. 257–308.

Swain, Carole. 2000. Chapter 13: Quantity and Number: Xilixana. In Mary Ritchie Key (ed.), *South American Indian Languages* (Intercontinental Dictionary Series 1). Irvine: Computer Database on CD-ROM, University of California .

Thomas, N. W. 1920a. Duodecimal Base of Numeration. *Man* 20(32). 56–60.

Thomas, N. W. 1920b. Duodecimal Base of Numeration. *Man* 20(14). 25–29.

Thomas, Northcote W. 1927. The Bantu Languages of Nigeria. In *Festschrift Meinhof*, 65-72. Hamburg: L. Friederichsen.

Tindale, Norman B. 1925, 1926, 1928. Natives of Groote Eylandt and of the West Coast of the Gulf of Carpentaria. *Records of the South Australian Museum* 3, 3, 52. 61–102, 103–134, 5–27.

Tormo, Jesús Galeote. 1993. *Manitana Auqui Besüro: Gramática Moderna de la lengua Chiquitana y Vocabulario Básico.* Santa Cruz de la Sierra, Bolivia: Los Huérfanos.

van der Voort, H. 2004. Review of The Amazonian Languages (R. M. W. Dixon and Alexandra Y. Aikhenvald, editors). *Anthropological Linguistics* 46(2). 210–215.

van der Voort, Hein. 2007. Proto-Jabutí: um primeiro passo na reconstrução da língua ancestral dos Arikapú e Djeoromitxí. *Boletim do Museu Paraense Emílio Goeldi: Ciências Humanas* 2(2). 133–168.

van Geluwe, H. 1960. *Les Bali et les Peuplades Apparentées (Ndaka-Mbo-Beke-Lika-Budu-Nyari)* (Ethnographic Survey of Africa: Central Africa: Belgian Congo: Part V). London: International African Institute.

Veerman-Leichsenring, Annette. 2000. *Gramática del Chocho de Santa Catarina Ocotlán, Oaxaca.* Netherlands: Research School of Asian, African and Amerindian Studies (CNWS), Universiteit Leiden.

Velder, Christian. 1963. Die Geister der gelben Blätter – ein Urvolk Thailands? Überblick über fünfzig Jahre Phi-Tong-Lüng Forschung. *Zeitschrift für Ethnologie* 88. 10–23.

Vellard, J. 1934, 1935. Les Indiens Guayakí. *Journal de la Société des Américanistes* XXVI, XXVII. 223–292, 175–246.

Vicedom, G. F. & H. Tischner. 1943-1948. *Die Mbowamb: Die Kultur der Hagenberg-Stämme im Östlichen Zentral-Neuguinea* (Monographien zur Völkerkunde I). Hamburg: Kommissionsverlag Cram, de Gruyter & Co. 3 vols.

Vidal, Alejandra. 2001. Pilagá Grammar (Guaykuruan Family, Argentina). University of Oregon doctoral dissertation.

Viegas Barros, José Pedro. 2004. Guaicurú no, macro-guaicurú sí. Una hipótesis sobre la clasificación de la lengua guachí (Mato Grosso do Sul, Brasil). Ms.

Villafañe, Lucrecia. 2003. Descripción de la lengua yuki. Katholieke Universiteit Nijmegen doctoral dissertation.

Vinci, Alfonso. 1956. *Samatari.* Bari, Italy: Leonardo da Vinci.

Vogt, P. 1902, 1903. Material zur Ethnographie und Sprache der Guayaki-Indianer. *Zeitschrift für Ethnologie* 34, 35. 30–45, 849–874.

von Chamisso, Adelbert. 1837. *Über die Hawaiische Sprache.* Leipzig: Weidmannischen Buchhandlung. Also published in Philosophische, philologische und historische Abhandlungen der königlichen Akademie der Wissenschaften zu Berlin aus dem Jahre 1837:1-79, 1839.

von der Gabelentz, G. & A. B. Meyer. 1882. *Beiträge zur Kenntnis der melanesischen, mikronesischen und papuanischen sprachen* (Abhandlungen der Philologisch-Historischen Klasse der Königlich-Sächsischen Gesellschaft der Wissenschaften 8(4)). Leipzig: S. Hirzel.

Weir, E. M. Helen. 1984. A Negação e outros Tópicos da Gramática Nadëb. Universidade Estadual de Campinas masters thesis.

Welmers, W. E. 1950. Notes on Two Languages in the Senufo Group II: Sup'ide. *Language* 26(4). 494–531.

Wilbert, J. 1962. Notes on a Sanema Vocabulary. *Journal de la Société des Américanistes* LI. 83–101.

Williamson, Kay. 1973. *Benue-Congo Comparative Wordlist* volume II. University of Ibadan, Nigeria: West African Linguistic Society.

Wilson, Patricia. 1989a. Ambulas-Wingei Statement. Unpublished Manuscript.

Wilson, Patricia. 1989b. Brief Ambulas-Wosera-Mamu Statement. Unpublished Manuscript.

Wilson, Patricia. 1990. Ambulas-Wosera-Kamu-K Statement. Unpublished Manuscript.

Wilson, Patricia R. 1976. Abulas dialect survey. In Richard Loving (ed.), *Surveys in five P.N.G. languages* (Workpapers in Papua New Guinea Languages 16), 51-79. Ukarumpa: Summer Institute of Linguistics.

Wilson, Patricia R. 1980. *Ambulas Grammar* (Workpapers in Papua New Guinea Languages 26). Ukarumpa, Papua New Guinea: Unpublished Typescript, The Summer Institute of Linguistics.

Wilson, W. A. A. 1961a. Numeration in the Languages of Guiné. *Africa* 31(4). 372–377.

Wilson, W. A. A. 1961b. Outline of the Balanta Language. *African Language Studies* 2. 139–168.

Wise, Mary Ruth & Eduardo Riggle. 1979. Terminología matemática y la ensenanza de conocimientos básicos entre los grupos étnicos de la Amazonía Peruana. *Lenguaje y Ciencias, Trujillo-Peru* 19(3). 85–103.

Wolf, Teodoro. 1892. *Geografía y geología del Ecuador.* Leipzig: F. A. Brockhaus.

Wolfers, Edward P. 1971. The Original Counting Systems of Papua and New Guinea. *The Arithmetic Teacher* February. 71–83.

Wolfers, Edward P. 1972. Counting and Numbers. In Peter Ryan (ed.), *Encyclopedia of Papua and New Guinea* volume 1, 216-220. Carlton: Melbourne University Press.

Wolff, E. 1974/1975. Sprachwandel und Sprachwechsel in Nordostnigeria. *Afrika und Übersee* LVIII. 187–212.

Zerries, Otto & Meinhard Schuster. 1974. *Mahekodotedi: Monographie eines Dorfes der Waika-Indianer (Yanoama) am oberen Orinoco (Venezuela)* (Ergebnisse der Frobenius-Expedition 1954/55 nach Südost-Venezuela II). München: Klaus Renner.

Zide, Norman H. 1978. *Studies in the Munda Numerals* (CIIL Occasional Monographs Series II). Central Institute of Indian Languages: Grammar Series.

Chapter VIII

The Status of the Least Documented Language Families in the World

Hammarström, H. (2009). The Status of the Least Documented Language Families in the World *Submitted.*

The Status of the Least Documented Language Families in the World

Harald Hammarström
Department of Computer Science and Engineering
Chalmers University of Technology
and University of Gothenburg
SE-412 96 Göteborg, Sweden
harald2@chalmers.se

1 Introduction

There are several legitimate reasons for pursuing language documentation, cf. Krauss (2007) for a fuller discussion. Perhaps the most important reason is for the benefit of speaker community itself – see van der Voort (2007) for some clear examples. Another reason is documentation for linguistic theory – through understanding the limits and distribution of diversity of the world's languages we can formulate and prove statements about the nature of language (Brenzinger 2007, Hyman 2003, Evans 2009, Harrison 2007). From the latter perspective, it is especially interesting to document the languages which are the most independent from the ones that are already documented, in other words, those which belong to unrelated families. We have conducted a world survey of the documentation of the language families of the world. In this paper we will list all the known least-documented language families of the world, with the aim to inform of their existence and documentational status. The ultimate hope is that this will increase their chances of being documented.

We have used the following criteria for inclusion in the present list.

1. The language family is known through at least a wordlist (i.e., languages known to exist, but for which there is no data, such as the language of 'isolados'[1], are not listed).

2. The language family, at the present knowledge, is not demonstrably related to any other known family.

3. There are no good grounds for concluding that the language is extinct (= does not have fluent speakers).

[1] See Hammarström (2007) for some listings of these.

4. All languages of the family are poorly documented, in the sense that there is less documentation than a rudimentary grammar sketch, and there is no ongoing documentation of any of its languages.

The listing is summarized in Table 1. In Section 2 we provide full bibliographic data, references, potential links to other families, endangerment status, and history of knowledge in each case. In Section 3 we clarify why a number of unclear cases are not judged to match the above criteria, and are thus dis-listed from the list in Table 1.

All the judgments as to family membership, i.e., what counts as demonstrably related and so on, were taken by the author based on inspection of all relevant sources. These judgments are more superficial than those of a potential specialist, but unavoidable since there are no active specialists for the bulk of the languages in question. Since there are frequent disagreements among scholars on such matters, the entry for each individual case[2] includes an explanation of the comparative situation and the choices taken.

Every languages and language family considered is cited with its iso-639-3 three-letter unique identifier for ease of comparison with standard reference works, e.g., the Ethnologue (Lewis 2009). The information in the present paper differs from standard reference works mainly in that the status of documentation is systematically investigated and used as the criteria for inclusion. For example, while the Ethnologue lists a certain amount of references to descriptive data, it does not aim to systematically list all (or the most extensive etc) references, so the status of documentation is not directly deducible from its listings. Other recent reference works, e.g., Brown and Ogilvie (2009), Brown (2006) do have rigid bibliographic listings but do not aim to be complete in their coverage of the language families of the world. The present listing also differs somewhat from general reference works in the genetic classification and speaker numbers. As explained above, we declare the reason for every innovative choice taken as to genetic classification, and give individual references to the relevant comparative-historical literature. For speaker numbers, we have consulted specialist literature wherever possible, and used the Ethnologue (Lewis 2009) figures in the rest of the cases or when there is reason to believe that it holds the most recent information. A final point of difference is that we mention the first appearance in the linguistic literature as an introduction to each language, in order to give a general sense of how accessible the language is and how long it has gone without documentation.

In total, 25 families are listed, all but two of them one-member families, i.e., isolates. In view of the fact that most of the world's language families are very small Hammarström (2007) – roughly half of them even isolates – it is not surprising that the set of the least documented ones is dominated by isolates.

The lack of geographical diversity among the resulting list is striking. The bulk are situated in the lowland rainforest area between the Sobger (Papua, Indonesia) and the Upper Sepik (Papua New Guinea) rivers of the island of New Guinea. A few more on the list are in other remote areas of Papua, Indonesia,

[2] We do this also for the unclear and dis-listed cases wherever relevant.

Name	Location	iso	Size	Endangerment	Documentation
Arara do Rio Branco	Brazil	axg?	Isolate?	HE	Short wordlist
Awaké	Venezuela	atx	Isolate	HE	Wordlists, some phrases
Sapé	Venezuela	spc	Isolate	HE	Wordlists, some phrases
Yurí	Colombia	cby?	Isolate	Unknown	Wordlists
Kujarge	Chad	vkj	Isolate	Unknown	200-item wordlist
Nihali	India	nll	Isolate	E	Long wordlists, a little text. Possibly more.
Shom Pen	Nicobar	sii	Isolate?	Unknown	Wordlists and phrases, but quality unclear
Afra	Papua, Indonesia	ulf	Isolate	E	ca 250 words and 15 sentences
Lepki	Papua, Indonesia	lpe	Isolate	Unknown	ca 200 words
Asaba	PNG	seo	Isolate?	NE	Short wordlist
Baiyamo	PNG	ppe	Isolate?	NE	Short wordlist
Busa	PNG	bhf	Isolate	Unknown	Short wordlist
Dem	Papua, Indonesia	dem	Isolate	NE	Wordlist and some sentences
Guriaso	PNG	grx	Isolate	HE	Short wordlist and short grammar notes
Kapauri	Papua, Indonesia	khp	Isolate	NE	ca 250 words and 15 sentences
Kimki	Papua, Indonesia	sbt	Isolate	NE	ca 250 words and 15 sentences
Mawes	Papua, Indonesia	mgk	Isolate	E	ca 250 words and 15 sentences
Mor	Papua, Indonesia	moq	Isolate?	HE	Short wordlist, possibly sentences
Murkim	Papua, Indonesia	rmh	Isolate	NE	ca 250 words and 15 sentences
Namla-Tofanma			2-family		
-Namla	Papua, Indonesia	naa		HE	ca 250 words and 6 sentences
-Tofanma	Papua, Indonesia	tlg		NE	ca 250 words and 15 sentences
Pyu	PNG	pby	Isolate	HE	Short wordlist
Sause	Papua, Indonesia	sao	Isolate	Unknown	ca 200 words
Tanahmerah	Papua, Indonesia	tcm	Isolate?	Unknown	Short wordlist
Walio			4-family		
-Pei	PNG	ppq		Unknown	Short wordlist
-Tuwari	PNG	tww		Unknown	Short wordlist
-Walio	PNG	wla		Unknown	Short wordlist
-Yawiyo	PNG	ybx		Unknown	Short wordlist
Yetfa/Biksi	Papua, Indonesia	yet	Isolate	NE	ca 250 words and 15 sentences

Table 1. HE = Highly Endangered, E = Endangered, NE = Not Presently Endangered, Unknown = No recent data (though in all cases, a good guess is that they are endangered.)

namely, Sause between the Tor and the Lakes Plain, Dem in a pocket of the highlands, Mawes on the northern coastline as well as Mor and Tanahmerah in the northeast of the Bomberai peninsula. In contrast, the listed cases from South America come from much more explored areas and no longer maintain their original territory. Kujarge, the only African case, was sighted on the Chad side in the very remote area near the border of Chad, Sudan and the Central African Republic. In India, Nihali is in an accessible location in central India, while Shom Pen requires travel to the Nicobar islands.

2 Listings

2.1 South America

Arara do Rio Branco [axg?]

First reported in Moore (2005) as a sidenote, based on an unpublished wordlist collected by Inês Hargreaves.

There is only a short wordlist for which the majority of entries are not Tupí cognates. Cognates for these words have so far not been uncovered in other families, including Arawak, but the search has been limited since the wordlist is not yet published (Souza 2008).

There is only a short unpublished wordlist, now posted on the internet (Hargreaves 2007). This wordlist was re-checked by Souza (2008).

There are were four rememberers left in 2001. In 2008, there were only two (Souza 2008). While this would normally mean that the language is practically extinct, there is great interest in language revivial with the 200 people strong ethnic group (Souza 2008).

Awaké [atx]

Awaké was first reported by Koch-Grünberg (1928) who also published a vocabulary.

The collected vocabularies bear no significant relations to neighbouring languages (Loukotka 1968, Migliazza 1980, 1983, 1985). In particular, there is no evidence for an Arutani-Sape family, and listings of such a family (Lewis 2009) is merely the result of basketting "left over" languages.

There are short vocabularies and some phrases (Koch-Grünberg 1928, de Matallana and de Armellada 1943, Migliazza 1978), a short unpublished vocabulary (Coppens 1976) and a minuscule amount of analyzed grammar (Migliazza 1980, 1983, 1985). On the other hand, there is a fairly extensive ethnographic study (Coppens 1983).

It is not clear how many competent speakers there are at present. Migliazza, with excellent knowledge of the region, counted only 5 speakers among 15 ethnic Awaké in 1964 (Migliazza 1978:135-136, cf. Migliazza 1972:20). The 2001 Venezuela census counts 29 members of the ethnic group (Amodio 2007), but it is not recorded how many of them speak Awaké (rather than Yanam). The situation on the Brazilian side appears to be similar (Lewis 2009, Fabre 2005).

Sapé [spc]

First reported by Koch-Grünberg (1928) who also published a vocabulary.

The collected vocabularies bear no significant relations to neighbouring languages (Loukotka 1968, Migliazza 1980, 1983, 1985). In particular, there is no evidence for an Arutani-Sape family, and listings of such a family (Lewis 2009) is merely the result of basketting "left over" languages.

There are short vocabularies and some phrases (Koch-Grünberg 1928, de Matallana and de Armellada 1943, Migliazza 1978), a short unpublished vocabulary (Coppens 1976) and a minuscule amount of analyzed grammar (Migliazza 1980, 1983, 1985). On the other hand, there is a fairly extensive ethnographic study (Coppens 1983).

It is not clear how many competent speakers there are at present. Migliazza, with excellent knowledge of the region, counted only 5 speakers among 25 ethnic Sapé in 1977 (Migliazza 1978:135-136), though census data has listed a number around 100 for the ethnic group (Fuchs 1967:78-79). The 2001 Venezuela census counts 6 members of the ethnic group (Amodio 2007) which conflicts with a slightly higher figure in Mosonyi (2003:109-110).

Yurí [cby?]

First reported by Wallace (1853), but no new data has been forthcoming for more than a century.

The vocabularies bear no significant resemblances to any other language in the region (Ortiz 1965, Loukotka 1968).

All materials (only vocabularies) can be found in Ortiz (1965:232-244). An additional wordlist is mentioned in Vidal y Pinell (1970) which may or may not be Yurí.

The language has not been sighted since the 19th century and therefore suspected extinct (cf. Ortiz 1965). However, there are persistent rumours of speakers or uncontacted peoples, though not precisely from the historical territory, who might be the descendendants of the century-old Yurí (Trupp 1974, Patiño Rosselli 2000, Fabre 2005, Vidal y Pinell 1970, Adelaar 2004:164, Landaburu 2000:30). Consequently, the status of endangerment is unknown – it could be extinct just as well as non-endangered. If the entry for Carabayo [cby] turns out to be Yurí, then the number of speakers, is estimated to be 150 from aerial observations (Lewis 2009).

2.2 Africa

Kujarge [vkj]

Kujarge was first reported by Doornbos and Bender (1983:59-60) with a 100-word list, and this remains the known sighting of the language[3].

The language was classified as Chadic, in the Mubi group, following 1979 personal communication (to Marvin Lionel Bender) from Paul Newman (Doornbos and Bender 1983:76)[4]. However, Paul Newman does not remember the precise details of this classification, in particular not whether it was

[3] Names similar to Kujarge occur in Lebeuf (1959:116) and MacMichael (1918:45) as well but without any language data and seemingly designating a different group or caste than Doornbos' Kujarge.

[4] Also, the remark "I am assuming that [...] Doornbos' Kujarke is Newman's Birgit, 1977:6." (Doornbos and Bender 1983:76) suggests that the classification is (partly) the result of a confused language identity. Comparing the actual language data shows that Newman's Birgit is not the same language as Doornbos' Kujarge and Newman's Birgit is definetely

based on a 100-word or 200-word list (p.c. Sep 2006). When shown the 200-word list Paul Newman sees Chadic elements in it but does not want to commit to Kujarge being a Chadic language (p.c. Sep 2006). Low numerals, pronouns etc look very un-Chadic so it is likely that the words in Kujarge shared with Chadic are loans from a Mubi group language.

The only published data is a 100-word list in Doornbos and Bender (1983) which were taken from 200-word list taken down by Paul Doornbos. The full 200-word list has been typed up by Paul Whitehouse and is available to interested linguists.

The number of speakers was estimated to 1000 in Doornbos and Bender (1983:59-60). Nothing further is known about its endangerment status.

2.3 Eurasia

Nihali [nll]

The Nahali ethnic group was first reported in 1868 (van Driem 2001:243-244) but the first language data is from Konow (1906).

It clear that Nahali has heavy overlay from neighbouring Munda, Dravidian and Indo-Aryan languages, but there is core vocabulary that remains distinct. There have been many attempts, old and new, to relate this chunk of vocabulary to other families, cf., a special issue of Mother Tongue in 1996, but so far there is no convincing case of for a relation (Shafer 1940, Kuiper 1962, 1966, van Driem 2001:242-253).

Published data on Nahali includes long wordlists, a little text and tiny amounts of grammar (Mundlay 1996, Bhattacharya 1957, Konow 1906).

Speaker numbers generally range between 1 000 and 2 000 (Lewis 2009), most bilinguals, though it is possible that there are also a few monolinguals (van Driem 2001:252). No further information on the endangerment status is known to the present author.

Shom Pen [sii]

The first linguistic data was collected in 1876 (Man 1886:432).

Shom Pen has been assumed as a Mon-Khmer language in the Nicobar group, but recently Blench (2007a) has argued convincingly that this affiliation is unjustified.

There is an old short wordlist (Man 1886) along with some data on counting (Temple 1907) as well as a more recent larger collection of words and short phrases (Chattopadhyay and Mukhopadhyay 2003), though there are question marks for the quality (Blench 2007a). There is also a very brief ethnographic study (Rizvi 1990).

Lewis (2009) gives a speaker number of 400 (from 2004) while Chattopadhyay and Mukhopadhyay (2003:3) estimate some 300 speakers and

an East Chadic language (Jungraithmayr 2004).

the general conditions indicate that the language is being transmitted to children.

2.4 Papua

Afra [ulf]

First reported (with wordlist) as Oeskoe by Galis (1956) whose information remained the only source for almost half a century.

Voorhoeve (1971) has Afra (under the name Usku) as "unclassified", by which he means that no significant lexical relations are found with its neighbours, or, in other words, a language isolate. In Voorhoeve (1975a), however, it is classified as Trans New Guinea, but no evidence or arguments were ever adduced. Hammarström (2008b) finds any link to Trans New Guinea premature.

Published wordlists are collected in Smits and Voorhoeve (1994). There is also an SIL Indonesia survey report to appear which contains 250 words and 15 sentences (Im and Lebold 2006). There is also a brief anthropological report (Dumatubun and Wanane 1989).

At present, there are about 115 speakers but the language is not immediately in danger. However, the younger generation is just as strong in Indonesian as in Afra (Im and Lebold 2006) which points to a weakeningen position of the vernacular.

Asaba [seo]

The Asaba language was probably first reported (under the name Suarmin) by Healey (1964:108).

Typological arguments are not sufficient to conclude a Leonard Schultze family with Walio (Laycock and Z'Graggen 1975). The lexical evidence does not show any conclusive genetic relationship either, be it inside or outside Leonard Schultze (Conrad and Dye 1975), with Sepik-Hill (as suggested in Lewis 2009), or with Baiyamo (as Papi) (Conrad and Lewis 1988). However, a higher figure (29%) of Baiyamo-Asabo (as Papi-Duranmin) lexicostatistical relations was quoted by Laycock and Z'Graggen (1975:753), before the later superseding lower figure (10%) of (Conrad and Lewis 1988:259), and some lexical data collected recently by anthropologists does contain matches between the two. It remains to be worked out whether these are loans or indicative of a genetic relationship.

There are some very brief notes on grammar in Laycock and Z'Graggen (1975). There are extensive anthropological studies on the people (Lohmann 2000, Little 2008).

There are about 180 speakers (Little 2008:2) and the language is still being transmitted to children (p.c. Roger Lohmann 2009).

Baiyamo [ppe]

Baiyamo was first reported by Laycock (1973) as Papi (a village name).

Typological arguments are not sufficient to conclude a Leonard Schultze family with Walio (Laycock and Z'Graggen 1975). The lexical evidence does not show any conclusive genetic relationship either, be it inside or outside Leonard Schultze (Conrad and Dye 1975), with Sepik-Hill (as suggested in Lewis 2009), or with Asaba (as Duranmin) (Conrad and Lewis 1988). However, a higher figure (29%) of Baiyamo-Asabo (as Papi-Duranmin) lexicostatistical relations was quoted by Laycock and Z'Graggen (1975:753), before the later superseding lower figure (10%) of (Conrad and Lewis 1988:259), and some lexical data collected recently by anthropologists does contain matches between the two. It remains to be worked out whether these are loans or indicative of a genetic relationship.

There is a wordlist in Conrad and Dye (1975) and some very brief notes in Laycock and Z'Graggen (1975:752-753).

Lewis (2009) cites a speaker number of 70. However, the language is still being transmitted to children (p.c. Jack Kennedy 2009).

Busa [bhf]

Busa was first reported by Loving and Bass (1964).

Busa lexicon bears no significant relations to any other language in the region (Laycock 1975a, Conrad and Dye 1975).

There is a wordlist in Conrad and Dye (1975) and some very brief notes on grammar in Laycock (1975a).

Lewis (2009) cites 240 (2000 census) speakers. In 1980, Busa was spoken by 238 people, and, though Tok Pisin usage was growing, Busa was not endangered (Graham 1981).

Guriaso [grx]

Guriaso was first reported in 1983 as a new language and named after a central village, Guriaso (Baron 1983:27). Previously, Guriaso and a few smaller villages were thought to speak (a dialect of) Kwomtari.

It was subsequently grouped with Kwomtari on very low cognate counts (3%-13%) and shared typological features (Baron 1983:27-29). In my judgment of the same data, these resemblances can just as well be explained by chance.

The only data (basic lexical and grammatical data) appears to be the 1983 unpublished SIL Survey (Baron 1983) and five numerals in Lean (1986).

The most recent data on speaker numbers is 160 (2003 SIL) as of Lewis (2009), but the language is endangered (p.c. Ian Tupper, SIL-PNG 2007).

Dem [dem]

The first data from Dem appears in Le Roux (1950) which remains the only source of information about the language.

Dem was grouped with neighboring highland languages based on lexicostastistical counts (Larson 1977), but the the cognation judgments involving Dem

are warped in that a match is judged if at least one segment matches, which obviously gives inconsistent sound correspondences. The lexicostatistic argument for relatedness is the only one offered so far, and apart from probable borrowings, I cannot find any cognates in vocabulary or morphology.

A quite extensive wordlist as well as sentences can be found in (Le Roux 1950). There is also a wordlist in (Stokhof 1983:219-221).

Lewis (2009) cites 1 000 speakers (1987 SIL) and the language is still strong in the community (p.c. Mark Donohue Aug 2008).

Kapauri [khp]

Published information (including wordlist) on Kapauri appears first in Voorhoeve (1971).

Voorhoeve (1971) grouped Kapauri with the Kaure languages based on some lexical correspondeces. However, a newer evaluation of the lexical relationships sheds doubt on a genetic relation between the Kaure languages and Kapauri (Rumaropen 2006).

Wordlists of Kapauri (as Narau) are collected in (Smits and Voorhoeve 1994). There 250 words and 15 sentences will appear in an SIL Indonesia survey report (Rumaropen 2006), which also mentions translated bible portions.

The number of native speakers is approximately 200, and the language is in good transmission to the younger generation. At present, Kapauri is not an endangered language (Rumaropen 2006).

Kimki [sbt]

A tiny 11-word list of what is probably Kimki (Hammarström 2008c) was taken up as early as 1914 (Langeler 1915) but this wordlist has lingered in the unknown. Otherwise, references to Kimki go back no earlier than to 1978 in unpublished SIL Indonesia survey mss (Silzer and Heikkinen 1984, Silzer and Heikkinen-Clouse 1991).

The language is listed as "unclassified" (Silzer and Heikkinen 1984, Silzer and Heikkinen-Clouse 1991) until between 1996 and 2000 when Grimes (2000) groups it with neighbouring Yetfa-Biksi. However, the lexical evidence is not sufficient for concluding a genetic relation between the two (Hammarström 2008b).

The only substantial data is an unpublished 250 word list and 15 sentences in an SIL survey report to appear (Rumaropen 2004).[5]

At this time, Kimki is being transmitted to children, and thus not an endangered language (Rumaropen 2004).

[5] Doriot (1991) refers to an unpublished wordlist of Kimki from Mot, but Mot is listed in survey maps as Murkim speaking (Wambaliau 2004).

Lepki [lpe]

A tiny 15-word list of what is probably Lepki (Hammarström 2008c) was taken up as early as 1914 (Langeler 1915) but this wordlist has lingered in the unknown. Otherwise, Lepki is first listed in Silzer and Heikkinen-Clouse (1991) presumably deriving from Doriot (1991).

Wherever it appears, the language is listed as "unclassified" (Lewis 2009, Silzer and Heikkinen-Clouse 1991). The label isolate is more appropriate as the wordlist shows no significant relation(s) with any of the neighbouring languages.

There is an unpublished hurriedly taken up wordlist by Donohue (no date) and a few songs plus a short wordlist contained in an unpublished anthropological report (Andersen 2007). Doriot (1991) also refers to another unpublished wordlist.

Lewis (2009) cites 530 speakers (1991 SIL). Andersen (2007) counts exactly 328 Lepki speakers and records clan membership of each one. Though there is no investigation of language shift, one may suspect that Lepki is under pressure from Ketengban and Indonesian in recently founded villages which have attracted many Lepki (Andersen 2007).

Mawes [mgk]

The language was first reported as early as (Robidé van der Aa 1879:112) but without accompanying data. Likewise, van der Leeden (1954) noted as separate identity of the language, but no actual language data surfaced.

At some point, the language was classified as a Tor language (Lewis 2009), probably on geographical/cultural grounds (cf. Voorhoeve 1975b:60) but already van der Leeden (1954) had noted its distinctness from Tor. Judging from recent and older lexical data (see below for sources), the there is little basis for classifying Mawes with Tor, nor with neighbouring languages (Wambaliau 2006).

A wordlist was published in Smits and Voorhoeve (1994) and numerals can be found in Galis (1955). 250 words and 15 sentences will appear in an SIL Indonesia survey report (Wambaliau 2006).

Though the speaker number is not low (ca 850), Mawes is under pressure from Indonesian and can be considered an endangered language (Wambaliau 2006).

Mor (of Bomberai) [moq]

As far as we are aware, the language was first reported by Anceaux (1958) and this remains the only source of information on the language.

Evidence for inclusion in Trans New Guinea is weak (Voorhoeve 1975a:431), both lexically and pronominally.

A wordlist can be found in Smits and Voorhoeve (1998) and judging from the note (note 34 p 18) there and comments in Anceaux (1958), Anceaux collected grammatical data as well. We searched the Anceaux Nachlass for these grammatical data 24 June 2008 at KITLV manuscripts Or 615, especially anvulling 4-23 but we could locate only wordlists for Bomberai Mor. There are

unpublished wordlists and some grammatical data from an SIL Indonesia survey in 1983[6], but exactly how much is not clear (p.c. Gilles Gravelle).

Lewis (2009) cites a figure of 25 speakers with Stephen Wurm 2000 given as the source. Since there is no record of Wurm having collected independent data for this region, the figure was presumably taken over from an earlier source, possibly an earlier Ethnologue edition. The present author has not had access to any other information, such as the 1983 survey data.[7]

Murkim [rmh]

Murkim was first reported in Silzer and Heikkinen-Clouse (1991).

Murkim is usually listed as an unclassified language (Silzer and Heikkinen-Clouse 1991, Lewis 2009). There are some lexical matches with neighbouring languages, but they look more like loans than the outcome of genetic inheritance (Hammarström 2008b).

The only data is 250 words and 15 sentences from an SIL survey not yet published (Wambaliau 2004).

At this time, Murkim is being transmitted to children, and thus not an endangered language (Wambaliau 2004).

Namla-Tofanma [naa,tlg]

Tofanma was first reported (with wordlist) by Galis (1956) whose information remained the only source for almost half a century. Namla was known to Galis (1956), Voorhoeve (1971) as well as Anceaux (nd) as a separate language. Since there was no published or unpublished wordlists or other language data to be found from these authors, the language failed to appear in subsequent listings. Much later, Doriot (1991) "re-discovered" the language and took up a wordlist, but since neither the wordlist or the survey was published the language still failed to appear in listings. Therefore, Namla was re-discovered "again" by chance in an SIL Indonesia survey more than a decade later (Lee 2005).

The only known data indicates that Namla and Tofanma are genetically related, because there are good matches in the basic lexicon, which are arguably not loans (Hammarström 2008b). Voorhoeve (1971) has Tofanma as "unclassified", by which he means that no significant lexical relations are found with its neighbours, or, in other words, a language isolate. In Voorhoeve (1975a), however, it is classified as Trans New Guinea, but no evidence or arguments were ever adduced. Hammarström (2008b) finds any link to Trans New Guinea premature.

Published wordlists for Tofanma are collected in Smits and Voorhoeve (1994) and there is also a 250-word list and 15 sentences in a survey report to appear (Wambaliau 2005). The only data of Namla is 250 words and 8 sentences from

[6] I wish to thank Mark Donohue for bringing my attention to this survey.

[7] However, the unpublished survey report *Roland Walker and Michael Werner 1978 Bomberai Survey Report MS, SIL Indonesia* contains no original information as the survey team did not travel to the Mor area (p.c. Roland Walker Aug 2008).

an SIL survey not yet published (Lee 2005) and an unpublished wordlist by Doriot (1991).

At present, Tofanma (251 speakers) is still being learnt by children, and so is not an endangered language (Wambaliau 2005). Namla speaker numbers are 25 speakers in Galis (1956), 27 in Anceaux notebooks (n.d.), 36 in Lee (2005). All speakers are bilingual in Tofanma, the young generation is more frequently using Indonesian and the children are not learning the Namla language fluently. Therefore, Namla is a highly endangered language.

Pyu [pby]

Pyu was first reported in the literature by Laycock (1972).

Pyu was grouped in the Kwomtari-Baibai-Pyu phylum, but no real evidence was ever presented (Laycock 1975b). There are no significant lexical links with neighbouring languages (Conrad and Dye 1975).

There are two short wordlists (Conrad and Dye 1975, Laycock 1972) and a sentence or two on grammar in Laycock (1975b:854).

Lewis (2009) cites a speaker number of 100 (2000 census) and more recent information suggests that the language is highly endangered. According to a 1992 report by Arjen Lock, the language is only spoken in the village of Biake 2 with its hamlets (north of the Sepik River and just east of the PNG - Indonesian border), together with an unlocated village on the bend of the Sepik within Indonesian territory. According to Lock's informant (who came from Biake 2), "people who are over 30 years and older are bilingual in Abau and [Pyu]. The children are claimed to lack fluency in both Abau and [Pyu]. They prefer to communicate in Tok Pisin." Although Lock's data did not come from observations in the language area, it seems very plausible that the language is highly endangered (p.c. Ian Tupper SIL-PNG Sep 2008).

Sause [sao]

Probably the first mention of Sause as a separate language is (Voorhoeve 1975b:45) based on Anceaux's collection of wordlists.

At some point, presumably on geographical grounds, the language started to be listed as a Tor language (Lewis 2009), but the lexical data available does not support this.

The only published data is a wordlist in Smits and Voorhoeve (1994). Mark Donohue has collected a minuscule unpublished wordlist from a transient speaker (p.c. Aug 2008).

Lewis (2009) cites 250 speakers. Nothing further is known to the present author about the endangerment status of Sause.

Tanahmerah (of Bomberai) [tcm]

As far as we can tell, the language was first reported by Galis (1955).

Links with Mairasi are unconvincing lexically and pronominally (Voorhoeve 1975a:424-431), at least for the given data.

There are some very scanty notes on grammar in (Voorhoeve 1975a:424-431) and numerals in Galis (1960). There are very short wordlists Galis (1955), Anceaux (1958) though a slightly longer one appears in Smits and Voorhoeve (1998) and has a note on p 18 that makes it clear that there was additional grammatical data collected by Anceaux. We searched the Anceaux Nachlass for these grammatical data 24 June 2008 at KITLV manuscripts Or 615, especially anvulling 4-23 but we could not locate anything beyond wordlists for Bomberai Tanahmerah. It is likely that Lloyd Peckham, an SIL member working with the nearby Mairasi languages, has newer data on the Tanahmerah (p.c. Mark Donohue Aug 2008). There are unpublished wordlists and some grammatical data from an SIL Indonesia survey in 1983[8], but exactly how much is not clear (p.c. Gilles Gravelle).

Lewis (2009) cites 500 (SIL 1978) speakers. The present author has not had access to any other information. The present author has not had access to any other information, such as the 1983 survey data.[9]

Walio [ppq,tww,wla,ybx]

The Walio languages were probably first reported by Healey (1964:108).

Typological arguments are not sufficient to conclude a Leonard Schultze family with Baiyamo (as Papi) (Laycock and Z'Graggen 1975:752-753) (Laycock 1973:32-33). The lexical evidence does not show any conclusive genetic relationship either, be it inside or outside Leonard Schultze (Conrad and Dye 1975, Conrad and Lewis 1988).

There are published wordlists (Conrad and Dye 1975, Conrad and Lewis 1988), numerals in Lean (1986) and some very minimal notes on grammar in Laycock and Z'Graggen (1975:752-753) and Laycock (1973:32-33).

Lewis (2009) cites speaker numbers of 50-360 emanating from the 2000 census. No further information on endangerment status is known to the present author.

Yetfa/Biksi [yet]

Biksi was first reported in the literature by Laycock (1972), who had met with transients from Papua, Indonesia [then West Irian] while doing fieldwork on the Papuan [then Australian] side in 1970. Yetfa is mentioned for the first time in the 2nd edition of the *Index of Irian Jaya languages* (Silzer and Heikkinen-Clouse 1991) as an unclassified language – without any references to data – but the information presumably derives from Doriot (1991) who trekked in parts of the Yetfa-speaking area in April-May 1991. Some time between the 14th edition of the Ethnologue (Grimes 2000) and the 15th (Gordon 2005), it was realised that Yetfa and Biksi are so close as to be regarded as one language.

[8] I wish to thank Mark Donohue for bringing my attention to this survey.

[9] However, the unpublished survey report *Roland Walker and Michael Werner 1978 Bomberai Survey Report MS, SIL Indonesia* contains no original information as the survey team did not travel to the Tanahmerah area (p.c. Roland Walker Aug 2008).

Biksi (by implication Biksi-Yetfa) was placed in the Sepik group languages by Laycock and Z'Graggen (1975:740-741) and this has often been repeated since (Lewis 2009). Biksi-Yetfa was not considered by Foley in his re-assessment of the Sepik family for lack of data (Foley 2005:126-127). The lexical matches adduced by Laycock to various Sepik are sporadic, and look more like loans or chance resemblance than the outcome of genetic inheritance (Hammarström 2008b). The lexical relations were also investigated independently by Conrad and Dye (1975:19) who found that Biksi shared no more than 4% probable cognates with any of the vicinity languages to the east, including Abelam.[10] (This lexical comparison includes numerals but no demonstratives or pronouns.) For the alleged connection with Kimki as in, e.g., Lewis (2009) see Kimki below.

Scanty notes on grammar can be found in (Laycock and Z'Graggen 1975:740-741) and short wordlists are published in Laycock (1972), Conrad and Dye (1975). An unpublished SIL Indonesia survey contains Yetfa 250-wordlists from five locations and 15 sentences (Kim 2006). There are further unpublished wordlists from several locations collected by Doriot (1991). Missionaries (SIL or UFM) in the area may have further unpublished materials from the past decade, the extent of which is not known to the present author.

At this time, Yetfa is still being transmitted to children and so is not an endangered language (Kim 2006).

3 Dis-listed and Unclear Cases

3.1 South America

Quite a lot of data, perhaps enough for a basic grammar sketch, was collected by Tastevin on Katawixi [xat] (dos Anjos 2005, Adelaar 2007), and this data makes a good case that Katawixi is related to Harakmbut, Katukina or both (Adelaar 2007).

The Máku language [iso-639-3 code lacking] isolate (Migliazza 1985) is reasonably well-documented in accessible documents (see Maciel 1991, Migliazza 1966 and references therein) and enough data for a 300-page grammar has been collected (p.c. Raoul Zamponi 2006). Also, the language is extinct as the last speaker died sometime between 2000 and 2002 (p.c. Raoul Zamponi 2006).

The isolate Taruma [iso-639-3 code lacking] was believed to be extinct but the last speakers have been located by Eithne Carlin who is actively working with the speakers to document it (p.c. Williem Adelaar 2009).

The Mako [wpc] language[11], is known only from a little more than 24 words

[10] The exaxt languages in question are Yerakai (0%), Chenapian (0%), Bahinemo (1%), Washkuk (1%), Yessan-Mayo (4%), Abelam (1%), Namie (0%), Abau (0%), May River Iwam (1%), Musan (0%), Amto (1%), Rocky Peak (0%), Ama (0%), Nimo (1%), Bo (0%), Iteri (0%), Owiniga (2%), Woswari (0%), Walio (0%), Paupe (0%), South Mianmin (0%), Nagatman (0%), Busan (1%), Pyu (1%).

[11] As Nimuendajú (1950:171-172) reminds us, there are no less than six distinct languages/ethnic groups so far referred to with the resemblant form #maku. The Mako language discussed here is the one defined by the vocabularies furnished by Vráz, Koch-

(see Loukotka (1949:56-57) and de Humboldt (1825:V7:155-157)) taken up a century or longer ago. This vocabulary shows considerable divergence from Piaroa, and does not seem close enough to be a dialect of it, but with so little data it is hard to say. More recent reports speak of a small ethnic group Mako living among the Piaroa who speak a variety mutually intelligible with Mako (Kaplan 1975, Migliazza 1985, Krute 1988). It could then be either that a) the Mako language as recorded by Vráz and Koch-Grünberg a century ago is extinct and different from what the ethnic Mako among the Piaroa speak in this century, or b) that the the Mako language as recorded by Vráz and Koch-Grünberg is the same as that of the ethnic Mako and that the perceived difference is due to imperfections of elicitation, analysis and data sparsity (cf. Fuchs 1967, Mosonyi and Mosonyi 2000).

3.2 Africa

Oropom [iso-639-3 code lacking] of Uganda (Wilson 1970) is probably not a genuine language (Fleming 1987:203), and even if so, it is both extinct and has too little non-Nilotic elements to be considered an isolate (Souag 2004).

The Mpra (= Mpre) [iso-639-3 code lacking] language in Ghana has lexical items with cognates in Atlantic-Congo as well as lexical items without plausible cognates (Goody 1963). Regardless of the genetic status of the language, the language is practically extinct (Blench 2007b).

A number of language families often subsumed under "Nilo-Saharan" have documentational status bordering that of a grammar sketch. We list the most unclear cases (only the most extensive data collections are mentioned):

Shabo [sbf]: There is a grammar sketch, though admittedly brief (Teferra 1991) and the languages is being documented by Tyler Schnoebelen (Stanford University).

Temeinan [teq,keg]: A phonological description of These (Yip 2004) as well as collections by Stevenson (Blench 2006) give the impression that enough data for a sketch has been collected.

Daju [byg,djc,daj,dau,njl]: There is also a forthcoming grammar and dictionary of Daju-Eref by Pierre Palayer (p.c. Pascal Boyeldieu 2007).

Eastern Jebel [soh,xel,zmo,tbi]: Enough data for a sketch has been collected (Stirtz 2006).

Tama [mgb,sjg,tma]: Two old, admittedly brief, sketches are available (Lukas 1938, 1933) and there are some very brief grammar notes on Miisiirii in Edgar (1989). Gerrit Dimmendaal (Cologne University) has field data on Tama from the 2000s, sufficient for a grammar sketch.

The Nuba mountains outlier language Warnang [wrn] is counted here as a divergent Heiban language (Schadeberg 1981) though this matter is not entirely

Grünberg (Loukotka 1949:56-57) and de Humboldt (1825:V7:155-157).

clear (partly due to paucity of data). The language has only been sighted once and one can suspect it is extinct or highly endangered by political turmoil in the past few decades (p.c. Thilo Schadeberg 2007).

Robin Thelwall collected a fair amount of grammatical data on Tegem/Lafofa [laf] (p.c. May 2008).

There are good reasons to believe that substantial amounts of data have been collected for the Mao [myf,gza,hoz,sze] languages (Yimam 2007), and there is an unpublished rudimentary Ganza grammar sketch (Hammarström 2008a).

Stefan Elders collected enough data on Bangeri Me [dba] for (more than) a grammar sketch (cf. Elders 2006) before he died, and this is posted online at `http://www.dogonlanguages.org/` (accessed 29 Sept 2008).

Jalaa [cet] is now presumably extinct (Kleinewillinghöfer 2001).

Smith (1897) mentions Dūmē [iso-639-3 code lacking], along with a tiny vocabulary, which is not obviously related to any of the neighbouring languages. Later surveys of the region have failed to find any trace of Dūmē and have therefore regarded the original information as suspect (Jensen 1952:57-58, Haberland and Jensen 1959). In any case, if it is genuine, Dūmē must be presumed extinct.

3.3 Eurasia

Jiamao [jio] has recently been suggested as a language isolate with heavy Hlai overlay (Norquest 2007), but the suggested residual vocabulary is not large enough in the opinion of the present author.

3.4 Papua

There is too little data on Kehu to say whether it is to be considered an isolate or related to some other language(s) as the only known wordlist contains, for example, neither numerals nor pronouns (Whitehouse nd).

The present author has had no access to data on Kembra [xkw] (Doriot 1991) and cannot vouch for its status either as isolate or related to some other language(s).

The present author has had no access to data on Yerakai [yra] (Laycock 1973) and cannot vouch for its status either as isolate or related to some other language(s).

The present author has had not access to substantial data on the Mongol-Langam [lnm,mgt,yla] languages (Laycock 1973) and cannot vouch for their status either as a small family or related to some other language(s).

The present author has had no access to substantial data on Dibiyaso [dby], Doso [dol] and Turumsa [tqm] (Tupper 2007) and cannot vouch for their status either as a stand-alone family.[12]

[12] Doso and Turumsa has 61% lexicostatistical similarity and Turumsa and Dibiyaso has 19% which, pace known caveats, indicates that the three form a small (sub-)family (p.c. Ian Tupper Sep 2008).

Foau [flh] and Diebroud [tbp] are related to the Lakes Plain languages (Clouse 1997) though not in an obvious way (p.c. Mark Donohue 2008).

The Eastern Pauwasi languages Zorop [wfg] and Emem [enr] are now understood to be genetically related to Karkar-Yuri [yuj], an insight probably due originally to Tim Usher (Whitehouse 2006). Therefore, since there are extensive materials on Karkar-Yuri (Price 1987, Price et al. 1994, Rigden no date) the family counts as reasonably well-documented. (There are some good lexical cognates that suggest that the Western Pauwasi languages Tebi [dmu] and Towei [ttn] form a bona fide family with the Eastern Pauwasi languages and Karkar-Yuri, but it is not impossible that the links are loans and that we are dealing with two small families with a lot of interaction (Hammarström 2008b)).

Mark Donohue has collected enough data for sketches of Abinomn [bsa], Bayono-Awbono [awh,byl], Powle-Ma [msl], Tanglapui [swt] (of the Kolana-Tanglapui family/subfamily), Masep [mvs], Elseng [mrf], Moraori [mok], Yoke [yki] and Damal [uhn] (p.c. Mark Donohue 2008).

Grammar sketches or other materials enough for a sketch have been collected by SIL members or others for the following language families (only the most extensive data collections are mentioned):

Pele-Ata [ata]: There is a dictionary (Hashimoto 2008) as well as an unpublished 'Ata grammar essentials' in the SIL (Ukarumpa) archives. Tatsuya Yanagida (Australian National University) is writing a PhD thesis on the language (Yanagida 2004).

Kol [kol]: There are unpublished manuscripts by Stellan and Eivor Lindrud (Akerson and Moeckel 1992) and a New Testament translation soon due for press (p.c. Stellan Lindrud 2006).

Pahoturi [kit,idi]: There is also an unpublished rudimentary (20-page) grammar sketch of Idi in the SIL archives (No Author Stated nd) and the raw data collected by Wurm in 1966 and 1970 may include enough for a grammar sketch (Wurm 1971).

Waia [knv]: There is an unpublished grammar (2004, 394 pages) in the SIL archives (p.c. Tim Schlatter 2006). Translations of the New Testament have appeared in both the Aramia river (No Author Stated 2006a) and Fly river dialects (No Author Stated 2006b).

Awin-Pa [awi,ppt]: In addition to bible text data, there is now an (otherwise unpublished) extensive grammar sketch of Kamula posted online (Routamaa 1994).

Somahai [mmb,mqf]: Martha Reimer has collected a large amount of data om Momuna, of which a little has been published (Reimer 1986).

Konda-Yahadian [knd,ner]: Enough data for a rudimentary grammar sketch has been collected by SIL Indonesia members (Berry and Berry 1987).

Porome [prm]: Martin Steer (Australian National University) is writing a PhD thesis on the language.

Yuat [bwm,buv,cga,kql,myd,mvk]: There are unpublished notes in the Mead/Fortune fieldnotes (McDowell 1991:23) and James McElvenny (Sydney University) did two months of fieldwork on Mudukumo and has written up a draft grammar sketch (p.c. James McElvenny 2008).

Tirio [aob,bmz,mcc,aup,wei]: An unpublished SIL survey from the 2000s has collected sociolinguistic, lexical and grammatical data (p.c. Ian Tupper 2007). The raw data collected by Wurm in 1966 and 1970 may include enough for a grammar sketch (Wurm 1971). (Ray 1923:360) mentions a Tirio grammar manuscript by the Reverend Riley of unknown size and location[13].

Yalë [nce]: There is an unpublished grammar sketch by SIL missionaries (Campbell and Campbell 1987).

Kayagar [aqm,kyt,tcg]: Unpublished short grammar sketches are referenced in Silzer and Heikkinen-Clouse (1991).

Arafundi [afd,afk,afp]: There are notes on Arafundi in the writings of William Foley (Foley 2006), who presumably has extensive fieldnotes. There is also an SIL survey report from 2005 (p.c. Ian Tupper 2005).

Suki [sui]: There is a New Testament translation (Bidri et al. 1981).

Gogodala [aac,ggw,wrv]: There is a New Testament translation (Partridge 1981).

Amto-Musan [amt,mmp]: Linda Krieg et al. of the New Tribes Mission is in the process of translating the bible into Siawi (= Musan) and there is so far unpublished phonemic and grammar sketch write-ups (p.c. Linda Krieg 2007).

Piawi [tmd,pnn]: There is an unpublished grammar sketch of Pinai (Melliger 2000) as well as published grammar aspects of Haruai (e.g., Comrie 1991).

4 Conclusion

We have conducted a world survey of the documentation of the language families of the world. In this paper we have listed all the known least documented language families of the world which are not yet known to be extinct. Borderline cases, unclear cases and cases for which there exists little-known data, frequently unpublished, are also listed. This we hope will be useful in setting priorities for documentational fieldwork, in particular for those documentational efforts which have understanding of linguistic diversity as an underlying goal.

[13] I could not find any further clues about the Tirio grammar manuscript in the Nachlass of Sidney H. Ray as SOAS Library (Aug 2008).

Acknowledgements

The author wishes to thank Willem Adelaar for discussion and data about Katawixi, Raoul Zamponi for discussion and data about Venezuela-Brazil border languages, Mark Donohue for sharing bits and pieces of his vast knowledge of the Papua, Indonesia language scene, Matthew Dryer for fruitful exchange about scripture materials in Papuan languages, Ian Tupper for updates on the endagerment of some upper Sepik languages, Randy Lebold for access to unpublished survey data (forming the basis for most of the Papua section), Paul Whitehouse for access to unpublished data, and to Øystein Lund Andersen for sharing his priceless anthropological report on Lepki.

References

Adelaar, Willem F. H. 2004. *The Languages of the Andes* (Cambridge Language Surveys). Cambridge University Press.

Adelaar, Willem F. H. 2007. Ensayo de clasificación del Katawixí dentro del conjunto Harakmbut-Katukina. In A. Romero-Figueroa, A. Fernández-Garay & Ángel Corbera Mori (eds.), *Lenguas indígenas de América del Sur: Estudios descriptivo-tipológicos y sus contribuciones para la lingüística teórica*, 159-169. Caracas: Universidad Católica Andres Bello.

Akerson, Paula & Bonita E. R. Moeckel. 1992. *Bibliography of the Summer Institute of Linguistics Papua New Guinea Branch 1956–1990.* Ukarumpa, Papua New Guinea: Unpublished Typescript, The Summer Institute of Linguistics.

Amodio, Emanuele. 2007. La república indígena. Pueblos indígenas y perspectivas políticas en Venezuela. *Revista Venezolana de Economía y Ciencias Sociales* 13(3). 175–188.

Anceaux, J. C. 1958. Languages of the Bomberai Peninsula. *Nieuw-Guinea Studiën* 2. 109–121.

Anceaux, J. C. n.d.. District Jafi-Jamas. KITLV Manuscripts and Archives, Leiden [Or 615 Anvulling 12].

Andersen, Øystein Lund. 2007. The Lepki People of Sogber [sic!] River, New Guinea. Unpublished.

Baron, Wietze. 1983. Kwomtari Survey. Unpublished manuscript, SIL Survey office, Ukarumpa, now posted at `http://www.kwomtari.net/kwomtari_survey.pdf` accessed 15 Dec 2008.

Berry, Keith & Christine Berry. 1987. A survey of the South Bird's Head Stock. *Workpapers in Indonesian Languages and Cultures* 4. 81–117.

Bhattacharya, S. 1957. Field-Notes on Nahāli. *Indian Linguistics* 17. 245–258.

Bidri, Midim, Ivy Lindsay & Grahame Martin. 1981. *Godte gi amkari titrum ine [Suki New Testament]*. Port Moresby: Bible Society Papua New Guinea.

Blench, Roger. 2007a. The language of the Shom Pen: a language isolate in the Nicobar islands. *Mother Tongue* XII. 179–202.

Blench, Roger. 2007b. Recovering data on Mpra [=Mpre] a possible language isolate in North-Central Ghana. Draft Manuscript March 10, 2007.

Blench, Roger M. 2006. Temein, Tese and Keiga Jirru wordlists from R. C. Stevenson's Nachlass. Manuscript.

Brenzinger, Matthias. 2007. Language Endangerment Throughout the World. In Matthias Brenzinger (ed.), *Language Diversity Endangered* (Trends in Linguistics: Studies and Monographs 181), ix-xviii. Mouton de Gruyter.

Brown, Keith (ed.). 2006. *Encyclopedia of Language and Linguistics.* 2nd edn. Amsterdam: Elsevier. 14 vols.

Brown, Keith & Sarah Ogilvie (eds.). 2009. *Concise Encyclopedia of Languages of the World.* Amsterdam: Elsevier.

Campbell, Carl & Jody Campbell. 1987. *Yade grammar essentials.* Ukarumpa: Unpublished Manuscript, Summer Institute of Linguistics.

Chattopadhyay, Subhash Chandra & Asok Kumar Mukhopadhyay. 2003. *The Language of the Shompen of Great Nicobar: a preliminary appraisal.* Kolkata: Anthropological Survey of India.

Clouse, D. A. 1997. Toward a reconstruction and reclassification of the Lakes Plain languages of Irian Jaya. In Karl J. Franklin (ed.), *Papers in Papuan linguistics No. 2* (Pacific Linguistics: Series A 85), 133-236. Canberra: Research School of Pacific and Asian Studies, Australian National University.

Comrie, Bernard. 1991. How Much Pragmatics and How Much Grammar: The Case of Haruai. In Jef Verschueren (ed.), *Pragmatics at Issue*, 81-92. Amsterdam: John Benjamins.

Conrad, Robert J. & T. Wayne Dye. 1975. Some language relationships in the Upper Sepik region of Papua New Guinea. In *Papers in New Guinea Linguistics 18* (Pacific Linguistics: Series A 40), 1-35. Canberra: Research School of Pacific and Asian Studies, Australian National University.

Conrad, Robert J. & Ronald K. Lewis. 1988. Some language and sociolinguistic relationships in the Upper Sepik region of Papua New Guinea. In *Papers in New Guinea Linguistics 26* (Pacific Linguistics: Series A 76), 243-273. Canberra: Research School of Pacific and Asian Studies, Australian National University.

Coppens, Walter. 1976. Breve Vocabulario: Sape y Arutani. (Manuscript) Fundación La Salle, Caracas.

Coppens, W. 1983. Los Uruak (Arutani). In W. Coppens (ed.), *Los Aborigenes de Venezuela, Vol II* (Monografía / Fundación la Salle 29), 407-426. Caracas: Fundación la Salle.

de Humboldt, Alexandre. 1815, 1815, 1816, 1817, 1820, 1820, 1822, 1822, 1825. *Voyage aux régions équinoxiales du Noveau Continent.* Paris: N. Maze. 9 vols.

de Matallana, B. & Cesareo de Armellada. 1943. Exploración del Paragua. *Boletín de la Sociedad Venezolana de ciencias naturales* VIII(53). 61–110.

Donohue, Mark. (no date). Lepki. Hurriedly filled in SIL-Indonesia 1998 wordlist form.

Doornbos, Paul & Lionel M. Bender. 1983. Languages of Wadai-Darfur. In Marvin Lionel Bender (ed.), *Nilo-Saharan language studies* (Monograph / Committee on Northeast African studies 13), 43-79. East Lansing: African Studies Center, Michigan State Univversity.

Doriot, Roger E. 1991. 6-2-3-4 Trek, April-May, 1991. Ms.

dos Anjos, Zoraide. 2005. Fonologia Katukina (Dialeto Katukina do Biá). Brasilia: Universidade de Brasília masters thesis.

Dumatubun, A. E. & Teddy K. Wanane. ca 1989. *Orang Usku di daerah batas Timur, Senggi, Irian Jaya.* Jayapura: s.n.

Edgar, John. 1989. *A Masalit Grammar: With Notes on other languages of Darfur and Wadai* (Sprache und Oralität in Afrika: Frankfurter Studien zur Afrikanistik 3). Berlin: Dietrich Reimer.

Elders, Stefan. 2006. Présentation du bangeri me. Atélier sur le projet dogon, vendredi 8 décembre 2006, Bamako.

Evans, Nicholas. 2009. *Dying Words: Endangered Languages and What They Have to Tell Us.* John Wiley & Sons.

Fabre, Alain. 2005. Diccionario Etnolingüístico y guía Bibliográfica de los Pueblos Indigenas Sudamericanos. Book in Progress at `http://butler.cc.tut.fi/~fabre/BookInternetVersio/Alkusivu.html` accessed May 2005.

Fleming, Harold C. 1987. Review article: Towards a definitive classification of the world's languages (review of A guide to the world's languages, by Merritt Ruhlen). *Diachronica* 4. 159–223.

Foley, William. 2006. Universal constraints and local conditions in Pidginization: Case studies from New Guinea. *Journal of Pidgin and Creole Languages* 21(1). 1–44.

Foley, William A. 2005. Linguistic prehistory in the Sepik-Ramu Basin. In Andrew Pawley, Robert Attenborough, J. Golson & R. Hide (eds.), *Papuan Pasts: Studies in the Cultural, Linguistic and Biological History of the Papuan-speaking Peoples* (Pacific Linguistics 572), 109-144. Canberra: Research School of Pacific and Asian Studies, Australian National University.

Fuchs, Helmuth. 1967. Urgent Tasks in Eastern Venezuela. *Bulletin of the International Committee on Urgent Anthropological Ethnological Research* 9. 69–103.

Galis, Klaas Wilhelm. 1955. Talen en dialecten van Nederlands Nieuw-Guinea. *Tijdschrift Nieuw-Guinea* 16. 109–118, 134–145, 161–178.

Galis, Klaas Wilhelm. 1956. *Ethnologische Survey van het Jafi-district (Onderafdeling Hollandia).* Hollandia [Jayapura]: Gouvernement van Nederlands Nieuw-Guinea, Kantoor voor Bevolkingszaken. No. 102.

Galis, Klaas Wilhelm. 1960. Telsystemen in Nederlands-Nieuw-Guinea. *Nieuw Guinea Studien* 4(2). 131–150.

Goody, J. R. 1963. Ethnological Notes on the distribution of the Guang Languages. *Journal of African Languages* 2(3). 173–189.

Gordon, Raymond G. Jr. (ed.). 2005. *Ethnologue: Languages of the World.* 15th edn. Dallas: SIL International.

Graham, Glenn H. 1981. A sociolinguistic survey of Busa and Nagatman. In Richard Loving (ed.), *Sociolinguistic surveys of Sepik languages* (Workpapers in Papua New Guinea Languages 29), 177-192. Ukarumpa: Summer Institute of Linguistics.

Grimes, Barbara F. (ed.). 2000. *Ethnologue: Languages of the World.* 14th edn. Dallas: SIL International.

Haberland, Eike & Adolf E. Jensen. 1959. *Altvölker Süd-Äthiopiens* (Völker Süd-Äthiopiens: Ergebnisse der Frobenius-Expeditionen 1950-52 und 1954-56 I). Stuttgart: Frobenius-Institut an der Johann Wolfgang Goethe-Universität (Frankfurt am Main).

Hammarström, Harald. 2007. *Handbook of Descriptive Language Knowledge: A Full-Scale Reference Guide for Typologists* (LINCOM Handbooks in Linguistics 22). München: Lincom.

Hammarström, Harald. 2008a. Notes on the Ganza Language from unpublished notes by Reidhead and Disney. Ms in preparation.

Hammarström, Harald. 2008b. A Reclassification of Some West Papua Languages. Paper Presented at the The International Workshop on Minority Languages in the Malay/Indonesian Speaking World, 28 June 2008 Leiden, The Netherlands.

Hammarström, Harald. 2008c. Two Hitherto Unnoticed Languages from Sobger River, West Papua, Indonesia. Ms.

Hargreaves, Inês. 2007. Lista de palavras transcritas por Inês Hargreaves, de dois grupos ao norte do Parque Aripuanã, RO. Manuscript made available with the help of Denny Moore.

Harrison, David K. 2007. *When Languages Die* (Oxford Studies in Sociolinguistics). Oxford University Press.

Hashimoto, Kazuo. 2008. *Ata - English dictionary with English - Ata finderlist.* Ukarumpa: Summer Institute of Linguistics.

Healey, Alan. 1964. The Ok Language Family in New Guinea. Canberra: Australian National University doctoral dissertation. Sometimes cited as *A Survey of the Ok Family of Languages* presumably because part of the thesis II-IV, which contains all linguistic data, carries this title.

Hyman, Larry M. 2003. Why describe African languages?. Keynote address at the World Congress of African Linguistics & Annual Conference of African Linguistics, Rutgers University, June 18, 2003.

Im, Youn-Shim & Randy Lebold. 2006. Draft Survey Report on the Usku Language of Papua. To appear in the SIL Electronic Survey Reports.

Jensen, Adolf E. 1952. Forschungsreise nach Süd-Abessinien. *Zeitschrift für Ethnologie* 77. 57–61.

Jungraithmayr, Herrmann. 2004. Das Birgit, eine Osttschadische Sprache – Vokabular und Grammatische Notizen. In Gábor Takács (ed.), *Egyptian and Semito-Hamitic (Afro-Asiatic) studies in memoriam W. Vycichl* (Studies in Semitic languages and linguistics 39), 342-371. E. J. Brill.

Kaplan, Joanna Overing. 1975. *The Piaroa: a people of the Orinoco basin: a study in kinship and marriage.* Oxford: Clarendon Press.

Kim, So Hyun. 2006. Draft Survey Report on the Yetfa Language of Papua, Indonesia. To appear in the SIL Electronic Survey Reports.

Kleinewillinghöfer, Ulrich. 2001. Jalaa - An Almost Forgotten Language of Northeastern Nigeria: A Language Isolate. In Derek Nurse (ed.), *Historical Language Contact in Africa* (Sprache und Geschichte in Afrika 16/17), 239-271. Köln: Rüdiger Köppe.

Koch-Grünberg, Theodor. 1928. *Sprachen* (Von Roroima zum Orinoco: Ergebnisse einer Reise in Nordbrasilien und Venezuela in den Jahren 1911-13 4). Stuttgart: Strecker und Schröder.

Konow, Sten. 1906. Nahālī. In G. A. Grierson (ed.), *Munṇḍā and Dravidian Languages* (Linguistic Survey of India IV), 185-189. Calcutta: Office of the Superintendent of Goverment Printing.

Krauss, Michael E. 2007. Mass Language Extinction and Documentation: The Race against Time. In O. Miyaoka, O. Sakiyama & M. Krauss (eds.), *Vanishing Languages of the Pacific Rim*, 3-24. Oxford University Press.

Krute, Laurence Dana. 1988. Piaroa nominal morphosemantics. Columbia University doctoral dissertation.

Kuiper, F. B. J. 1962. *Nahali: A Comparative Study* (Mededelingen der Koninklijke Akademie van Wetenschappen, afdeling letterkunde, nieuwe reeks, deel 25 nr 5). Amsterdam: Noord Hollandsche Uitgeversmatschappij.

Kuiper, F. B. J. 1966. The sources of the Nahali Vocabulary. In Norman H. Zide (ed.), *Studies in Comparative Austroasiatic Linguistics*, 57-81. Mouton de Gruyter.

Landaburu, Jon. 2000. Clasificación de la lenguas indígenas de Colombia. In María Stella González de Pérez & María Luisa Rodríguez de Montes (eds.), *Lenguas indígenas de Colombia: una visión descriptiva*, 25-50. Santafé de Bogotá: Instituto Caro y Cuervo.

Langeler, J. W. 1915. Sobger-Rivier: 'Journaal loopende van 22 October 1914 t/m 7 Januari 1915. Betreffende eene patrouille te water tot exploratie van de twee groote linker zijrivieren der Idenburg-rivier op ongeveer 140 10/4 O.L. en tot aanpeiling van het hooggebergte'. KITLV Manuscripts and Archives, Leiden [D H 1163].

Larson, Gordon F. 1977. Reclassification of Some Irian Jaya Highlands Language Families: A Lexicostatical Cross-Family Subclassification with Historical Implications. *Irian* VI(2). 3–40.

Laycock, Don. 1972. Looking Westward: Work of the Australian National University on Languages of West Irian. *Irian* 1(2). 68–77.

Laycock, D. C. 1973. *Sepik Languages: Checklist and Preliminary Classification* (Pacific Linguistics: Series B 25). Canberra: Research School of Pacific and Asian Studies, Australian National University.

Laycock, D. C. 1975a. Isolates: Sepik Region. In Stephen A. Wurm (ed.), *New Guinea Area Languages and Language Study Vol 1: Papuan Languages and the New Guinea linguistic scene* (Pacific Linguistics: Series C 38), 879-886. Canberra: Research School of Pacific and Asian Studies, Australian National University.

Laycock, D. C. 1975b. Sko, Kwomtari and Left May (Arai) Phyla. In Stephen A. Wurm (ed.), *New Guinea Area Languages and Language Study Vol 1: Papuan Languages and the New Guinea linguistic scene* (Pacific Linguistics: Series C 38), 849-858. Canberra: Research School of Pacific and Asian Studies, Australian National University.

Laycock, Donald C. & J. Z'Graggen. 1975. The Sepik-Ramu Phylum. In Stephen A. Wurm (ed.), *New Guinea Area Languages and Language Study Vol 1: Papuan Languages and the New Guinea linguistic scene* (Pacific Linguistics: Series C 38), 731-764. Canberra: Research School of Pacific and Asian Studies, Australian National University.

Le Roux, C. C. F. M. 1950. 25: Taalkundige Gegevens. In *De Bergpapoea's van Nieuw-Guinea en hun Woongebied* volume II, 776-913. Leiden: E. J. Brill.

Lean, Glendon A. 1986. *Sandaun Bay Province* (Counting Systems of Papua New Guinea 7). Port Moresby: Papua New Guinea University of Technology. Draft Edition.

Lebeuf, Annie M.-D. 1959. *Les populations du Tchad (Nord du 10e parallèle)* (Ethnographic survey of Africa, Western Africa, French series, part 8). Paris: Presses Universitaires de France for the International African Institute (IAI).

Lee, Myung Young. 2005. Draft Survey Report on the Namla Language of Papua. To appear in the SIL Electronic Survey Reports.

Lewis, Paul M. (ed.). 2009. *Ethnologue: Languages of the World.* 16th edn. Dallas: SIL International.

Little, Christopher A. J. L. 2008. Becoming an Asabano: The socialization of Asabano children, Duranmin, West Sepik Province, Papua New Guinea. Canada: Trent University masters thesis.

Lohmann, Roger Ivar. 2000. Cultural Reception in the Contact and Conversion History of the Asabano of Papua New Guinea. University of Wisconsin-Madison doctoral dissertation.

Loukotka, Čestmír. 1949. Sur Quelques Langues Inconnues de l'Amerique du Sud. *Lingua Posnaniensis* I. 53–82.

Loukotka, Čestmír. 1968. *Classification of the South American Indian Languages* (Reference Series 7). Los Angeles: Latin American Center, University of California.

Loving, Richard & Jack Bass. 1964. *Languages of the Amanab sub-district.* Port Moresby: Department of Information and Extension Services.

Lukas, Johannes. 1933. Beiträge zur Kenntnis der Sprachen voon Wadái (Marar´ēt, Mába). *Journal de la Société des Africanistes* 3(1). 25–55.

Lukas, J. 1938. Die Sprache der Sungor in Wadai. *Mitteilungen der Ausland-Hochschule an der Universität Berlin* XLI(III). 171–246. Mitteilungen der Ausland-Hochschule an der Universität Berlin is the continuation of MSOS, and this publication is filed under MSOS in various libraries.

Maciel, Iraguacema. 1991. Alguns aspectos fonológicos e morfológicos da língua Máku. Brasilia: Universidade de Brasil masters thesis.

MacMichael, H. A. 1918. Nubian Elements in Darfur. *Sudan Notes and Records* I. 30–48.

Man, E. H. 1886. A Brief Account of the Nicobar Islanders. with Special Reference to the Inland Tribe of Great Nicobar. *Journal of the Royal Anthropological Institute of Great Britain and Ireland* 15. 428–451.

McDowell, Nancy. 1991. *The Mundugumor: From the Field Notes of Margaret Mead and Reo Fortune* (Smithsonian Series in Ethnographic Inquiry). Washington, D.C.: Smithsonian Institution Press.

Melliger, Markus. 2000. Pinai-Hagahai. In John Brownie (ed.), *Sociolinguistic and literacy studies: highlands and islands* (Data papers on Papua New Guinea languages 45), 64-122. Ukarumpa: Summer Institute of Linguistics.

Migliazza, Ernest. 1972. Yanomama grammar and intelligibility. Indiana University doctoral dissertation.

Migliazza, Ernesto C. 1966. Esbôço sintático de um corpus da língua Makú. *Boletim do Museu Paraense Emílio Goeldi, Série Antropologia* 32. 1–38.

Migliazza, E. C. 1978. Maku, Sape and Uruak Languages: Current Status and Basic Lexicon. *Anthropological Linguistics* XX(3). 133–140.

Migliazza, E. C. 1980. Languages of the Orinoco-Amazon Basin: Current Status. *Antropológica* 53. 95–162.

Migliazza, E. C. 1983. Lenguas de la Región Orinoco Amazonas: Estado Actual. *América Indígena* 43. 703–784.

Migliazza, E. C. 1985. Languages of the Orinoco-Amazon Region: Current Status. In Harriet E. Manelis Klein & Louisa Stark (eds.), *South American Indian Languages: Retrospect and Prospect*, 17-139. Texas University Press.

Moore, Denny. 2005. Classificação interna da família lingüística Mondé. *Estudos Lingüísticos* XXXIV. 515–520.

Mosonyi, Esteban Emilo. 2003. Situación actual de las lenguas indígenas de Venezuela. In Esteban Emilo Mosonyi, Arelis Barbella & Silvana Caula (eds.), *Situación de las lenguas indígenas en Venezuela*, 86-116. Caracas: Casa de Las Letras-Casa de Bello.

Mosonyi, Esteban Emilo & Jorge Carlos Mosonyi. 2000. Nociones Generales Sobre Las Lenguas Indígenas en Venezuela. In Esteban Emilo Mosonyi & Jorge Carlos Mosonyi (eds.), *Manual de Lenguas Indígenas de Venezuela* (Serie Origenes), 28-115. Caracas: Fundación Bigott.

Mundlay, Asha. 1996. Nihali Lexicon. *Mother Tongue* II. 17–48.

Nimuendajú, Curt. 1950. Reconhecimento dos rios Içána, Ayarí e Uaupés. *Journal de la Société des Américanistes* 39(1). 125–182.

No Author Stated. 2006a. *Godokono Hido Tabo: Aramia River Tabo Testament.* Port Moresby: Bible Society of Papua New Guinea.

No Author Stated. 2006b. *Godokono Wade Tabo: Fly River Tabo New Testament.* Port Moresby: Bible Society of Papua New Guinea.

No Author Stated. n.d.. The Dibla:g Language. Ms, SIL, Ukarumpa.

Norquest, Peter K. 2007. A phonological reconstruction of Proto-Hlai. University of Arizona doctoral dissertation.

Ortiz, Sergio Elías. 1965. *Prehistoria Tomo 3: Lenguas y Dialectos Indígenas de Colombia* (Historia Extensa de Colombia I). Bogotá: Ediciones Lerner.

Partridge, Edna. 1981. *Sa:lenapa wala gilala dote bata ete miyana gi kanika:.* Port Moresby: Bible Society Papua New Guinea.

Patiño Rosselli, Carlos. 2000. Lenguas aborígenes de la Amazonia meridional de Colombia. In María Stella González de Pérez & María Luisa Rodríguez de Montes (eds.), *Lenguas indígenas de Colombia: una visión descriptiva*, 169-170. Santafé de Bogotá: Instituto Caro y Cuervo.

Price, Dorothy. 1987. Some Karkar-Yuri orthography and spelling decisions. In John M. Clifton (ed.), *Studies in Melanesian orthographies* (Data Papers on Papua New Guinea Languages 33), 57-76. Ukarumpa: Summer Institute of Linguistics.

Price, Dorothy, Veda Rigden & Maramia Nkonifa. 1994. *Kwaromp kwapwe kare kar (God's truly good talk) [New Testament].* USA: The Bible League, South Holland, Illinois.

Ray, Sidney H. 1923. The Languages of the Western Division of Papua. *Journal of the Royal Anthropological Institute of Great Britain and Ireland* 53. 332–360.

Reimer, Martha. 1986. *The notion of topic in Momuna narrative discourse* (Pacific Linguistics: Series A 74). Canberra: Research School of Pacific and Asian Studies, Australian National University.

Rigden, Veda. (no date). Karkar Grammar Essentials. Ukarumpa: Unpublished Manuscript, SIL.

Rizvi, S. N. H. 1990. *The Shompen: A Vanishing Tribe of the Great Nicobar Island.* Calcutta: Seagull Books.

Robidé van der Aa, Pieter Jan Baptist Carel. 1879. *Reizen naar Nederlandsch Nieuw-Guinea ondernomen op last der Regeering van Nederlandsche Indie in de jaren 1871, 1872, 1875-1876 door de Heeren P. van Crab en J.E. Teysmann, J.G. Coornengel, A.J. Langeveldt van Hemert en P. Swaan.* The Hague: Martinus Nijhoff.

Routamaa, Judy. 1994. Kamula grammar essentials. Ms. Available at http://www.sil.org/pacific/png/abstract.asp?id=50209 accessed 1 August 2008.

Rumaropen, Benny. 2004. Draft Survei Sosiolinguistik pada ragam Bahasa Kimki di Bagian Tenggara Gunung Ji, Papua, Indonesia. To appear in the SIL Electronic Survey Reports.

Rumaropen, Benny. 2006. Draft Survey Report on the Kapauri Language of Papua. To appear in the SIL Electronic Survey Reports.

Schadeberg, Thilo C. 1981. *A Survey of Kordofanian Vol 1: The Heiban Group* (Sprache und Geschichte in Afrika: Beiheft 1). Hamburg: Helmut Buske.

Shafer, Robert. 1940. Nahālī: A Linguistic Study in Paleoethnography. *Harvard Journal of Asiatic Studies* 4. 346–371.

Silzer, P. J. & H. Heikkinen. 1984. Index of Irian Jaya Languages. *Irian* XII. 1–124.

Silzer, Peter J. & Heljä Heikkinen-Clouse. 1991. *Index of Irian Jaya Languages* (Special Issue of Irian: Bulletin of Irian Jaya). 2nd edn. Jayapura: Program Kerjasama Universitas Cenderawasih and SIL.

Smith, Donaldson A. 1897. *Through Unknown African Countries: The First Expedition from Somaliland to Lake Lamu.* New York: Edward Arnold.

Smits, L. & C. L. Voorhoeve. 1994. *The J. C. Anceaux collection of wordlists of Irian Jaya languages B: Non-Austronesian (Papuan) languages (Part I)* (Irian Jaya Source Material No. 9 Series B 3). Leiden-Jakarta: DSALCUL/IRIS.

Smits, L. & C. L. Voorhoeve. 1998. *The J. C. Anceaux collection of wordlists of Irian Jaya languages B: Non-Austronesian (Papuan) languages (Part II)* (Irian Jaya Source Material No. 10 Series B 4). Leiden-Jakarta: DSALCUL/IRIS.

Souag, Lameen M. 2004. Oropom Etymological Lexicon: Exploring an extinct, unclassified Ugandan language. Paper in progress posted at http://lameen.googlepages.com/oropomed.pdf accessed 20 Sep 2008.

Souza, Larissa da Silva Lisboa. 2008. O processo de revitalização de uma língua: Mecanismos para documentação e clasificaç ao da língua dos Arara do Rio Branco. *Língua, Literatura e Ensino* 3. 555–561.

Stirtz, Timothy M. 2006. Possession of Alienable and Inalienable Nouns in Gaahmg. In Al-Amin Abu-Manga, Leoma Gilley & Anne Storch (eds.), *Insights into Nilo-Saharan Language, History and Culture: Proceedings of the 9th Nilo-Saharan Linguistic Colloquium, Institute of African and Asian Studies, University of Khartoum, 16-19 February 2004* (Nilo-Saharan 23), 377-392. Köln: Rüdiger Köppe.

Stokhof, W. A. L. (ed.). 1983. *Holle Lists: Vocabularies in Languages of Indonesia, Vol.5/2: Irian Jaya: Papuan Languages, Northern Languages, Central Highlands Languages* (Pacific Linguistics: Series D 53). Research School of Pacific and Asian Studies, The Australian National University.

Teferra, Anbessa. 1991. A Sketch of Shabo Grammar. In M. Lionel Bender (ed.), *Proceedings of the Fourth Nilo-Saharan Linguistics Colloquium* (Nilo-Saharan: Linguistics Analyses and Documentation 7), 371-387. Hamburg: Helmut Buske.

Temple, R. C. 1907. A Plan for Uniform Scientific Record of the Languages of Savages Applied to the Languages of the Andamanese and Nicobarese. *Indian Antiquary: A Journal of Oriental Research* XXXVI. 181–203, 217–251, 317–347, 353–369.

Trupp, F. 1974. Una Tribu Desconocida en la Amazonia Colombiana. *Bulletin of the International Committee on Urgent Anthropological Ethnological Research* 16. 109–110.

Tupper, Ian. 2007. Endangered Languages Listing: TURUMSA [tqm]. Document posted at `http://www.pnglanguages.org/pacific/png/show_lang_entry.asp?id=tqm` accessed 1 May 2007.

van der Leeden, Alexander Cornelis. 1954. *Verslag over taalgebieden in het Sarmische van de Ambtenaar van het Kantoor voor Bevolkingszaken* volume 35. Hollandia: Gouvernement van Nederlands-Nieuw-Guinea Dienst van Binnenlandse Zaken Kantoor voor Bevolkingszaken.

van der Voort, Hein. 2007. Theoretical and social implications of language documentation and description on the eve of destruction in Rondônia. In Peter K. Austin, Oliver Bond & David Nathan (eds.), *Proceedings of Conference on Language Documentation and Linguistic Theory*, 251-259. London: SOAS.

van Driem, George. 2001. *Languages of the Himalayas* (Handbuch der Orientalistik: Section Two: India 10). E. J. Brill. 2 Vols.

Vidal y Pinell, Ramón. 1969-1970. Identificación de la tribu de los yuríes en el Amazonas de Colombia. *Amazonía Colombiana Americanista* 7. 95–109.

Voorhoeve, C. L. 1971. Miscellaneous Notes on Languages in West Irian, New Guinea. In *Papers in New Guinea Linguistics 14* (Pacific Linguistics: Series A 28), 47-114. Canberra: Research School of Pacific and Asian Studies, Australian National University.

Voorhoeve, C. L. 1975a. The Central and Western Areas of the Trans-New Guinea Phylum: Central and Western Trans-New Guinea Phylum Languages. In Stephen A. Wurm (ed.), *New Guinea Area Languages and Language Study*

Vol 1: Papuan Languages and the New Guinea linguistic scene (Pacific Linguistics: Series C 38), 345-460. Canberra: Research School of Pacific and Asian Studies, Australian National University.

Voorhoeve, C. L. 1975b. *Languages of Irian Jaya, Checklist: preliminary classification, language maps, wordlists* (Pacific Linguistics: Series B 31). Canberra: Research School of Pacific and Asian Studies, Australian National University.

Wallace, Alfred Russell. 1853. *A narrative of travels on the Amazon and Rio Negro.* London: Reeve & Co.

Wambaliau, Theresia. 2004. Draft Laporan Survei pada Bahasa Murkim di Papua, Indonesia. To appear in the SIL Electronic Survey Reports.

Wambaliau, Theresia. 2005. Draft Laporan Survei pada Bahasa Tofanma di Papua, Indonesia. To appear in the SIL Electronic Survey Reports.

Wambaliau, Theresia. 2006. Draft Laporan Survei pada Bahasa Mawes di Papua, Indonesia. To appear in the SIL Electronic Survey Reports.

Whitehouse, Paul. 2006. The "Lost" Paper: A Belated Conference Postscript. *Mother Tongue* XI. 262–274.

Whitehouse, Paul. n.d.. Type-up of anonymous Kehu wordlist from SIL Indonesia. Ms.

Wilson, J. G. 1970. Preliminary Observations on the Oropom People of Karamoja, their Ethnic Status, Culture, and Postulated Relation to the Peoples of the Late Stone Age. *The Uganda Journal* 34(2). 125–145.

Wurm, Stephen A. 1971. Notes on the Linguistic Situation of the Trans-Fly Area. In *Papers in New Guinea Linguistics 14* (Pacific Linguistics: Series A 28), 115-172. Canberra: Research School of Pacific and Asian Studies, Australian National University.

Yanagida, Tatsuya. 2004. Socio-historic overview of the Ata language, an endangered Papuan language in New Britain, Papua New Guinea. In Shibata Norio & Toru Shionoya (eds.), *Kan minami Taiheiyoo no gengo 3 [Languages of the South Pacific Rim 3]* (ELPR Publications Series A1-008), 61-94. Suita: Faculty of Informatics, Osaka Gakuin University.

Yimam, Baye. 2007. Mao of Bambasi. In Siegbert Uhlig (ed.), *Encyclopaedia Aethiopica* volume III, 760-761. Wiesbaden: Otto Harrassowitz.

Yip, May. 2004. Phonology of the These language. *Occasional Papers in the Study of Sudanese Languages* 9. 93–117.

Can a computer extract a description of word conjugation in a natural language using only written text in the language? For example, if we present a computer with lots of English text, just by looking at frequencies, the computer can tell that words end in -ing much more frequently than some random sequence like -xpq and, further, that those words which end in -ing often occur with -s and -ed endings in place of the -ing. This thesis systematically explores the feasibility of looking at such cues to approximate human-developed inflection engines. If successful, the same technique can be used for a wide variety of different languages, and thus jumpstart them in further Natural Language Processing tasks, such as Natural Language Processing tasks such as Machine Translation, Document Categorization and Information Retrieval.

There are around 7 000 languages in the world, exhibiting a bewildering structural diversity. Many of the languages in the world today are spoken only by relatively small groups of people and are threatened by extinction. This thesis makes contributions for our understanding very little known languages by surveying which are at the highest priority be adequately recorded and by surveying the particular language feature of numeral systems, where one can find such rarities as base-6-36 systems.

UNIVERSITY OF
GOTHENBURG

ISBN 978-91-628-7942-6