# Developing an annotated corpus for Gĩkũyũ using language-independent machine learning techniques

**Peter W. Wagacha*, Guy De Pauw† and Katherine W. Getao***

*School of Computing & Informatics
University of Nairobi - Box 30197-00100, Nairobi, Kenya
{waiganjo,kgetao}@uonbi.ac.ke

†CNTS - Language Technology Group
University of Antwerp - Universiteitsplein 1, 2610 Antwerpen, Belgium
guy.depauw@ua.ac.be

**Abstract**

Networking the development of computational resources for African languages can be greatly advanced if researchers aim to develop tools that are to a large extent language-independent and therefore reusable for other languages. In this paper we describe a particular case study, namely the development of an annotated corpus of Gĩkũyũ, using language-independent machine learning techniques. The general aim of our work on Gĩkũyũ is two-fold: on the one hand we wish to digitally preserve this resource-scarce language, while on the other hand it serves as a feasibility study of using language-independent machine learning techniques for linguistic annotation of corpora. To this end we investigate established annotation induction techniques like unsupervised learning and knowledge transfer. These methods can provide interesting perspectives for the linguistic description of many other resource-scarce languages.

## 1. Introduction

Many languages in Africa are resource-scarce. This means that their computerization in this digital world is nearly impossible for the moment. The languages on this continent are as diverse as its people. Many researchers are addressing this though at various levels and stages. Networking the disparate nodes of research on the continent working on African languages is therefore crucial. Through a network, the distributed knowledge and expertise, tools, corpus building strategies can be shared. This unique situation requires that researchers should aim to develop tools that are to a large extent language-independent and therefore reusable. Such a network of researchers should be instrumental in enhancing inter-disciplinary collaboration and synergy between linguists and computer scientists, who traditionally do not always collaborate.

The general aim of our work on Gĩkũyũ is two-fold: on the one hand we simply wish to preserve this resource-scarce language, which is decreasingly being used in verbal and written communication. An annotated corpus for this language can consequently serve to stimulate interest in the language from researchers from linguistics as well as computer science. On the other hand the development of the Gĩkũyũ corpus acts a case study that researches the feasibility of using the aforementioned language-independent machine learning techniques for linguistic processing of text corpora.

To this end we investigate established annotation induction techniques like unsupervised learning and the relatively novel approach of knowledge transfer. These methods aim to minimize human annotator cost and language expert knowledge during corpus construction and can provide interesting perspectives for the linguistic description of many other RSLs. Through the use of these techniques the process of developing an annotated corpus can be significantly expedited.

## 2. The Gĩkũyũ language

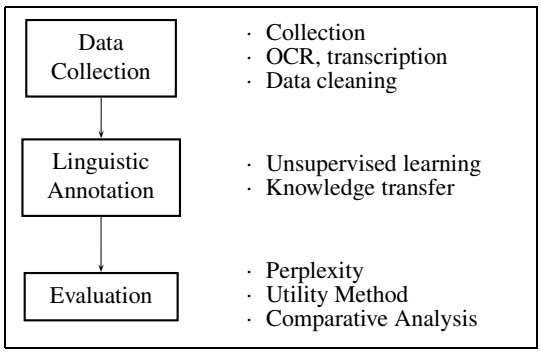
Gĩkũyũ is a language spoken by the Gĩkũyũ people (often written as Kĩkũyũ) who are approximated at 6 million people. Most of these live in the central part of Kenya in East Africa. It is a Bantu language and has been classified as E51 (Guthrie, 1967). The Bantu linguistic group covers the area from South Cameroon to almost the whole of Southern Africa, including Eastern and Central Africa. Gĩkũyũ belongs to the Kamba-Kikuyu subgroup of Bantu. There are six dialects, namely Southern Gĩkũyũ (Kiambu, Southern Murang'a), Ndia (Southern Kirinyaga), Gichugu (Northern Kirinyaga), Mathira (Karatina), and Northern Gĩkũyũ (Northern Murang'a, Nyeri). The Gĩkũyũ language is lexically similar to some closely related languages such as Chuka, Embu, Kamba and Meru.

### 2.1. Socio-linguistic challenges of Gĩkũyũ

Gĩkũyũ nowadays is more spoken than read. The primary reason for this is the fact that Kenya has two official languages, English and Kiswahili, which have become the default languages of communication both written and spoken. In school, English is the language of instruction. This is indeed a good thing in a country that has about 40 local languages. These two languages unify all and serve as a standard communication medium.

A few decades back, basic primary education was taught in the mother tongue in the rural areas. This is not the case today. Typically for the younger generation, one speaks his/her mother tongue at home (or Kiswahili), Kiswahili to the general public and English in school. This has resulted in a generation that is neither fluent in any language. Moreover, there is an emergent language, Sheng - initially a distortion of Kiswahili and English, but now a murky concoction of languages. It is evident that good native speakers of the mother tongue are lacking, and there is a marked decrease in young language users. There is nevertheless a

Figure 1: Corpus-Building Strategy




great interest in local languages within the country as has been clearly noted from the high number of vernacular radio broadcasts and some vernacular periodic publications. This work is motivated by the need to develop resources and tools that can be used in the language's computerization, as well as its preservation.

It is a fact that language, culture and knowledge are entwined. African languages contain a significant part of the world's historical, cultural, social, botanical, ... knowledge and wisdom, which needs to be preserved. Unlike in South Africa where there is in place a language policy that enshrines local languages in the constitution, this is not the case in Kenya, for its 40 odd languages, indicating a possible lack of political will to preserve local languages.

### 2.2. Gĩkũyũ as a resource-scarce language

We have adopted the following definition for a resource-scarce language (RSL): *a language for which few digital resources exist; a language with limited financial, political, and legal resources; and a language with very few linguistics experts.* A language that has very few digital resources, poses unique challenges with regard to computerization. With these resources, it would be easy to develop technologies and tools to support a language such as machine translation, speech recognition and text-to-speech systems.

Limited political and financial resources is another aspect that surrounds many local languages. This largely contributes to the lack of significant quantities of language resources. To our knowledge, there are very few trained linguists for many of our local languages. It is interesting to note that, apart from Mugane (1997), there seems to be no other recent and easily available work describing this language. Most of the literature about Gĩkũyũ was developed between 1900 and 1960. This may indicate a heightened lack of interest in Gĩkũyũ within linguistics circles. This is puzzling, since the number of speakers of this language is still relatively high. Academic work on the languages is also not readily available if it exists. Moreover, over time a language changes, necessitating further research. This is indeed the case for Gĩkũyũ.

## 3. Corpus Construction

We define three phases during the construction of our annotated Gĩkũyũ corpus, illustrated in Figure 1. First of all we have a data collection and clean-up phase, during which we gather as much raw data as possible. Next, we annotate the collected data, where we initially concentrate on providing part-of-speech tag information for the words in the corpus. Finally, we evaluate the corpus using quantitative and qualitative measures.

### 3.1. Data collection and clean-up

For Gĩkũyũ today, there are few resources available. As in many African languages, the Holy Bible is the default language reference for Gĩkũyũ. Most available material is in print media, such as religious literature, hymn books, poems, basic language learning aides, short stories, novels by renown authors e.g. Ngũgũ wa Thiong'o, etc. There is hardly any ready material in electronic form.

In our data collection phase, we located available source documents, which were run through a scanner and later an optical character recognition (OCR) system. A small collection of the collected data was manually transcribed. The source documents included: religious texts, simple readers, songs, poetry, transcribed radio programs, books, newspapers and magazines. We were also able to find a few web pages.

Gĩkũyũ incorporates diacritics (or accents), to provide two extra vowels (ĩ and ũ). In a typical Gĩkũyũ text more than 50% of words require at least one of the two diacritics. The fact that these diacritics are not readily available on the computer keyboard means that words are typically transcribed and OCRed without the correct diacritic, resulting in 'i' representing ĩ and 'u' representing ũ. Even in the print medium, there is some data that that does not have the diacritics. We therefore developed a language-independent diacritic restoration tool based on machine learning techniques (Wagacha et al., 2006). This tool was able to provide the correct diacritics for more than 90% of the words in a text.

Another technique that can be used in the collection of data is the automatic collation of web pages from the internet. An example of one such application has been used for Kiswahili (Geato and Miriti, 2005). The unavailability of diacritics on the computer keyboard is responsible for the lack of web content in Gĩkũyũ language. The above two examples illustrate that even for the data collection phase, suitable tools need to be developed.

### 3.2. Corpus Annotation using language-independent induction techniques

Linguistic annotation of large text corpora is nowadays most often done semi-automatically. Computational annotation systems generate a first (automatic) classification, which is manually corrected by human annotators. Not only does this approach increase the speed with which these corpora can be annotated, it also improves the consistency of the linguistic description. A typical first annotation task in corpus construction is part-of-speech (POS) tagging, as this is a primordial component on the one hand for linguistic description, on the other hand as a first processing step for almost all language technology applications. We therefore chose to concentrate our research on this particular annotation task.

The current state-of-the art methods for POS tagging are trained on annotated corpora (De Pauw et al., 2006 submitted). If there is no previously annotated data available however, which is by definition the case for RSLs, these traditional algorithms for POS tagging are largely useless. Rather than go through a costly human annotator phase, we opt to investigate two alternative techniques that can provide an initial annotation of the Gĩkũyũ corpus: (1) "unsupervised learning" and (2) "knowledge transfer".

In **unsupervised learning** classification is performed on the basis of free text. The standard technique is that of conceptual clustering, which describes the linguistic objects as data points in a feature space, which can be subdivided in a number of clusters. These clusters can then a posteriori be interpreted as placeholders for linguistic classes (e.g. parts of speech). Experiments for other languages have shown that this technique is very suitable as a first initialization of POS tag annotation (Schütze, 1993; Yarowsky, 1995). Furthermore, it can also provide interesting novel insights from a (psycho)linguistic point of view as it approaches the data without preconceived linguistic notions (Clark, 2001).

**Knowledge transfer** techniques try to apply the annotation properties of an adequately described language (= source language, e.g. Kiswahili) to a language for which there is no annotated data available (= target language, e.g. Gĩkũyũ). Knowledge transfer makes use of parallel corpora (the same text in two different languages) which have been aligned on the sentence and word level. The direct correspondence assumption (Hwa et al., 2002) consequently allows for the annotation of the words in the source language to be projected unto the text in the target language. Even though lexical and structural differences between languages prevent a simple one-to-one mapping, knowledge transfer is often able to generate a well directed initial annotation of the target language with a minimal amount of resources (Cucerzan and Yarowsky, 2002). The knowledge transfer technique could be particularly fruitful for transfer of linguistic annotations among Bantu languages since they are known to have grammatical similarities (Nurse and Philippson, 2003).

Unsupervised learning techniques can be directly applied to the existing Gĩkũyũ corpus. For the knowledge transfer methods, we need a parallel corpus. For this purpose we can use the bible and quran, available in both source and target language. The POS tag annotation of the source language can consequently be induced from the annotated Helsinki Corpus of Swahili (Hurskainen, 2004; De Pauw et al., 2006 submitted) and transferred to the target language.

Although unsupervised learning and knowledge transfer methods typically serve the same purpose, the literature hardly describes comparative efforts. Furthermore, they are not often applied as tools in the linguistic description of RSLs, although they are often touted as providing interesting perspectives for them. During the development of the annotated Gĩkũyũ corpus, we aim to evaluate and possibly combine these two techniques. This will give us an idea of their feasibility as language-independent annotation induction techniques. In the next section, we discuss how this feasibility can further be gauged in an extensive evaluation phase.

### 3.3. Evaluation of the annotated corpus

As noted before, the methods described in the previous section only provide a first initialization of annotation, not unlike established supervised methods would during semi-automatic corpus construction. This means that there is still a certain amount of postprocessing needed by human annotators. Evaluation of the annotation induction methods can be described in terms of how well they are able to minimize this human postprocessing effort.

Again, the evaluation is met with specific challenges when dealing with RSLs. They may have no prior formal linguistic abstractions for parts of speech or faulty or outdated linguistic abstractions for parts of speech. These issues may prevent standard evaluation against an existing gold-standard. Evaluation and postprocessing must therefore be performed manually, preferably by experts of the language. It may however also be challenging to identify and recruit human annotators for RSLs because:

- Typically, the general literacy level is low (few people may be competent in reading and writing the language).
- The language may not be standardized and thus may be represented by several dialects.
- There is a lack of formal linguistic expertise in the language.
- Computer literacy may be low so that capacity building would be needed before human annotators could make full use of automated tools.
- Linguistic expertise in the language is more likely to reside in places where the cost of labor is high.

One way to circumvent this problem is by performing a purely quantitative evaluation on the basis of perplexity measures. These are commonly used in language modeling to evaluate the generalization properties of an annotation system for natural language (Charniak, 1993). While this can provide a solid quantitative estimation of the adequacy of the induced annotation, it still does not guarantee a well annotated corpus, nor does it solve remaining annotation errors. We therefore propose two additional evaluation/postprocessing methods that are intended to minimize human intervention as much as possible.

The first method, the **utility method**, involves the annotation of a small test set using the discovered (or transferred) tags. This test set should also have a manual annotation. The two annotations can then be compared using normal tag evaluation methods. This method requires much less human intervention since the discovered tags do not have to be named. Instead quantitative measures of tag correspondence on the corpus annotation task can be used to match tags (for example how often is a manually-annotated tag, VERB, matched with an automatically induced tag, C1, when using the same test set). Different manually-produced tag sets can be compared with the automatically-produced tag sets to reduce the problem of bias (where an automatically-produced tag set is poorly scored because of a particular linguistic choice of tags.) This evaluation method may be a rich area for research study since learning algorithms, RSLs and tag choice methodologies can all be

research parameters. It is hoped that this research direction will produce some results that can be generalized.

The second method, the **comparative analysis**, compares a manually produced tag set for an RSL with a tag set produced through unsupervised learning or knowledge transfer. This could be done by human intervention, requiring the diagnosis and naming of the discovered tag set and the manual production of an optimized tag set so that the two can be compared. Both quantitative (such as the percentage of optimized tags discovered) and qualitative (such as the 'meaningfulness' of the tags) metrics would apply in this case. One method of comparative analysis uses tag instance lists. An example tag instance list for is: instance-PRONOUN = (he, she, it, I, ?). The tag instance lists produced from manually-annotated corpora can be compared with tag instance lists generated during the course of unsupervised learning. Further drill down analysis can be performed by comparing the automatically induced instance-context lists of all instances of a particular tag for correspondence with the context (bigram or trigram) present in corpus data.

An advantage of the proposed evaluation approach is that it can also be implemented using an exemplar database instead of a hand-annotated corpus. An exemplar database consists of a list of POS tags, each associated with a list of examples. It may require less expertise to manually generate such an exemplar database than it would to manually annotate a sizable corpus. The interaction between the manual and automatic annotation process can be iterative, with each iteration leading to a refinement of the exemplar database.

## 4. Conclusion

We defined a resource-scarce language as a language for which few digital resources exist; a language with limited financial, political, and legal resources; and a language with very few linguistics experts. We have used Gĩkũyũ as an example of such a language. Since this language has some very closely related languages, we believe that the knowledge, tools and expertise can be easily 'transfered' to these languages and indeed other Bantu languages.

We have proposed the use of state-of-the art unsupervised and knowledge transfer techniques. The proposed evaluation techniques, the utility and comparative analysis methods, combined with an iterative approach of developing an exemplar database, are also designed to automate linguistic processing and minimize expert human intervention.

In summary, the corpus-building strategy that we propose combines data collection, with automated linguistic annotation and semi-automated evaluation of annotated corpora to achieve the following:

- Minimize the use of scarce human resources;
- Maximize the potential of limited linguistic data;
- Develop techniques that speed RSL linguistic annotation by: (a) using learning techniques that do not require a large amount of prior information; and, (b) facilitating transfer of knowledge between already-annotated languages and similar languages that have not yet been well annotated.

If these aims are achieved within this corpus building strategy, it will go a long way towards facilitating the entry of RSLs into the digital world, thus assisting in the preservation of culture, experience and knowledge that is embodied in these languages.

## Acknowledgments

The research presented in this paper was made possible through the support of the Flemish Inter-university Council (VLIR), under the VLIR-IUC-UON program.

## 5. References

E. Charniak. 1993. *Statistical Language Learning*. The MIT Press, Cambridge, MA.

A. Clark. 2001. *Unsupervised Language Acquisition: Theory and Practice*. Ph.D. thesis, COGS, University of Sussex.

S. Cucerzan and D. Yarowsky. 2002. Bootstrapping a multilingual part-of-speech tagger in one person-day. In *Proceedings of CoNLL-2002*, pages 132–138, Taipei, Taiwan.

G. De Pauw, G. M. de Schryver, and P. W. Wagacha. 2006 (submitted). Data-driven part-of-speech tagging of Kiswahili. In *Proceedings of the Ninth International Conference on TEXT, SPEECH and DIALOGUE*.

K. Geato and E. Miriti. 2005. Process for building a Kiswahili corpus from the world wide web. In *Proceedings of the 1st Annual International Conference and Workshop on Sustainable ICT capacity in developing countries 2005.*, pages 148–152, Makerere University, Kampala.

M. Guthrie. 1967. *The Classification of Bantu Languages*. Dawsons of Pall Mall, London.

A. Hurskainen. 2004. *HCS 2004 - Helsinki Corpus of Swahili*. Compilers: Institute for Asian and African Studies (University of Helsinki) and CSC - Scientific Computing.

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

J. M. Mugane. 1997. *A Paradigmatic Grammar of Gĩkũyũ*. CLSI Publications, Stanford California.

D. Nurse and G. Philippson. 2003. *The Bantu Languages*. Routledge Language Family Series. Routledge, London, UK.

H. Schütze. 1993. Part-of-speech induction from scratch. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 251–258, Columbus, OH, USA.

P. W. Wagacha, G. De Pauw, and P. W. Githinji. 2006. A grapheme-based approach for accent restoration in Gĩkũyũ. In *Proceedings of fifth international conference on Language Resources and Evaluation, LREC 2006*.

D. Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 189–196, Cambridge, MA, USA.