Facebook Corpus Creation and Language Identification

Richard Littauer

Presented for the CL4LRL Course

April 11, 2012

1 Introduction

Half of the world's 7,000 languages have been predicted to go extinct within this century (Krauss 1992). This is a loss not only for the communities that speak dying languages, but also for linguists, who fundamentally seek to understand language phenomenon. One of the many tasks that linguists can perform in the face of this loss is that of corpus creation. By compiling corpora of minority or low resource languages, linguistic data can be analysed by linguists interested in theoretical questions, utilised by data scientists and computational linguists to provide better tools and applications, and archived for posterity. This paper is concerned with these issues, and outlines two tools that can be used towards this end: a corpus building mechanism for harvesting data from public Facebook groups, and a shallow phonotactic inducer useful for language identification in noisy, multilingual corpora.

Minority language speakers are increasingly using social networks, micro-publication sites, and web forums as venues for communicating in their own languages. This presents an opportunity for computational linguists: by utilising tools that can sort, annotate, and archive existing digital data, the long and arduous process of data collection in the field can be enhanced and supplemented. However, there are legal, privacy, and storage concerns that arise from using online corpora. In this paper, I examine a Facebook group set up by speakers of the Bantu language Rangi in order to speak their own language. I will discuss the process of harvesting data from this group, and will cover, in particular, the legality of harvesting data from Face-

book. I will also address privacy concerns for the harvested data. I furthermore present an XML schema, developed to properly annotate and archive the data and metadata.

However, in order to create a coherent corpus in a single language from a language-based user group, harvesting and XML formatting alone are not enough. Often, the harvested data may be noisy, as several languages can be used simultaneously by a single community. There are several different techniques normally used to identify the language of a document, but there are few to identify noisy text. Here, I will outline and discuss a tool that builds on previous work using trigram analyses to identify languages on the web. The tool I suggest works by utilising a trigram classifier along with phonotactic information that can be deduced from a training corpus.

2 Corpus creation via Facebook harvesting

2.1 Legality

The legality of using Facebook for corpus creation is an open question. The following section covers information regarding harvesting off of Facebook using Automatic Data Collection (ADC) services, such as scrapers, as well as legal issues surrounding harvesting an open group, and ultimately suggests the alternative of using source files and not ADC techniques for corpus creation. This topic is remarkably sensitive, and as such I could find no previous linguistic work using or creating Facebook corpora.

In Facebook's Statement of Rights and Responsibilities, section 3.2 states: "You will not collect users' content or information, or otherwise access Facebook, using automated means (such as harvesting bots, robots, spiders, or scrapers) without our permission."[1] This applies only to users. Facebook has a broader document, Automated Data Collection Terms, detailing explicit allowances for scraping.[2] They stipulate that all automated processes on the site are forbidden, unless there is express written consent. In previous cases where scraping has been done - using the same technique as google uses - the Facebook legal department has issued statements that "prior written consent" must be obtained before scraping anything.[3] If applied to the rest of the internet, this would mean that, legally, every robot or crawler - including search engines - would need prior consent. The fact that this is not the case is largely because no

[1] http://www.facebook.com/legal/terms?ref=pf
[2] http://www.facebook.com/apps/site_scraping_tos_terms.php
[3] http://petewarden.typepad.com/searchbrowser/2010/04/how-i-got-sued-by-facebook.html

precedent for this stance has been set in a court of law. Barring legal precedent, discouragement alone may be enough to stop most researchers from using automated technology on the site: the terms stipulate that "You agree that any violation of these terms may result in your immediate ban from all Facebook websites, products and services. You acknowledge and agree that a **breach or threatened breach** of these terms would cause irreparable injury..."[4] (Emphasis added).

However, there are possibilities available to the researcher to circumvent this issue: using only 'public' information, utilising the EU directive 96/9/EC regarding database protection, fair use, implied license, and finally not explicitly using a crawler or scraper. Facebook states that posting 'public information' means that "anyone, including people off of Facebook, will be able to see it."[5] In an open group, all information is publicly available. This means that, barring ACD, all of the data is legally publicly available. This conflicts with the Legal Terms article 5, which states that "If you collect information from users, you will: obtain their consent, make it clear you (and not Facebook) are the one collecting their information, and post a privacy policy explaining what information you collect and how you will use it."[6] I will address language documentation privacy concerns below; for now, in an open group, the previous statement that all public information is public necessitates that asking for prior permission is unnecessary.

Under the EU Directive 96/9/EC, any user can retrieve or re-use a non-substantial part of any database, "a collection of independent works, data or other materials arranged in a systematic or methodical way and individually accessible by electronic or other means."[7] This would hypothetically allow saving information - as from a group page - from a database such as Facebook, using the *sui generis* rights provided. This would apply only to researchers operating in the EU, as Facebook's EU offices are centred in Dublin, Ireland, although Facebook Inc. is officially registered in Delaware, USA.[8] It is worth noting that the legal situation in Tanzania might also apply to any information gathered from Tanzanian citizens posting from within the country.

Another way to approach harvesting Facebook data is by using fair use and implied license. According to Title 17 USC section 107, the use of copyrighted work, "for purposes such as criticism, comment, news reporting, teaching (including multiple copies for classroom use),

[4] http://www.facebook.com/apps/site_scraping_tos_terms.php
[5] http://www.facebook.com/full_data_use_policy
[6] http://www.facebook.com/legal/terms?ref=pf
[7] http://europa.eu/legislation_summaries/internal_market/businesses/intellectual_property/l26028_en.htm
[8] http://www.facebook.com/legal/terms

scholarship, or research, is not an infringement of copyright."[9] Fair use has a long and vibrant history, and can be used as a valid justification for using material posted by users on Facebook. Implied license differs slightly, by stating that website owners imply a license allowing people to crawl, cache, RSS, or otherwise manipulate contents. Facebook itself allows this: a diagnostic run of Network Mapper[10] shows that ports 80 and 443 are open on Facebook, ports which are normally used by search engines such as Google. Modifying the robot to collect data is not allowed under the Terms, but caching and using already publicly available data such as names, profile pictures, and gender is allowed. Where the line between publicly available data and data made public is drawn is unfortunately unclear, and how implied license and fair use would potentially play out in a European court is, unfortunately, difficult to say without precedent.

All of these possible circumventions or justifications for harvesting a Facebook group depend on both a researcher's judgement and the *post hoc* decisions of Facebook's legal team. However, they also depend on the research using ADC tools and techniques. As can hopefully be seen, this is not a decision to be taken lightly. However, there is another possible way to collect corpora from a Facebook group page that would not need ADC techniques: namely, not automating the process.

2.2 A Non-ADC Technique

The tool presented here utilises the source code for an open group. The target group used as an example is called T~~ulu~~usike K~~i~~laangi, available at `http://www.facebook.com/groups/TuluusikeKilaangi/`. In order to prevent infractions of copyright law, the main group page was loaded into a browser normally. In such a case, the source code has already been collected into the system, and automation is not necessary for retrieving more URLs. However, the immediate page presents a limited amount of data, as a limited amount of post comments are loaded initially for speed reasons. By clicking on "Display more posts..." (or scrolling down) or "View all *n* comments", an Ajax query is sent to the database, and the posts are loaded in the browser. In order to avoid using an ADC tool, each such comment was opened and expanded manually. This source code was then copied and saved as a plain text file, in HTML format. This HTML format was then parsed using the Beautiful Soup[11] Python library, which, while originally meant for scraping or harvesting (which can be defined as explicitly

[9] http://www.law.cornell.edu/uscode/text/17/107
[10] http://nmap.org/
[11] http://www.crummy.com/software/BeautifulSoup/

accessing a webpage through a port and parsing the HTML from there), can also be used for local text mining. As there is no explicit scraping or harvesting using this method, and as this text was posted by users under the knowledge that they were publishing it publicly, the source should be open to the public for generic use. However, as a safeguard, the data gathered using this technique should not be used for anything but research, and should not be hosted online or shared.

2.3 Privacy

As mentioned above, there could be privacy concerns regardless of the public nature of the data. Facebook makes it clear that written consent should be gotten from each person if information is gathered from them. This is standard procedure in language documentation, and for any sort of data gathering in the social sciences. Many universities, and some countries (for instance, the UK) require approval from an Ethics Board before creating, using, and especially funding databases of this sort. In this case, I feel the privacy concerns are negligible, for several reasons: the data being gathered has already been released into the public domain according to the Terms of Use provided by Facebook, the data will not be shared or monetized, all names and personal data will be anonymised if the data is made public, and the data is being used purely for research.

Furthermore, the entire existence of the group itself stemmed originally from a desire for the language in question, Rangi, to be more present, both online and in research. The founder of the group, Oliver Stegen[12], is an experienced linguist who is the main expert on the Rangi language in Tanzania. He set up the group as a way to create online corpora and to promote the use of Rangi on the internet.[13] If there are any privacy issues raised by concerned members, it would be easy to remove all of their public posts. The group's wish for this data be used has been independently checked (by Stegen) through personal communication with the users most active in the group.

Finally, the Facebook User Operations team in Dublin responded to a personal query regarding gathering the data in the manner specified above with the following statement:

> Thanks for your email. Please note that we do not permit data scraping, but some site content may be publicly available depending on privacy settings.

[12] http://oliverstegen.net/

[13] Verified through personal communication with the author.

Given this, I feel that the data gathered here can be safely used as a corpus, and that future researchers may also use this technique, if the above precautions are taken, without fear of legal repercussions. It is important to note that in this particular case there are strong links between the linguists studying the group and the members of the group, and such community links are strongly encouraged.

3 XML format for online discourse

```
<text id=``text_id" source_id=``input_file" title=``title">
  <metadata idref=``text_id">
    <!-- incorporate OLAC metadata standard -->
  </metadata>
<body>
  <thread id=``thread_id">
    <comment id=``thread_id.comment_id" idref=``thread_id">
      <author id=``author_id" idref=``thread_id.comment_id" url=``author_href">
        author_name
      </author>
      <timestamp id=``thread_id.comment_id.timestamp_id" \
        idref=``thread_id.comment_id" utime=``utime">
        time_readable
      </timestamp>
      <likes id=``thread_id.comment_id.like_id" idref=``thread_id.comment_id">
        <like_author id=``author_id" idref=``thread_id.comment_id.like_id" \
          url=``like_href">
          author_name
        </like_author>
        <likes_total idref=``thread_id.comment_id.like_id">
          likes_total
        </likes_total>
      </likes>
      <source id=``source_id" idref=``thread_id.comment_id">
        mobile
      </source>
      <plaintext>
        message_text
      </plaintext>
      <phrase id=``thread_id.comment_id.phrase_id" lang="code"\
        idref=``thread_id.comment_id" text=``phrase" />
      <word id=``thread_id.comment_id.word_id" lang="code"\
        idref=``thread_id.comment_id" text=``word" />
    </comment>
  </thread>
</body>
</text>
```

Table 1: XML schema for Facebook data storage

The group data, gathered from Facebook through the manual technique described above, is exceedingly large. When it is output using standard structured HTML format (with a break after

each tag), the file from Tuluusike Kilaangi covering one year (February 11, 2011 to February 17, 2011) is almost 300k lines of HTML. Mining this HTML format is not trivial. In order to make this process easier, and to store the data in a format that is both machine- and human-readable, an XML schema to store the data has been created, and is shown in Table 1.

An XML (extensible markup language) format is convenient for several reasons. XML is not reliant on any single, particular program, and is widely used for data storage already. XML works by conforming to a schema, or prototype, which means that it is easy to check for errors in formatting. XML has the advantage of being easily converted into RDF and other useful storage formats. It is easy to understand for both humans and machines. XML schemas can also be stored independently of the data, as in databases like ISOcat,[14] meaning that others can utilise previously existing data formats, without reinventing the wheel. Here, the independent creation of a new XML schema was necessary, as prior work using Facebook data has not been undertaken due to legal concerns, to my knowledge.

The schema provided above records all of the relevant data that can be gathered or might be needed to a reasonable extent by researchers - this means that it stores information about where the comment is in the page, as well as when it was made, who made it, how they made it, and what it was in response to. This may seem excessive - however, metadata is crucial for documentation, as it provides a means of analysis after the fact and can be studied in and of itself (Nathan and Austin 2004). It stores this information using layered XML embedding. Each element requires inline annotations - for most elements, this involves a unique ID, and a referent to the unique id for the enveloping layer. This is not easily flexible. However, due to the non-ADC process used in the initial gathering of the HTML source file, the XML formatting itself is a one-time event that does not require post-processing. Some elements are optional: for instance, <likes> and <source> occur only if there are likes for a comment, or if a comment is posted using a mobile device.

The suggested XML schema stores items based upon the thread and comment format of Facebook and other networking sites. An initial comment is stored in a <comment> element, and each responding comment is stored within the same <thread> element. It is possible that it can be adapted to micro-publication sites, such as Twitter, or to other online forums or social networks, as the thread and comment format is common. The initial comment is embedded at the same level as the responding comments, as the inline annotation attribute 'id'

[14] http://www.isocat.org/

sequentially increases in value for each comment. Each <comment> element must minimally have a <author> and a <timestamp> element, both of which refer back to the comment and thread id, and a < plaintext > element, which does not. This is due to the nature of the corpus; ideally, the plain text would be post processed and language identified. As the block test may contain several languages, but as there cannot be multiple identical elements, this element is used as a cover-all for the entire comment. Other XML formats used for linguistic data also follow this system (Palmer and Erk 2007).

The elements <phrase> and <word> are optional, and depend on sorting the contents of the < plaintext > into separate phrases and words. As can be seen in Table 1, if this step is undertaken, there is an attribute 'lang' which can be used to identify which language or languages is represented in the ' text ' attribute. This would allow for storage of a parallel multilingual corpus using this XML schema, after processing. As language identification is not a trivial issue, these attributes are not included in the < plaintext > element.

In line with the previous discussion on privacy concerns, the <author> element can be made optional. It can also be anonymised by removing the contents and the ' url ' attribution, bot of which point to the authors identify. The unique ID attribute should provide anonymity enough to satisfy any worried parties, as it would not then be attached to any other external information that could identify the author. However, the author information is included for legitimate reasons: it shows the context, and it enables the researcher to check on the language proficiency of the speaker. For instance, by following the URL and examining user profiles, it would be possible to check if someone is a native speaker, and to remove or include their comments in the corpus based on this evaluation.

The schema above does not allow for other linguistic annotation, such as part-of-speech tagging, or morphological or syntactic annotation. It is meant primarily as a storage format, to maintain the context of each comment and all detail that may be relevant to linguists from the original page. A different annotation format would need to be used for further annotations, but that is beyond the scope of this paper.

The largest corpus currently on the net for Rangi is from the Án Crúbadán crawler: this corpus is 108 documents large, and is comprised of 17,908 words and 123,354 characters.[15] The corpus that this paper has compiled and stored contains 990 threads, 64,891 words and 571,182 characters.

[15] http://borel.slu.edu/crubadan/stadas.html

4 Previous work

It is rarely the case that the minority language is the only language used online. Minority language speakers are often from under privileged areas in countries with little online infrastructure; such is the case with Rangi in Tanzania. The *de facto* language of the internet is English, and minority language speakers must have some knowledge of English in order to access certain websites. Although this is changing, and it may now be possible for many speakers of different languages (such as Chinese or Russian) to access operating systems, browsers, and websites in their own language, it is estimated that only around thirty languages currently enjoy full technological resources, and only a 100 or so have basic resources such as dictionaries, spellcheckers, or parsers (Scannell 2007; Krauwer 2003).

In many cases, tools and infrastructure have been developed for only officially recognised national languages: in Tanzania, both Swahili and English are official languages. In order for a Rangi speaker to use the internet, they must be in effect trilingual. Although software like Google Swahili[16] is available and becoming more widespread, any corpus gleaned from a Rangi online community will likely have instances of all three languages, if not more. This creates a problem where there is noisy data - any corpus will have all three languages within it, unless the languages are sorted afterwards. Doing this manually takes time; automation is key, but the automation process is both complex and difficult, especially for non-annotated data.

One way to automate the process is to identify sections that are likely to be in a given language, using a dictionary from that language to measure this probability. This depends upon a pre-existing dictionary, however, and is very computationally heavy. A short option is to use only the most commonly appearing words in a language, as Zipf's law (Zipf 1949) states that shorter word length is correlated to higher frequency. These words can be identified in a training corpus, and then the algorithm could test shorter words alone to find a language. This works for large corpora, but may not work for sections that have fewer words. A third option is using *n*-grams of orthographical regularities in a language from a training corpus, and testing unknown data on the *n*-grams frequencies from all gathered languages to see which is most likely. This has been used widely (Scannell 2007), but a trigram analysis has been shown to not perform adequately on strings smaller than five, and it may not be efficient in distinguishing dialects or sister languages. Here, I suggest another technique: basic phonological and phonotactic induction combined with the trigram (*n*-gram) approach.

[16] https://www.google.com/webhp?hl=sw

As mentioned above, there is a large literature surrounding language identification, and many possible ways to ascertain the language of a given text. For this corpus, there is a constraint which limits the use of many of these techniques - the length of individual comments. Each comment is language independent. Identification must then occur on a smaller level than either the corpus or the thread. It is also possible, however, that a comment may code-switch between languages. In such cases, single words may be used from different languages, and identifying the languages of these words is not trivial. Using a dictionary may not produce the best results: for instance, there may be identical forms with entirely separate meanings between differing languages. There are very few approaches in the literature to this particular granularity problem in multilingual texts, as most language identification works at only the document level. (Hughes *et al.* 2006).

This problem is particularly present in cases such as this Rangi Facebook group. Rangi shares much of its lexicon with Swahili, as they are closely related (Stegen 2003). In many cases the language will be clear due to Rangi orthography: e.g. Swahili <nyumba> and <bata> against Rangi <nyuumba> and <ibáta> (following standard linguistic conventions, < and > are used here to represent orthographical form - XML elements will be indicated explicitly in the text). Swahili loans into Rangi are easily recognisable as Swahilis penultimate stress is reinterpreted as a falling tone on a long vowel, as in Swahili <pete> to Rangi <péete> or Swahili <zawadi> to Rangi <zawáadi> (Stegen, personal communication). But this may not always be the case. Similar to this problem is that the tradition of writing in Rangi is new - an orthography has only been developed in the past couple of decades, and linguistic analysis of the language has been intermittent for only a century (Seidel 1898; Dempwolff 1916; Akhavan-Zandjani 1990). Without a strong literary tradition in Rangi, Swahili (and English) influence may be accentuated further than would be expected for a minority language (even a minority language with more than 300k speakers) (Stegen 2004; Stegen 2005).

Scannell presents a possible solution to the problem of language identification for small strings. Án Crúbadán (Irish for 'crawler'), has created corpora for over 400 languages, and in some cases the only corpora known for those languages. (One of these language is also Rangi, using text supplied by Stegen[17].) The crawler relies on expert opinions to identify and suggest training corpora, and to suggest work arounds for orthographical or other issues. Then, it searches using a trigram analysis - by looking only at three character length strings, and

[17]http://borel.slu.edu/crubadan/stadas.html

creating a list of the most common strings, text can be easily identified as coming with a reliable probability from a specific language. This works relatively well for different languages, but does not apply well to language pairs and dialects, such as Norwegian-Danish or Indonesian-Malay, nor for cases where there are conflicting orthographies, such as in reconstructed Cornish (Scannell 2007, 8). While this technique is exceptional at identification for small strings and without unnecessarily expensive training, as can be seen by it's successful identification of 1029 (currently) languages,[18] this trigram analysis does not work well for strings smaller than 5 in length (Scannell, personal correspondence).

5 Methodology

I suggest a system that builds on the trigram classifier used by Scannell. A trigram analysis of a text fails to comprehend the differences between different characters in word formation, depending entirely upon collocations. However, there are different units that together make up a word above the level of the character which are language specific - namely, phonotactic constraints and rules. On the simplest level, vowels and consonants (and glides) interact in predictable ways. As was mentioned above, for instance, syllable construction reflects stress, which is above the level of trigram analysis as it involves two syllables. This reflection can be seen in differences between character doubling to represent falling tone in Rangi as opposed to falling tone in Swahili which is not reflected in the orthography. There are other examples: for instance, Rangi and Swahili words are predominately constructed with CV syllables, and closed syllables are not allowed.

In order to accurately capture the phonotactic rules of a language, a sophisticated knowledge of the sound system is necessary. This would involve expert advice for each language. However, as long as there are regular orthographic conventions for a given language, surface-level analysis may be enough. For this study, only vowels, glides, and consonants were used to analyse phonotactic constraints.

The training data used for this script was identical for each of the three languages: this was done by using the book of Genesis, one of the few books to be translated and publicly available for Rangi. Ideally, the Facebook group would have been used; however, as it has not been annotated for language identification, there would be no way to test the accuracy. As the Bible

[18] http://indigenoustweets.blogspot.de/2011/12/1000-languages-on-web.html

is often the first book to be translated into a low resource language, mostly through the efforts of linguists (such as Stegen) working for SIL,[19] this is a source that would be likely to be usable by other minority languages. For each language, an inventory of letters used in the orthography was decided upon. In each case, all non-alphabetic characters were thrown out, unless they were considered orthographically important. This exception occurred only for Rangi, which uses <ʉ> and <ɨ>, as well as acute accents. Two factors may have influenced the data: the use of standard ascii formatting in the original documents but not in the Python code, where the latin-1 charset was used, and the process of ripping the Rangi text from the Genesis .pdf file (the only available source), as the accents were misplaced. Regex search and replaces were able to spot fix most issues of this nature, but not all.

A training corpus was created using an extemporaneous XML <plaintext /> tag for testing accuracy after identification. This corpus was created by forming semi-random length strings from each of the three corpora and combining them in one, language annotated corpus. The identifier used here was blind to these annotations. For each word, < and > tags were wrapped around the edges to show word boundaries as a character: this is the same process used in Án Crúbadán. Each character was then converted into a <C>,<y>, or <V>, (brackets not included in the code) depending on matching the language-specific lists of vowels and consonants. Finally, the manipulated words were stored in a list, with the most common constructions sorted in a descending order, normalised by dividing by the total amount of occurrences of words in the training corpus.

This system can work in conjunction with the trigram counts, to provide more accurate identifications on the word, phrase, or paragraph level. The trigram analysis approach was replicated, and the output was normalised by dividing by the total amount of trigram appearances in the training corpus.

newpage

6 Results

The trigram analysis replicated hovered in the range of 22-17% accuracy. The final corpus, compiled from 1000 random strings from the three corpora, was 62k characters long. This was more than enough time for the results to stabilise.

[19] http://www.sil.org/

Analysis	Accuracy
Trigram	20%
Syllable or Phonotactic	20%

Table 2: Results for Language Identification Algorithms

The phonotactic analysis also hovered at around 20%. It also performed much slower.

7 Discussion

Both of these results are surprising, as they are so low. Other language identifiers of noisy data do not have nearly the same results: For instance, Án Crúbadán generally has an identification rate that is near perfect. Rosner and Farrugia's (2008) Markov Model identifier performs at almost 99% ; Cavner's (1994) *n*-gram identification at 99% ; Hayati's (2004) *n*-gram, Fisher, and cosine model is close, at 33%, but that is both not helpful in pointing out what went wrong, and not as poor as this run . Hoogeveen and de Pauw's (2011) runs at 89%, but they also cite their worst comparison as Google at 39% correct identification.

There can be only three possible conclusions: that there is something wrong with the implementation used here, or that there is something wrong with the data. The data is not at fault, barring the two possible sources of contamination I mentioned earlier; namely, copying and pasting from the .pdf causing issues with accent placement and formatting, or unicode and ascii mismatches. The code, then, is probably at fault. This can be seen clearly as the replication of the trigram analyses is not near what the original results are.

The other possible solution is that the metric is off. Accuracy was measured here as the amount of words correctly identified as from a certain language, over all words in the corpus. This seems to be the most logical way of measuring accuracy, so I must conclude that the implementation is poor. The fact that the trigram scored just as poorly as the phonotactic analysis does not rule out the possibility that the shallow vowel, consonant, and glide analysis was unsuitable for language identification, but there is no way of knowing without other tests.

Future work would necessarily include discovering why the results are so low; and if they are, in fact, a systematic problem with the implementation, or a larger problem that may shed light upon the field of identifying language for small strings. Neither of the analyses provided here looked deeply at larger strings - anything more than 20 characters. This may explain the

low results without embarrassment. More work should be put into this. Future work would also include running different algorithms simultaneously, in sequence, and live instead of over a corpus. Integrating language identification with parsing or tagging models is another exciting area of possible research.

8 Conclusion

While the results of the two analysis presented for language identification are disappointing, the final product is not limited by this alone. As mentioned above, the corpus created from the Facebook group studied here is the largest currently in existence for Rangi; If, at a low estimate, only a third of this Facebook corpus is written in Rangi, it still represents a corpus almost twice as large as any other currently in existence. This is a remarkable achievement, and it is hoped that the legal ground covered behind using this corpus and sharing it with other researchers and the language community will enable Rangi to be used and studied more than it is today.

References

Akhavan-Zandjani, Firouzeh (1990). *Untersuchungen zur Grammatik des Irangi anhand des Materials aus dem Nachlass Dr Paul Berger*.

Dempwolff, Otto (1916). Beiträge zur Kenntnis der Sprachen in Deutsch-Ostafrika, 8: Irangi (Langi). *Zeitschrift für Kolonialsprachen* 6: 102–123.

Hughes, Baden, Baldwin, Timothy, Bird, Steven, Nicholson, Jeremy, and Mackinlay, Andrew (2006). Reconsidering language identification for written language resources. In: *Proceedings of LREC2006*. (pp. 485–488).

Krauss, Michael E. (1992). The world's languages in crisis. *Language* 68: 4–10.

Krauwer, Steven (2003). The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources Roadmap. In: *Proceedings of the 2003 International Workshop Speech and Computer (SPECOM 2003)*. (pp. 8–15). Moscow State Linguistic University.

Nathan, David, and Austin, Peter K. (2004). Reconceiving metadata: language documentation through thick and thin.. *Language Documentation and Description* 2: 179–187.

Palmer, Alexis, and Erk, Katrin (2007). IGT-XML: An XML Format for Interlinearized Glossed Text. Proceedings of the Linguistic Annotation Workshop. Prague.

Scannell, Kevin (2007). The Crúbadán Project: corpus-building for under-resourced languages. In: C. Fairon and H. Naets and A. Kilgarriff and G-M de Schryver (ed.) *Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop*. (pp. 5–15). Louvain-la-Neuve, Belgium: PUL.

Seidel, A. (1898). Grammatik der Sprache von Irangi. In: Werther, C. W. (ed.) *Die mittleren Hochländer des nördlichen Deutsch-Ostafrika: Wissenschaftliche Ergebnisse der Irangi-Expedition 1896-1897 nebst kurzer Reisebeschreibung* (pp. 387–434). Berlin: Hermann Paetel.

Stegen, Oliver (2003). First steps in reconstructing Rangi language history. Paper presented at the 33rd Colloquium on African Languages and Linguistics at Leiden, August 25-27, 2003.

Stegen, Oliver (2004). A pilot study of writing in Rangi society. *Edinburgh working papers in applied linguistics* 13: 102–111.

Stegen, Oliver (2005). Editing Rangi narratives: a pilot study in literature production. *Edinburgh working papers in applied linguistics* 14: 68–98.

Zipf, George K. (1949). *Human behavior and the principle of least effort*. Cambridge, Mass.: Addison-Wesley.