Focused Crawling for Automated IsiXhosa Corpus Building

Cael Marquard[1][0009−0003−9420−6199] and Hussein Suleman[2][0000−0002−4196−1444]

[1] University of Cape Town, Cape Town, Western Cape, South Africa
cael.marquard@gmail.com
[2] University of Cape Town, Cape Town, Western Cape, South Africa
hussein@cs.uct.ac.za

Abstract. IsiXhosa is a low-resource language, which means that it does not have many large, high-quality corpora. This makes it difficult to perform many kinds of research with the language. This paper examines the use of focused Web crawling for automatic corpus generation. The resulting corpus is characterised using statistical methods: its vocabulary growth has been found to fit Heaps' Law, and its word frequency has been found to be heavy-tailed. In addition, as expected, the corpus statistics did not match expectations from non-agglutinative languages.

Keywords: Corpus · IsiXhosa · Web Crawling · Low Resource Languages.

1 Introduction

IsiXhosa is a low-resource language, meaning that there are few large, high-quality corpora available for research purposes. This makes it difficult to perform many kinds of research with the language, such as the testing of novel information retrieval algorithms, the training of machine-learning models on isiXhosa, and corpus linguistics research.

Web crawling is the process of visiting pages on the Internet and then recursively visiting the pages that they link to, in order to build a view of the content on the Web. The Web has been used in the past as a source for corpus building for isiXhosa and other languages, as it contains many freely-accessible documents. Other methods of building corpora include paying participants to answer prompts [3], or manually curating collections of documents for inclusion. These methods require more time and money in comparison to Web crawling, which can be done as a largely automated process. Therefore, Web crawling arguably may be used to address the lack of isiXhosa corpora in a cost-effective manner and enable further research with the language. Additionally, a corpus of documents obtained from the Web is likely to be more representative of the kind of documents that an isiXhosa Web search engine would need to process, and thus may be more suited to information retrieval research in Web indexing.

Focused crawling is an optimised approach to Web crawling, which is applied when the aim is to retrieve only pages that satisfy specific criteria [7].

While broad Web crawling may require large amounts of storage, bandwidth, and CPU time, focused Web crawling uses fewer resources, meaning it can be applied at a lower cost. The broad Web crawling approach has been modified through a directing algorithm that decides which links to crawl and how, in order to maximise the number of relevant pages crawled per total number of pages crawled. The focused crawler decides to save a page or crawl links based on the number of isiXhosa sentences found on the page. This is ascertained using a language identification algorithm, which estimates the probability that a given piece of text is written in isiXhosa.

This research sought to investigate this approach to corpus building. In particular, the following research question was posed:

To what degree is it possible to build a high quality isiXhosa corpus automatically through focused Web crawling?

To answer this question, a corpus of isiXhosa documents was collected using focused Web crawling. This resulting corpus has been analysed and compared to linguistic benchmarks such as Heaps' [19] and Zipf's laws [15] in order to characterise the text. The resulting metrics are compared to other corpora, both in isiXhosa and in other languages. The veracity of the isiXhosa language identification approach has been manually verified, and the sites included have been categorised as well as classified as likely machine translated or likely human-written. The source-code of the crawler and the analysis tools is provided as a Git repository, hosted on GitHub.[3] This paper is based on research conducted as part of coursework undertaken by the first author.

2 Related Works

Barnard et al. [3] built a corpus of South African languages by recording the utterances of speakers in response to prompts generated from pre-existing corpora. With this manual approach, an isiXhosa corpus of 56 hours of speech (136 904 words) was constructed, in addition to similar length corpora for the other languages targeted by the researchers.

As an alternative to manual approaches, focused crawling has been applied to building corpora of specific languages [13, 14]. In the case of Swiss German (GSW), over 500 000 sentences were collected over a period of approximately three months [13]. The algorithm used to direct the crawler was remarkably simple — if the page had more than two GSW sentences on it, outgoing links were also crawled.

By comparison, Corpulyzer uses a more complex seeding and directing algorithm [18]. The process begins with filtering the Common Crawl Corpus[4] to extract only the web pages in the target language. Then, sites are prioritised based on the percentage of target-language content found on the site as a whole, and on its individual pages.

[3] Source code available at https://github.com/Restioson/isixhosa-crawler
[4] https://commoncrawl.org/

Focused crawling has been proposed in order to capture documents most relevant to a given topic [7]. The initial crawler design proposed by Chakrabarti et al. [7] relies on breaking down Web pages into a taxonomy based on their topic. This, however, does not suit the application of focused crawling to language-specific corpus building, as it is unclear how the output of a language identification algorithm could fit into a topic taxonomy.

Gaustad and Puttkamer's dataset [9] is an example of a corpus of isiXhosa created from the Web. The original dataset was obtained by randomly selecting documents from official South African government websites, and then using an existing language identification tool in order to filter them. This approach is similar in that it also uses documents from the Web, but it is limited to one website, and will not crawl links from government websites to other sites in order to find more documents.

Crawling in combination with language identification has been applied to the task of collecting isiXhosa documents for the purpose of building an isiXhosa search engine [11]. However, the approach taken by Kyeyune [11] was to use a broad, undirected crawl and filter the documents using language identification after the fact. Using a focused crawling approach, it is hypothesised that documents can be more efficiently crawled from the web than by using a broad crawl, as many non-isiXhosa pages could be filtered out, leading to a higher harvest rate.

In order to decide whether to save a page or not, the language in which the page is written must be identified. There are many pre-existing language detection software packages such as Lingua[5], but many of them suffer from a lack of coverage of South African languages, which can lead to issues such as confusing text in Chichewa for text in isiXhosa. Since Chichewa and isiXhosa are not mutually intelligible, this kind of error is not acceptable. Fortunately, there are other approaches to classifying South African languages, such as using rank order statistics [8], which perform far better at this task. In the end, the NCHLT South African Language Identifier [16] was chosen, since it is fairly accurate, is provided as pre-built software, and is simple to interface with.

The corpus gathered by the crawler may be characterised through a variety of statistical measures. Zipf's law predicts that, in a corpus sufficiently large, the rth most common term will have frequency proportional to $\frac{1}{r^\alpha}$, where $\alpha \approx 1$ [2, 15]. Heaps' law predicts that, in a corpus sufficiently large, the vocabulary size of the corpus will grow with respect to its total size in words N according to the power law $k \times N^\beta$, with k and β being parameters to the curve [2,19]. These predictive models have been shown to hold for the vast majority of languages. Thus, the quality of the corpus can also be judged by how well these laws hold.

IsiXhosa is an agglutinative language. This means that each word may consist of several, clear-cut morphemes [1]. Morphemes are the smallest unit of words which carry meaning [17]. Many of the methods for analysing corpora are based on languages which are not agglutinative, such as English. Hence, they may not work as well for a language like isiXhosa. Thus, instead of segmenting the corpus

[5] https://github.com/pemistahl/lingua

on word boundaries for analysis, it may make sense to segment the corpus by word and then by each morpheme in each word (morphological decomposition). Another approach could be to segment the corpus into n-grams, which are character sequences n characters long [12]. N-grams have been used before to identify the language of a given text, and it has been suggested that Zipf's law holds for n-gram frequency in corpora [5].

3 Methodology

3.1 Design of Crawler

The crawling application itself is comprised of three main components:

1. **Seeding**. The seeding algorithm is responsible for creating a list of pages used by the crawler to begin its search.
2. **Language identification**. This allows the crawler to ascertain if a certain page is written in isiXhosa and should be added to the corpus.
3. **Direction**. This allows the crawler to decide which pages to crawl and in what order.

The crawler was written in Python using the Scrapy[6] framework. Scrapy is a mature framework for writing Web crawlers in Python and has many utilities for controlling the crawling process, such as automatically respecting `robots.txt` and rate-limits in order not to be a burden to the websites included in the corpus. The source code of the crawler is available on GitHub[7].

Seeding Process To begin the crawling process, the algorithm was supplied with a list of initial URLs from which to begin the recursive search of the graph of Web pages.

Linder et al. [13] leverage Google search for the creation of a list of seed URLs. By generating combinations of Swiss German words, Google's indexed archive was searched for documents likely to contain Swiss German. Medelyan et al. [14] used a similar approach, adapted to their more specific criteria, which included both topic and language.

A similar approach was used for the isiXhosa crawler, in which single isiXhosa words (from the dataset of the IsiXhosa.click live dictionary[8]) were searched through Google's Custom Search JSON API[9]. URLs were kept as seeds if the returned snippet was identified to be isiXhosa and they were not contained within the blocked sites list. It was hypothesised that higher quality seeds could be generated this way, since pages that happened to contain a heteronym of an isiXhosa word but no isiXhosa text would be excluded, speeding up the

[6] https://scrapy.org/
[7] https://github.com/Restioson/isixhosa-crawler
[8] https://isixhosa.click
[9] https://developers.google.com/custom-search/v1/overview

initial stages of the crawl. Additionally, the website of I'solezwe lesiXhosa[10] was included manually in the seed list, as it is a well known hub of isiXhosa content.

Sites in the block list were machine translated sites, dictionary websites, sites with isiXhosa content in navigation elements only, and sites from the WikiMedia projects[11]. The WikiMedia projects were excluded, since these sites are available as public data downloads and are organised by language, which therefore makes them uninteresting to crawl. Although, they may have links to other isiXhosa language material which is not available as public data downloads, they were excluded so as to streamline the seeding process, with the assumption that these materials would be linked elsewhere on the internet, too.

In order to comply with Google's rate limit of 100 searches per day, the API was queried four times on four separate days, each time with a new, random set of words, selected with replacement. The seed URLs were then deduplicated, and sites in the block list removed. In total, 235 unique seed URLs were gathered.

Language Identification First, the human readable text from each page was scraped using the BeautifulSoup library[12]. Then, NLTK [4] was used to tokenize the text into sentences, using the Punkt algorithm [10]. Since an isiXhosa sentence tokenizer was not available within NLTK, the English model was used. Because isiXhosa also terminates sentences with the same punctuation marks as English, this was not anticipated to be an issue. Sentences were then split into subdivisions of 300 characters using the standard library text wrapping function since, if the sentence is too long, it is rejected by the language classifier. After this, the segmented sentences were sent to the language classifier and deemed to be isiXhosa if the classifier had a confidence value of at least 0.5 and isiXhosa was the most likely language determined for the text.

Direction Based on the hypothesis that isiXhosa pages link to other isiXhosa pages more often than a random page links to an isiXhosa page, the prevalence of isiXhosa text on the page can be used to determine whether the pages it links to should be crawled or not. This is similar to the approach that was used by Linder et al. [13] — pages that had at least one Swiss German sentence had their links crawled.

Separate heuristics were used to decide whether to save a page and whether to crawl a given link. A page was saved if it contained over five isiXhosa sentences, or over 40% of its sentences were in isiXhosa. Sites known to be using the GTranslate[13] machine translation plugin were excluded. This plugin is very widely used on the web in order to provide automated translations of content on websites. However, since the translation quality is often quite poor, sites containing text translated by GTranslate were excluded. A site was determined to be

[10] https://isolezwelesixhosa.co.za
[11] https://wikimediafoundation.org/our-work/wikimedia-projects/
[12] https://crummy.com/software/BeautifulSoup/
[13] https://gtranslate.io/

using GTranslate if an HTML comment starting with "delivered by GTranslate" was present in the document.

All links from a page were added to a list to be crawled if the page contained at least one isiXhosa sentence. Additionally, links would be added to the list if the anchor text of the link was identified as isiXhosa, or if it included the substring "xhosa". This was to account for pages that were written in English or other languages, but had links to isiXhosa versions of the page. Links were prioritised in the crawl if they contained "xhosa" or had their anchor text written in isiXhosa. Otherwise, they were assigned default priority. This likely did not affect the result of the crawl since, by the end, the sites were crawled almost completely

The list of links was filtered to remove websites in the block list before being added to the crawl queue.

4 Analysis

4.1 Basic Statistics and Validation

The crawler was run from 29 September to 20 October 2022. Table 1 lists some basic statistics about the corpus obtained.

Pages	202 646
Total words	75 807 261
Unique words	672 460
Total sentences	4 663 036
Unique sentences	1 002 714
Websites	90

Table 1. Basic statistics of the corpus.

The largest five domains (containing the most isiXhosa documents) crawled are listed in Table 2, along with the number of documents obtained from each, the category of the website, and whether it was determined as being machine translated.

Domain	Pages	Category	Machine translated
seals.ac.za	107 021	Academic	No
jw.org	67 888	Religious and News	No
churchofjesuschrist.org	10 696	Religious	No
fundza.mobi	4 814	Literature	No
isolezwelesixhosa.co.za	3 882	News	No

Table 2. The top five sites in the corpus, ranked by number of documents

It is worth noting that out of the five top sites, two of the websites are religious. Out of the top 15, six are religious. Specifically, these sites contain isiXhosa translations of the Bible, as well as other texts about Christianity. This could be one of the reasons for the high number of duplicate sentences in the corpus — each website may be hosting the same translation of the Bible.

Jw.org in particular contains translations of religious text, but it also contains (religious) isiXhosa news media, so it has been categorised as both religious and news. The largest domain, seals.ac.za, contains mostly documents from the SEALS Digital Commons[14], which is a collection of academic output from Eastern Cape universities. According to the 2011 census [6], isiXhosa is the most common first language in the Eastern Cape, so it is no surprise that Eastern Cape universities produce much of the isiXhosa content available on the internet.

4.2 Statistical Distributions of the Corpus

The text in the corpus was evaluated using standard corpus characterisation techniques and compared against benchmarks such as Zipf's law and Heaps' law [2,15,19]. It can then be evaluated whether the corpus fits general expectations of natural language corpora. This was done by plotting the data against manually selected Zipf's and Heaps' curves. In order to evaluate the corpus, the order of the documents in the corpus was randomized using the `shuf` program from the GNU Core Utilities, and it was then passed to a utility written to process the JSON Lines format output by Scrapy[15]. Since the raw data was very large (22GB), it was processed in parallel, with each document being sent to a worker thread. In order to parallelise the language identification process, the NCHLT classifier server [16] was launched 12 times — one instance per logical CPU core on the machine used for analysis.

Agglutinativity of IsiXhosa It should be noted that since isiXhosa is an agglutinative language, relying on word segmentation may yield results inconsistent with broad expectations, since the number of distinct words is likely to be higher, given that meaning is often created by appending morphemes to words in a sentence. Therefore, in addition to a standard word-based analysis, a modified trigram-based analysis has also been performed, which segments the text into character n-grams of length three [5, 12].

Duplicate Texts Since many websites crawled contain the same headers and footers, it was expected that many sentences would be duplicated across the corpus. Since isiXhosa translations of the Bible was a notably large source of documents in the corpus, this increases the likelihood of duplicate sentences occurring in the corpus.

[14] https://vital.seals.ac.za/vital/access/manager/Index
[15] https://scrapy.org/

A simple duplicate check based on case-insensitive sentence-level equality reveals that only 22% of the sentences (1 002 714) in the corpus are unique (which represents roughly 19% of the corpus's overall word count). However, duplicates are not uncommon in online data sources and many potential uses for crawled corpora, such as search enginge indexing algogirthms, exploit this fact. Hence, despite the potential to skew the analysis, the duplicate sentences have been included in the dataset.

Zipf's Law Zipf's law is a predictive model that estimates that, in a sufficiently large corpus, each word has a frequency in the text that is inversely proportional to its rank [2, 15]. For example, the second most common word is estimated to occur roughly half as frequently as the first most common word. The harvested corpus can be compared to this idealised frequency falloff curve in order to determine whether it matches this general benchmark of word usage.

When word frequency and rank is plotted on a log-y graph (Figure 1), it can be seen that the frequency of a word versus its rank fits a heavy-tailed distribution.

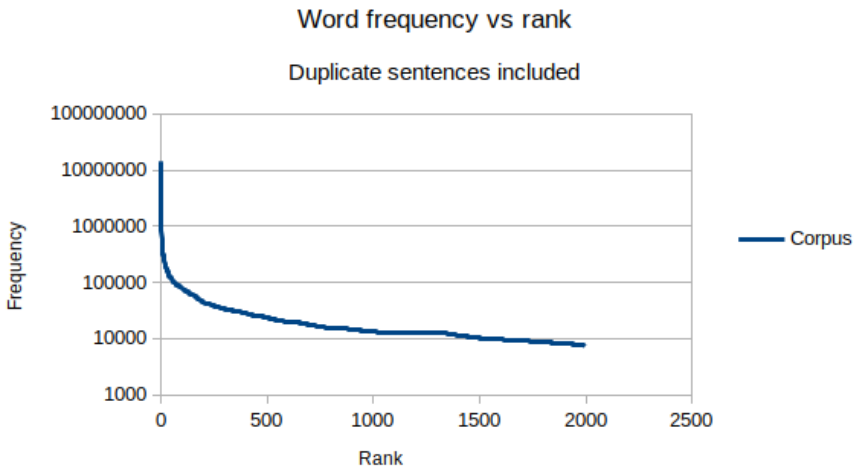

Fig. 1. Curve for the corpus on a log-y graph, including duplicate sentences.

Token-to-Type Ratio The Token-to-Type ratio (TTR) of the corpus can be measured in order to estimate lexical variety in the corpus. The TTR of a given corpus can be calculated as the corpus size in words divided by the number of unique words in the corpus [2]. This ratio can be computed for the corpus in order to compare its lexical variety to other corpora.

When duplicate sentences are present in the corpus, the lexical variety appears to grow slowly with the corpus size (Figure 2). However, when they are excluded, the lexical variety grows much faster. The TTR of the full corpus is 21.5 when duplicate sentences are removed. This is quite different to the figure obtained in the work by Ali et al. [2], which has a TTR of 27.17 for English and 26.71 for Arabic at 800 000 words. By comparison, the isiXhosa corpus obtained has a TTR of 4.81 with duplicate sentences excluded or 11.81 with duplicate sentences included at a corpus size of 800 014 words. This could be due to the fact that isiXhosa is an agglutinating language, meaning that a higher lexical variety is expected.

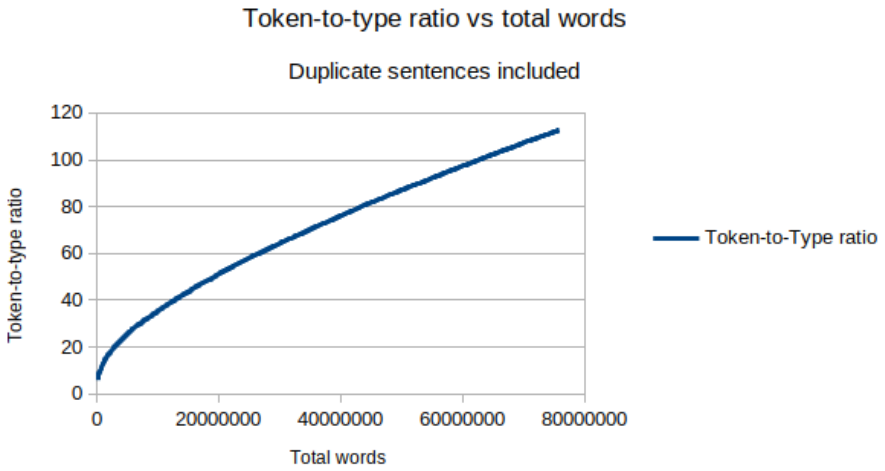

Fig. 2. Token-to-Type ratio vs corpus size in words

Heaps' Law Heaps' law is a predictive model for estimating the size of the vocabulary of a given corpus. For a corpus of size N words, with k and β being parameters to the curve, the size of the vocabulary v is estimated to be $k \times N^{\beta}$ [2, 19]. Typically, k will be between 10 and 100, while β will be approximately 0.5 [2]. The vocabulary growth of the corpus can be modelled and then compared with a Heaps' curve of suitable parameters k and β in order to ascertain if the corpus fits the estimated trend.

The corpus's vocabulary growth fits the prediction of Heaps' law very well, as can be seen in Figure 3. The values of K and β are within the typical ranges of $10 \leq k \leq 100$ and $\beta \approx 0.5$ [2].

However, Heaps' law does not predict the growth of the number of unique trigrams in the corpus well. As seen in Figure 4, the growth of the number of unique trigrams does not seem to fit a power graph.

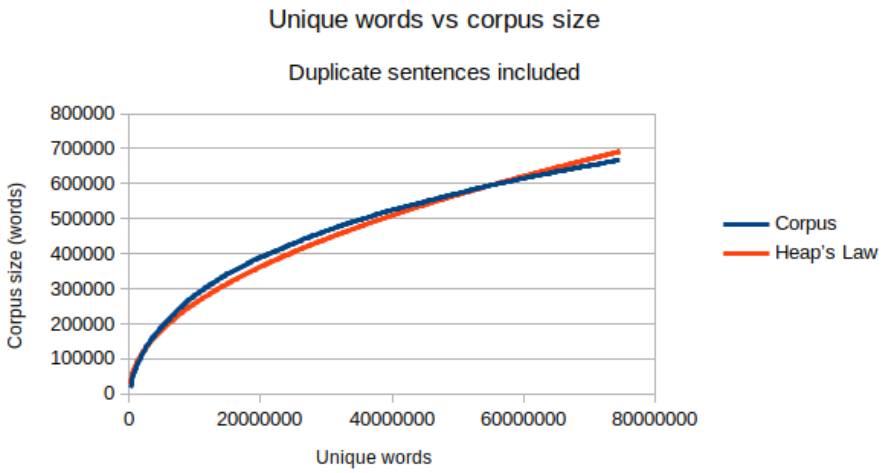

Fig. 3. Heaps' law graph of the corpus for words with $k = 96$ and $\beta = 0.46$

5 Discussion

5.1 Effects of Agglutination

Since isiXhosa is an agglutinative language in which subjects, objects, and verbs can all be combined into a single word, it is possible that Zipf's and Heaps' laws are not good predictors of the statistics of isiXhosa corpora. For instance, in English, the five most common words are "the", "be", "and", "of", and "a" [20]. The articles "the" and "a" are missing from isiXhosa, as it does not use articles. The other words are translated to isiXhosa using grammatical constructs, which are affixed to other words in the sentence. Specifically, "be" is translated using the copulative concord, "of" is translated using the possessive concord, and "and" is translated using the prefix "na", possibly in conjunction with the auxiliaries "kunye" or "kwaye", though it is mostly present on its own. The absence of these common words as separate words could contribute to the distribution of word frequency in the corpus being not perfectly Zipfian.

In order to account for the agglutinativity of isiXhosa, terms may be segmented in other ways, such as by n-grams or morphologically. When the text is decomposed into trigrams instead of words, the distribution is even less Zipfian. This could be due to the fact that there is a finite (and relatively small) number of total possible trigrams. This may therefore contribute to the tail of the distribution being heavier than would be expected for words, which would change the shape of the distribution and may explain why it does not match a Heaps' power law graph. Morphological decomposition may represent a more interesting way to segment the text, but this is still the subject of research in isiXhosa.

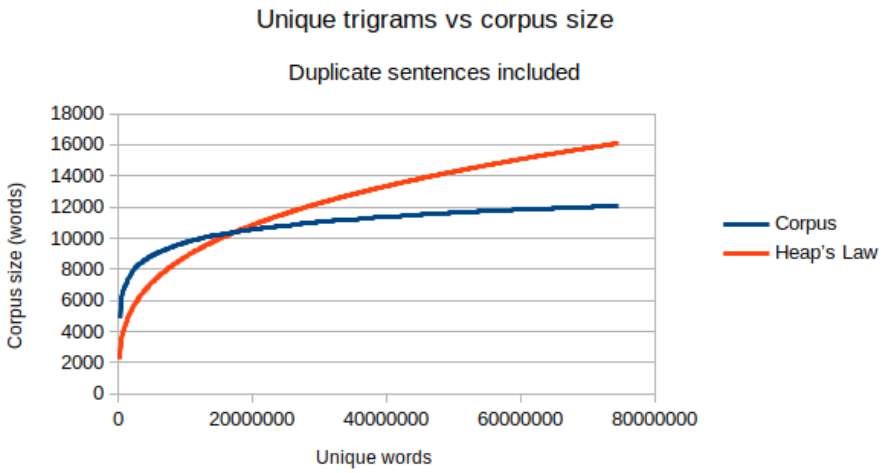

Fig. 4. Heaps' law graph of the corpus for trigrams with $K = 70$ and $\beta = 0.3$

It may be that some of these statistical models do not hold for isiXhosa text. If this is the case, then it is likely that they would also not hold for languages closely related to isiXhosa, such as the Nguni languages isiZulu, siSwati, and isiNdebele. Indeed, Barnard et al. [3] found that, for an isiXhosa corpus of 136 904 words, 29 130 unique words occurred. This represents a Token-to-Type ratio of 4.7. In comparison, when sentences are deduplicated, the corpus obtained through Web crawling had a Token-to-Type ratio of 3.7 at a size of 129 462 words. These values are much closer to each other than the values obtained by Ali et al. [2] for Arabic and English.

5.2 Effect of Seed URLs on Final Crawl

Only four new domains (4.54% of all domains) were discovered that were not included in the list of seed URLs. These domains accounted for only 0.098% of documents in the corpus. One of these domains (jw-cdn.org) was the Content Delivery Network (CDN) for another one of the sites (jw.org). This accounted for 196 documents, whereas the other new domains only accounted for one document each. This further demonstrates that sites on the isiXhosa web are likely to link either to themselves or websites that they are directly affiliated with, but not external, unaffiliated websites.

Therefore, the seed URLs greatly determined which websites ended up being crawled in the end. This suggests that the isiXhosa web is sparse and fairly difficult to discover. This may mean that a focused crawl is not the best approach to discover new websites containing isiXhosa content, although it is a good fit for extracting isiXhosa documents from known hubs of isiXhosa.

6 Conclusions and Future Work

Through the use of a focused Web crawling algorithm, a corpus of isiXhosa documents has been collected, totalling 202 646 documents containing 4 million sentences, 1 million of which are unique. The corpus matches various statistical models, such as Heaps' law and Zipf's law. It should, however, be noted that some of the analysis methods may not be suited to isiXhosa, given that it is an agglutinative language. The websites crawled were also almost entirely determined by the list of seed URLs, which suggests that focused Web crawling may be a good strategy to extract content from known isiXhosa websites, but is not a good strategy for discovering new websites that contain isiXhosa content.

Future work may investigate alternative statistical distributions to better fit isiXhosa data. This could then be extended to other isiXhosa corpora and corpora in related languages, such as isiZulu, to explore if these characteristics are unique to isiXhosa or if they apply to languages similar in grammatical structure and lexicon. Additionally, future work may attempt to verify analytically how well the corpus fits Zipfs' and Heaps' laws if it is large enough and sufficiently similar in characteristics to other corpora.

Some of the deviations from Zipf's and Heaps' laws may be due to the agglutinative nature of the isiXhosa language, as has been discussed. Therefore, future work may aim to segment the corpus morphologically and then ascertain whether it fits these models more closely.

Since the seed URLs made up 95.45% of all sites in the corpus, the seeding approach may be refined in future to yield better results for future crawls. While the WikiMedia projects' pages were excluded from this research, future work could investigate the possibility of using them to assist in generating the list of seed URLs.

Acknowledgements This research was partially funded by the National Research Foundation of South Africa (Grant number: 129253) and University of Cape Town. The authors acknowledge that opinions, findings and conclusions or recommendations expressed in this publication are that of the authors, and that the NRF accepts no liability whatsoever in this regard.

References

1. Aikhenvald, A.Y.: Typological distinctions in word-formation. In: Shopen, T. (ed.) Language Typology and Syntactic Description, vol. 3, p. 1–65. Cambridge University Press, 2 edn. (2007). https://doi.org/10.1017/CBO9780511618437.001
2. Ali, M., Mohammed, Suleman, H.: Building a Multilingual and Mixed Arabic-English Corpus. In: Proceedings of Arabic Language Technology International Conference (ALTIC). Alexandria, Egypt (2011)
3. Barnard, E., Davel, M., van Heerden, C., Wet, F., Badenhorst, J.: The NCHLT speech corpus of the South African languages pp. 194–200 (Jan 2014)
4. Bird, S., Klein, E., Loper, E.: Natural language processing with Python: analyzing text with the natural language toolkit. " O'Reilly Media, Inc." (2009)

5. Cavnar, W., Trenkle, J.: N-gram-based text categorization. Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval (05 2001)
6. Census in brief. Statistics South Africa, Pretoria (2011), https://www.statssa.gov.za/census/census_2011/census_products/Census_2011_-Census_in_brief.pdf
7. Chakrabarti, S., Berg, M., Dom, B.: Focused crawling: A new approach to topic-specific Web resource discovery. Computer Networks **31**, 1623–1640 (Apr 2000). https://doi.org/10.1016/S1389-1286(99)00052-3
8. Dube, M., Suleman, H.: Language Identification for South African Bantu Languages Using Rank Order Statistics. In: Jatowt, A., Maeda, A., Syn, S.Y. (eds.) Digital Libraries at the Crossroads of Digital Information for the Future. pp. 283–289. Springer International Publishing, Cham (2019). https://doi.org/10.1007/978-3-030-34058-2_26
9. Gaustad, T., Puttkammer, M.J.: Linguistically annotated dataset for four official South African languages with a conjunctive orthography: IsiNdebele, isiXhosa, isiZulu, and Siswati. Data in Brief **41**, 107994 (Apr 2022). https://doi.org/10.1016/j.dib.2022.107994, https://www.sciencedirect.com/science/article/pii/S2352340922002050
10. Kiss, T., Strunk, J.: Unsupervised Multilingual Sentence Boundary Detection. Computational Linguistics **32**(4), 485–525 (2006). https://doi.org/10.1162/coli.2006.32.4.485, https://aclanthology.org/J06-4003
11. Kyeyune, M.J.: IsiXhosa Search Engine Development Report. Technical report CS15-01-00, University of Cape Town (2015), https://pubs.cs.uct.ac.za/id/eprint/1035/
12. Lecluze, C., Rigouste, L., Giguet, E., Lucas, N.: Which granularity to bootstrap a multilingual method of document alignment: Character n-grams or word n-grams? Procedia - Social and Behavioral Sciences **95**, 473–481 (10 2013). https://doi.org/10.1016/j.sbspro.2013.10.671
13. Linder, L., Jungo, M., Hennebert, J., Musat, C.C., Fischer, A.: Automatic Creation of Text Corpora for Low-Resource Languages from the Internet: The Case of Swiss German. In: Proceedings of the 12th Language Resources and Evaluation Conference. pp. 2706–2711. European Language Resources Association, Marseille, France (May 2020), https://aclanthology.org/2020.lrec-1.329
14. Medelyan, O., Schulz, S., Paetzold, J., Poprat, M., Markó, K.: Language Specific and Topic Focused Web Crawling. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06). pp. 865–868. European Language Resources Association (ELRA), Genoa, Italy (May 2006), http://www.lrec-conf.org/proceedings/lrec2006/pdf/228_pdf.pdf
15. Piantadosi, S.T.: Zipf's word frequency law in natural language: a critical review and future directions. Psychon. Bull. Rev. **21**(5), 1112–1130 (Oct 2014)
16. Puttkammer, M., Hocking, J., Eiselen, R.: NCHLT South African Language Identifier (2016), https://repo.sadilar.org/handle/20.500.12185/350, accepted: 2018-02-05T20:22:40Z Publisher: North-West University
17. Sims, A., Haspelmath, M.: Understanding Morphology. Routledge, 2 edn. (05 2010). https://doi.org/10.4324/9780203776506
18. Tahir, B., Mehmood, M.A.: Corpulyzer: A Novel Framework for Building Low Resource Language Corpora. IEEE Access **9**, 8546–8563 (2021). https://doi.org/10.1109/ACCESS.2021.3049793

19. van Leijenhorst, D., van der Weide, T.: A formal derivation of heaps' law. Information Sciences **170**(2), 263–272 (2005). https://doi.org/https://doi.org/10.1016/j.ins.2004.03.006, https://www.sciencedirect.com/science/article/pii/S0020025504000696
20. Zhukovskyi, S.: Word Frequency List of American English (2010), https://www.academia.edu/29501273/Word_Frequency_List_of_American_English