An exploration of the computational identification of English loan words in Sesotho

Mmasibidi Setaka
South African Centre
for Digital Language Resources
North-West University
Potchefstroom
South Africa
Mmasibidi.Setaka@nwu.ac.za

Menno van Zaanen
South African Centre
for Digital Language Resources
North-West University
Potchefstroom
South Africa
Menno.vanZaanen@nwu.ac.za

Abstract

South Africa, with its twelve official languages, is an inherently multilingual country. As such, speakers of many of the languages have been in direct contact. This has led to a cross-over of words and phrases between languages. In this article, we provide a methodology to identify words that are (potentially) borrowed from another language. We test our approach by trying to identify words that moved from English into Sesotho (or potentially the other way around). To do this, we start with a bilingual Sesotho-English dictionary (Bukantswe). We then develop a lexicographic comparison method that takes a pair of lexical items (English and Sesotho) and computes a range of distance metrics. These distance metrics are applied to the raw words (i.e., comparing orthography), but using the Soundex algorithm, an approximate phonological comparison can be made as well. Unfortunately, Bukantswe does not contain complete annotation of loan words, so a quantitative evaluation is not currently possible. We provide a qualitative analysis of the results, which shows that many loan words can be found, but in some cases lexical items that have a high similarity are not loan words. We discuss different situations related to the influence of orthography, phonology, syllable structure, and morphology. The approach itself is language independent, so it can also be applied to other language pairs, e.g., Afrikaans and Sesotho, or more related languages, such as isiXhosa and isiZulu.

1 Introduction

In multilingual environments, speakers from different languages are regularly in direct contact with each other, which often influences their languages. In a multilingual country such as South Africa, where there are twelve official languages (Afrikaans, English, Sepedi, Sesotho, Setswana, Siswati, Tshivenḓa, Xitsonga, isiNdebele, isiXhosa, isiZulu, and, recently added, South African Sign language) and several other non-official languages, language change through a process of borrowing is inevitable (Campbell, 2013).

Language contact can have a wide range of influences on the languages. For example, borrowing of various aspects such as lexical items, morphological, or phonological aspects (Hock, 1991), but also code switching or code mixing are possibilities. Of these aspects, borrowing lexical items typically happens most easily (Bergerson, 2011). The degree to which lexical items of one language can be adapted is influenced by various factors, such as social status of the language and usability in daily life (among others).

Given that borrowing of lexical items occurs most frequently, it will be useful to have an automatic, objective way of identifying items that may have been borrowed. Although it may be impossible to decide in an automated way whether lexical items are borrowed or not, or from which language the items are borrowed, having an automated way to identify potentially borrowed lexical items will greatly help with the identification of such items.

As borrowed lexical items show similarities on the orthographic or phonemic level, it may be possible to identify them using a computational similarity measure. We take a similar approach as that of computing text similarity. Whereas text similarity measures compute similarity between texts (see, for instance, Majumdar (2025)) by considering lexical overlap between two texts, we are interested in the overlap between lexical items. In other words, we will use similarity measures to compute the overlap of letters between two lexical items.

If we want to identify the extend of lexical borrowing between two languages, we need to compare the lexicons of the languages. Ideally, we would like to know which lexical items in the two languages are related to each other (most likely semantically as loan words typically have a very similar meaning). For this, a bilingual dictionary is useful, because the lexical items in the two languages are linked based on their meaning. In this research, we use the Sesotho-English Bukantswe online dictionary (Olivier, 2016), which contains over 10,000 entries. Lexical similarity between pairs of Sesotho-English lexical items can then be computed using a range of similarities measures.

The aim of this article is to investigate the feasibility of identifying loan words from English into Sesotho using computational means. Even though we do not expect the computational approach to provide final decisions on whether lexical items are loan words (as that most likely requires further linguistic investigation), the approach should help identify potential loan words, which will allow for a more efficient approach to a more complete investigation of loan words from English into Sesotho.

This research takes place in the context of the South African languages, which, given the intense interaction between speakers of the different languages, form a perfect environment for the investigation into language contact and language change. Having automated ways to help identify language change, such as lexical borrowing or loan words, will allow for a more focused and time-efficient research on these matters. For instance, Zulu (2013) mentions that limited work has been done on distance measures and this is even more pressing for the South African languages. Even though the focus here is on Sesotho-English interaction, this work can also be applied to other (South African) language pairs.

The article is structured as follows. We start with a short background on language contact and lexical similarity. Next, we provide the methodology where we discuss the data collection used as well as the proposed metrics to compute lexical similarity. We then apply these metrics to the data collection and provide an overview of the results. This is followed by a discussion of the results and finally we provide a conclusion.

2 Background

In this article, we will look at the identification of loan words in Sesotho from English. Sesotho is one of the official languages of South Africa, Lesotho, Zimbabwe, Zambia, and Botswana and is closely related to Sepedi (Northern Sotho) and Setswana. In contrast, English, a part of the Germanic (Gmc) language group, is an official language in many countries, including South Africa, Lesotho, and many others. English is considered a global language (Eventhough, 2012; El Garras, 2025; Mnguni, 2021; Kamwangamalu and Moyo, 2003; Kula and Marten, 2008).

Sesotho and English have been in close contact in South Africa for a while. In South Africa, English was introduced during the 19th century, and by the end of the century, it was used in many black communities as well (Silva, 1997). Against this background, language contact between English and Sesotho was unavoidable. Due to the power dynamics between the two languages, Basotho (Sesotho-speaking communities) had to learn English, which led to Sesotho borrowing words from English. Additionally, modifications (phonological, morphological, grammatical) of English loan words took place to properly fit the English loan words in Sesotho (Kunene, 1963). Due to the nature of the relationship between English and Sesotho, it was mostly Sesotho borrowing from English (and not vice versa). Note, that similar processes took place for Afrikaans (Majola and Lemeko, 2024).

In *Southern Sotho Words of English and Afrikaans Origin*, Kunene (1963) investigates, compares, and discusses methods of adapting words of English and Afrikaans origin to the grammatical and phonemic structures of Southern Sotho (Sesotho). A number of examples are provided that illustrate the influence of English (and Afrikaans) on the Sesotho lexicon. This shows the extensive lexical borrowing between the languages even though the languages are very different.

The examples in Kunene (1963) are hand picked to illustrate the different types of adaptations of loan words in Sesotho. Such an approach is suitable for describing the variety of loan words and their modifications, but it does not allow for a more quantitative analysis of loan words. For this, a most automated way to identify possible loan words is required.

2.1 Language Contact

Language contact is a phenomenon that occurs when speakers of different languages interact and their languages influence one another (Matras, 2020). It may lead to language change, for instance, through the borrowing of words, but it may even lead to new languages. In general, the interactions have the ability to influence languages as we known them, resulting in observable additions, subtractions, changes, and similarities in their lexicons. As such, information on how the interaction between the languages took place may also teach us about the history of these languages.

Historically, in Africa, Southeast Asia, and Oceania, Germanic languages came into contact with other languages, particularly as a result of colonization and trading activities. Riehl (2025) has seen many languages adapting and adopting certain sections of one language and then assimilating it into their own. In the context of the influence of the Germanic languages on South African languages, it is useful to have tools that can help identify these influences (semi-)automatically.

2.2 Language similarity

There are several ways to measure the impact of language contact on the different languages. With language contact come different types of similarities and measurements, which can be used on languages, such as lexical similarities, dialectometry, lexicostatistics, and many others. For the sake of this article, we will briefly share a few of them.

Lexical similarity is a natural language processing concept that focuses on similarities that exist between words and texts of different languages. According to Ethnologue[1], it is the percentage of lexical similarity between two linguistic varieties, which is determined by comparing a set of standardized word lists and counting those forms that show similarity in both form and meaning. This is particularly interesting for many low-resource languages (such as Sesotho) that share certain lexical similarities with some high-resource languages (such as English) (Maurya et al., 2023).

Many systems have been proposed to build resources automatically based on lexical similarity (Jadi et al., 2016). To provide a few examples of previous research in the area of lexical similarities, we mention work comparing German and Afrikaans (Bergerson, 2011), Khelobedu, Tshivenḓa, and Sepedi languages (Tebogo and Mandende, 2023), English, Sesotho, and Afrikaans (Kunene, 1963), English and Portuguese (Günther et al., 2019), and English and German (García and de Souza, 2014).

Related work can also be found in the field of dialectometry (which aims to measure the distance between dialects of a language). Dialects of a language are often mutually intelligible, meaning that speakers of one dialect can understand speakers of the other dialect (Gooskens and Van Heuven, 2022). This means that the research in this field can focus on the linguistic variations between the dialects. The lexical variation is often clear between dialects and can be measured. As such, there has been a wide range of work in this field, see for example, Heeringa and Nerbonne (2001); Nerbonne and Kretzschmar (2003); Nerbonne and Kretzschmar Jr (2013); Wieling and Nerbonne (2015) for an overview of the field. If we look at such research on the official South African languages, we find that some work has been done comparing isiZulu and isiXhosa (Zulu, 2013).

Phrasal similarities check the extent to which phrases in different languages are similar. Various data models have been developed such as the Document Index Graph, which indexes web documents based on phrases, rather than single terms only (Hammouda and Kamel, 2002). As such, this measure calculates the similarities based on the ratio of how much they overlap and rewards phrases with high frequency, significance level and length (Momin et al., 2006).

Semantic similarity is a relation between concepts on the level of meaning (Kolb, 2009). In semantic similarity, the likelihood of occurrence of a concept is determined by its frequency in a corpus (Li et al., 2003). Though it has been a part of natural language processing and information retrieval discussions for many years, it is has been a problem for many applications of computational linguistics and artificial intelligence (Li et al., 2003).

On a more general level, the field of lexicostatistics measures the similarity or difference between languages which shows a degree of relatedness and this provides a value that describes the distance between the languages. Based on these distances, family tries of languages can be constructed, which indicate relatedness between languages and may even be used to create proto-languages, which indicate historical relationships (Dyen, 1964).

Throughout the previous research described in this section, several computational techniques are used regularly. In particular, measures that describe the similarity or distance between lexical items, such as the edit-distance or Levenshtein distance (Wagner and Fischer, 1974) or Damerau distance (Damerau, 1964), are used. Here we also use the Levenshtein distance, but also consider a number of additional metrics coming from the field of in-

[1] See https://www.ethnologue.com/methodology/.

formation retrieval, where such measures are used to evaluate the set of documents found by the information retrieval system.

3 Methodology

To identify which English words may have influenced or are incorporated in Sesotho, we need access to lexical items in both languages. Ideally, the lexical items should be translations of each other. Bilingual dictionaries offer exactly this: the lexical items in the dictionary are paired based on their meaning. Given pairs of lexical items, we can then investigate whether the lexical items in both languages are similar enough.

Starting with a pair of lexical items we need to decide if the words are similar enough that they could be related. To decide whether a word may be a loan word, we assume that, in such a case, the words should be orthographically or phonologically similar (and at the same time that words that are not loan words are orthographically and phonologically dissimilar). We implement a range of distance and similarity metrics to help identify the loan word pairs.

Even though the methodology proposed is language independent, in this article, we will focus on Sesotho-English word pairs. We already know that Sesotho has borrowed from English as several lexical items are marked as such in the data collection that we are using (see below). Note that in this article, we assume that there are words that were incorporated into Sesotho from English, but it may also be the case that some Sesotho words have been taken over in (South African) English. This method does not identify the direction.

3.1 Data collection

To investigate which words are (potentially) loan words, we require lexical items in the two languages. These lexical items could be extracted from (large) corpora in the languages, which can provide lists of lexical items. However, in such a case it is unclear which words in the two languages have similar meaning. (We assume that loan words have similar meaning.) Aligning words with similar meaning in two languages is challenging. However, bilingual dictionaries provide lexical items in two languages where the lexical items are directly linked to each other based on their meaning.

Here, we make use of the Bukantswe Sesotho-English Bilingual Dictionary, which is developed by Olivier, Masilo, and Thejane and is publicly available[2]. In this data collection, each line contains a pair of lexical items (separated by a tab) with the Sesotho lexical item in the first column and corresponding English in the second. The English column sometimes also contains some additional information in brackets, such as part of speech (e.g., adj. v., n.) and whether the Sesotho lexical item is borrowed from English (<Eng) or Afrikaans (<Afr). In total, the dictionary contains 10,084 entries, but some of the entries contain multiple (comma separated) lexical items. After splitting these items and removing duplicate entries, the data collection contains 12,958 entries.

3.2 Metrics

To compare how similar two lexical items are, we compute their distance. There are many metrics that can be used for this purpose. Table 1 provides an overview of the metrics used in this article. Note that the metrics used here focus on the orthography (and through Soundex on the pronunciation). For example, metrics that take into account contexts of the use of the lexical item, such as word vector models (Turney and Pantel, 2010), are not used here. Several of the used metrics come from the field of information retrieval (Manning et al., 2008) to compare search results of different search systems. In Table 1 we provide the name of the metric, the implementation, whether the metric is a distance or similarity metric, whether the metric takes the ordering of letters in a word into account, and the range of the values of the metric.

To compute word similarity or distance[3], we compare words by looking at their individual letters. A relatively simple metric to compute letter overlap is the Jaccard metric. This metric does not take the order of letters into account. The formula is applied after the lexical items are converted into a bag (which allows for letters to occur multiple times), which is slightly different from the regular definition of the Jaccard metric, which works on sets. Similarly, the Euclidean and Cosine metrics are computed based on vectors that describe the bag of letters of the lexical items. This essentially compares the number of occurrences of the letters

[2]This resource can be downloaded from SADiLaR's repository: `https://hdl.handle.net/20.500.12185/419`.

[3]Similarity metrics are have high values when words are similar and low when words are dissimilar, whereas distance metrics have low values when words are similar and vice versa.

in the lexical items one by one. The Hamming distance, Levenshtein distance, Levenshtein ratio, and the Jaro metric do take the order of the letters of the words into account. The Hamming distance requires lexical items of the same length, which is attained by padding the shorter lexical item. For the Levenshtein distance and ratio, we use the standard weights for the edit operations (one for insertion, deletion, and substitution).

Several other metrics could be implemented and experimented with as well, such as the Damerau–Levenshtein distance (Damerau, 1964) or the Jaro-Winkler metric (Winkler, 1990), just like a number of variations of the current metrics (e.g., the standard set-based Jaccard metric, different weights for the Levenshtein distance). However, due to space constraints we will focus on the metrics in Table 1 only here.

The metrics described so far are applied to the orthographic representation of the lexical items. However, some words may have a similar pronunciation even if their orthographic representation is different. The Soundex algorithm (Knuth, 1998) maps phonetically similar orthographic letters onto common symbols. (Note, however, that Soundex is developed for English.) The metrics described in Table 1 can be applied to the Soundex converted sequences as well.

The different metrics and Soundex conversion are implemented in Python. We used the Levenshtein[4] package for most of the different metrics, and the Soundex[5] package to map orthography to Soundex sequences.

4 Results

For all the lexical items that we find in the Bukantswe bilingual dictionary, we compute the distance metrics as mentioned in Table 1. We first look at the correlations between the different metrics to understand the general behavior of the different metrics with respect to each other.

Following that, we look at a sample of the lexical items in the dictionary. We can try to identify loan words by sorting the items based on the computed distance values. The assumption is that lexical items that have a low distance are very similar (in orthography or in phonetic representation in the Soundex case). These words are possibly lexical items that have gone from one language to the other. We provide examples where indeed loan words are identified, but also cases where the metrics do not provide the right information. Unfortunately, the information in Bukantswe on whether lexical items come from English is not complete, so a fully automated evaluation is currently not possible.

4.1 Correlations

To investigate how the different metrics behave, we first investigate the correlations between the calculated values. In total we implemented seven metrics (Jaccard, Euclidean, Cosine, Levenshtein, Levenshtein ratio, Hamming, and Jaro), but these are also applied to the Soundex representation, so in total we have fourteen variants to compare.

Creating a correlation matrix shows that all correlations are significant ($\alpha = .05$), except for the Levenshtein and Cosine metrics ($p = .058$) and Hamming and Jaro ($p = .072$). The correlation matrix (shown in Table 2) illustrates that there is quite some variation between the results of the different metrics. Depending on the metrics, the correlation can range from very weak (e.g., Levenshtein versus Cosine with $r = -.02$) to very strong (e.g., Levenshtein versus Hamming with $r = .97$ and $r = .94$ with Soundex, Jaccard versus Cosine with $r = .92$ and $r = .93$ with Soundex, or Jaccard versus Levenshtein ratio with $r = .92$ and $r = .95$ with Soundex). This means that we should consider the effectiveness of the different metrics separately.

4.2 Analysis of lexical items

As mentioned before, the information on English loan words in Sesotho as marked in Bukantswe is not complete. As such, we cannot perform a computational evaluation of the entire data collection. Here we will look at a sample of the lexical items that show low distance (based on a few metrics).

Words that are very different, e.g., “bonolo” versus “ease” get the lowest value for the similarity metrics (0), and high values (e.g., 4.24 for Euclidean, 6 for Hamming and Levenshtein) for the distance metrics. Similar values are found for the Soundex versions of the metrics (2.24 for Euclidean, 3 for Hamming and Levenshtein Soundex). In contrast, words that are exactly the same, e.g., “radar” have the highest value for similarity metrics (1) and the lowest value for distance metrics (0). The Soundex versions are exactly the same for these.

[4] See https://pypi.org/project/python-Levenshtein/.

[5] See https://pypi.org/project/soundex/.

Table 1: Overview of metrics to compute lexical distance. Where A and B are bags of letters of the lexical items, whereas $\vec{A}$ and $\vec{B}$ describe sequences of letters. We provide the implementation, type (distance or similarity), whether the order of letters is taken into account and the range of the values.

Metric	Implementation	Type	Order of letters	Range
Jaccard	$\frac{A \cap B}{A \cup B}$	Similarity	Unordered	$[0, 1]$
Euclidean	$\sqrt{\Sigma_{i=A \cup B}(A_i - B_i)^2}$	Distance	Unordered	$[0, \infty)$
Cosine	$\frac{A \cdot B}{\|\|A\|\|\,\|\|B\|\|}$	Similarity	Unordered	$[0, 1]$
Hamming	Hamming (1950)	Distance	Ordered	$[0, \infty)$
Levenshtein	Wagner and Fischer (1974)	Distance	Ordered	$[0, \infty)$
Levenshtein ratio	$1 - \frac{Levenshtein(\vec{A}, \vec{B})}{\|\vec{A}\| + \|\vec{B}\|}$	Similarity	Ordered	$[0, 1]$
Jaro	Jaro (1989)	Similarity	Ordered	$[0, 1]$

For words like "diesel" (English) and "disele" (Sesotho), which have the same letters, but in a different order, we see that for the unordered metrics, the values are optimal, but, as expected, the ordered metrics show slight variation (4 for Hamming, 2 for Levenshtein, 0.83 for Levenshtein ratio, and 0.89 for Jaro).

The Soundex algorithm is effective in words such as "afrika" (Sesotho) versus "africa" (English). The orthography is very similar (e.g., Jaccard 0.71, Euclidean 1.41, Cosine 0.88, Levenshtein 1, Levenshtein ratio 0.83, Hamming 1), we have optimal scores for the Soundex versions of the metrics as the "k" and "c" are mapped to the same symbol. In situations like these, the Soundex algorithm has a major impact on the identification of similar words. Similar situations happen for words like "leaforikanere" versus "afrikaner", "disetatemente" versus "statement", "faele" versus "file", and "yunifomo" versus "uniform" (all Sesotho versus English words).

Given these results, it seems that the metrics work well, with the Soundex having advantages where the orthography is slightly different. However, there are also situations where the approach may not work. For example, according to the Bukantswe dictionary, "sekete" or "dikete" (Sesotho) comes from "circuit" (English). If we look at the metrics, we find low scores for the similarity (0.08 Jaccard, 0.09 Cosine, 0.15 Levenshtein ratio, and 0.44 Jaro) and high values for distance (4.58 Euclidean, 7 Hamming and Levenshtein) with somewhat better scores for the Soundex variants (0.4 Jaccard, 0.58 Cosine, 0.57 Levenshtein ratio, and 0.72 Jaro, 1.73 Euclidean, 4 Hamming and 2 Levenshtein) with slighly higher values for "dikete".

Looking at the words that have a high similarity (or low distance), especially using the Soundex metrics, we can identify a number of reasons for differences in the words. First, there are relatively simple orthographic differences. For example, "histori" (Sesotho) versus "history" (English), or "mengo" (Sesotho) versus "mango" (English). Second, there are words with orthographic changes based on the phonology or the structure of Sesotho syllables in words such as "boroto" (Sesotho) versus "board" (English), "keresemese" (Sesotho) versus "christmas" (English), or "potefoliyo" (Sesotho) versus "portfolio" (English). Third, there are morphological influences in cases like "dibanana" (Sesotho) versus "banana" (English), or "ditonki" (Sesotho) versus "donkey" (English). (Note that some words may show a combination of these, such as "okesejene" (Sesotho) from "oxygene" (English) as Sesotho does not use "x" in its orthography and the "kese" is added due to syllable structure.)

Whereas many "modern" words (like "visa", "radio", "bikini", "sms") are borrowed from English as well as names of countries (e.g., "Zambia", "Zimbabwe"), we also see names that may have been borrowed by English (e.g., "Kgalahadi" (Sesotho) and "Kalahari" (English).

Unfortunately, there are also a number of words that have a high similarity (or low distance) even though the Sesotho words do not come from the English words. For example, "ba batle" (Sesotho) versus "beautiful" with (all Soundex metrics) a Jaccard and Cosine values of 1, and Levenshtein value of 2, Levenshtein ratio of 0.75 and Jaro of 0.92. Similarly, it is unlikely that “nko” (Sesotho) comes from the English word “nose” (English), similarly for “amohetse” (Sesotho) versus “accommodated” (English).

We also see a number of words that have a high similarity and low distance that most likely come from Afrikaans, although the English word is similar as well. Examples of these are “tanki” (Sesotho) versus “thanks” (English) and “dankie” (Afrikaans), or “sekere” (Sesotho), “shears” (English) and “skêr” (Afrikaans).

Table 2: Correlation matrix (r) of all metrics.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Jaccard (1)	1.00	-.42	.92	-.18	.92	-.05	.75	.71	-.46	.66	-.30	.68	-.18	.56
Euclidean (2)	-.42	1.00	-.29	.90	-.44	.85	-.24	-.23	.71	-.06	.74	-.22	.73	-.05
Cosine (3)	.92	-.29	1.00	-.02	.85	.12	.77	.62	-.30	.63	-.13	.61	-.02	.53
Levenshtein (4)	-.18	.90	-.02	1.00	-.28	.97	-.10	-.14	.72	.06	.83	-.14	.83	.03
Levenshtein ratio (5)	.92	-.44	.85	-.28	1.00	-.13	.79	.71	-.50	.64	-.38	.71	-.26	.56
Hamming (6)	-.05	.85	.12	.97	-.13	1.00	.02	-.05	.66	.15	.78	-.04	.82	.11
Jaro (7)	.75	-.24	.77	-.10	.79	.02	1.00	.58	-.30	.56	-.19	.59	-.09	.52
SOUNDEX														
Jaccard (8)	.71	-.23	.62	-.14	.71	-.05	.58	1.00	-.64	.93	-.41	.95	-.28	.79
Euclidean (9)	-.46	.71	-.30	.72	-.50	.66	-.30	-.64	1.00	-.47	.90	-.60	.83	-.36
Cosine (10)	.66	-.06	.63	.06	.64	.15	.56	.93	-.47	1.00	-.22	.93	-.06	.84
Levenshtein (11)	-.30	.74	-.13	.83	-.38	.78	-.19	-.41	.90	-.22	1.00	-.43	.94	-.22
Levenshtein ratio (12)	.68	-.22	.61	-.14	.71	-.04	.59	.95	-.60	.93	-.43	1.00	-.27	.84
Hamming (13)	-.18	.73	-.02	.83	-.26	.82	-.09	-.28	.83	-.06	.94	-.27	1.00	-.10
Jaro (14)	.56	-.05	.53	.03	.56	.11	.52	.79	-.36	.84	-.22	.84	-.10	1.00

5 Discussions

This article provides an automatic way to identify similar words in bilingual word lists. The underlying assumption is that such words are likely to represent words that are borrowed from one language to the other. Different metrics have been proposed that rely on orthographic or (through Soundex) phonologic similarities. These metrics have been applied and investigated on Bukantswe, a Sesotho-English dictionary.

This approach illustrated that many loan words can automatically be identified. The data collection contains a number of word pairs where the words are exactly the same, but also many words that are similar in the sense that orthographic, phonologic or morphologic differences can be filtered out.

This research extends the work by Kunene (1963) in the sense that the proposed approach can (semi-)automatically identify words that are borrowed from another language. Whereas Kunene (1963) focused on Sesotho words from either English or Afrikaans, we have focused on English only, although (as illustrated in Section 4.2) some loan words of Afrikaans origin were also identified. Since some Afrikaans and English words are quite similar (due to Afrikaans and English being members of the Germanic language group), it is not always clear what the original language is.

The work by Kunene (1963) also shows that a human, manual analysis of the results is still needed. Even though the proposed techniques identify similar words, it does not explicitly indicate borrowings. Furthermore, even in the case of borrowings, it does not indicate the direction of borrowing. While for the Sesotho-English case (in particular for “modern” words) this is often clear, when applying these techniques to other language pairs (e.g., isiZulu-isiXhosa), the picture may be much more difficult.

The results also indicated that there are some structural differences as Sesotho has more complex morphological structures when compared to the English words. For example, Sesotho's noun class information is encoded as prefixes on nouns (Moloi and Thetso, 2014), which has no clear equivalent in English. These structural differences can potentially be handled in a consistent way. The current metrics do not take this explicitly into account. Such an approach would make the method language dependent, however.

6 Conclusion

In this article, we have described an approach to identify a computational approach to identify lexical items that are similar in two languages. We have applied this to lexical items from a Sesotho-English bilingual dictionary with this assumption that such similar words are most likely words that have been incorporated in Sesotho from English.

We have implemented seven distance metrics with different properties (such as taking the order of letters in the word into account or not) and we have applied the same metrics to a Soundex representation of the lexical items. The Soundex representation is an approximate phonological representation, contrasting from the orthographic dictionary entries. The metrics showed varied correlations, which indicates that they do describe alternative distances.

Unfortunately, we could not find a data collection that provides information on which words are actually incorporated from English into Sesotho, so a structural, computational analysis of the different metrics was not possible. However, the metrics did help to identify words that are quite similar, which may potentially save much time when trying to find borrowed words. Further (linguistic, historic) research needs to take place to properly identify those words that are incorporated into Sesotho from English.

The computational approach described in this article shows promising results, but a more controlled evaluation is needed. This, however, requires annotated data (e.g., containing word pairs in addition to a label that indicates whether that the lexical items are related).

The currently investigated metrics do not take any morphological information into account. The Sesotho lexical items show a different morphological structure, for instance, including morphemes that describe the noun class. English and Sesotho morphological structure is different, leading to (almost) regular differences which could be implemented as part of the distance metrics.

Limitations

The research in this article has a number of limitations. First, the data collection used (Bukantswe's bilingual Sesotho-English dictionary) contains not only words, but also some multi-word expressions. Even though these are interesting to investigate, comparing these as lexical items is challenging as some words may be loan words, but the entire expression does not necessarily have to.

Second, Bukantswe does not have complete tagging of English loan words. Even though some lexical items have a tag indicating this, there are many lexical items that are loan words from English without the tag. Complete information allows for a more robust evaluation of the methodology used.

Third, we assume that if the lexical items are similar enough (according to a particular metric), there is a relation between the words. As the lexical items are orthographically similar (or potentially similar on the phonologic level when using Soundex), this does not necessarily mean that these words are loan words. Even if they are, the direction of the relationship is not identified. Even though in this case it is more likely that words are taken from English and incorporated in Sesotho, there may be words taken from Sesotho and incorporated in South African English.

References

Jeremy Bergerson. 2011. *Apperception and linguistic contact between German and Afrikaans*. Ph.D. thesis, University of California, Berkeley.

Lyle Campbell. 2013. *Historical linguistics*, 3rd edition. Edinburgh University Press.

Fred J Damerau. 1964. A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7(3):171–176.

Isidore Dyen. 1964. Lexicostatistics in comparative linguistics. *Lingua*, 13:230–239.

Hassan El Garras. 2025. English, the global language: Its strength, status, and future. *International Journal of Linguistics, Literature and Translation*, 8:122–130.

Ndlovu Eventhough. 2012. Mother tongue education in the official minority languages in zimbabwe. *South African Journal of African Languages*, 31:229–242.

Maria Isabel Maldonado García and Ana Maria Borges de Souza. 2014. Lexical similarity level between english and portuguese. *Elia*, 0(14):145.

Charlotte Gooskens and Vincent Van Heuven. 2022. *Mutual intelligibility*, pages 51–95. Cambridge University Press.

Fritz Günther, Eva Smolka, and Marco Marelli. 2019. 'understanding' differs between english and german: Capturing systematic language differences of complex words. *Cortex*, 116:168–175.

Richard W Hamming. 1950. Error detecting and error correcting codes. *The Bell system technical journal*, 29(2):147–160.

Khaled M Hammouda and Mohamed S Kamel. 2002. Phrase-based document similarity based on an index graph model. In *2002 IEEE International Conference on Data Mining, 2002. Proceedings*, pages 203–210. IEEE.

Wilbert Heeringa and John Nerbonne. 2001. Dialect areas and dialect continua. *Language variation and change*, 13(3):375–400.

Hans Henrich Hock. 1991. *Principles of historical linguistics*, 2nd edition. Mouton de Gruyter.

Grégoire Jadi, Vincent Claveau, Béatrice Daille, and Laura Monceaux-Cachard. 2016. Evaluating lexical similarity to build sentiment similarity. In *Language and Resource Conference, LREC*.

Matthew A Jaro. 1989. Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. *Journal of the American Statistical association*, 84(406):414–420.

Nkonko Kamwangamalu and Themba Moyo. 2003. Some characteristic features of englishes in lesotho, malawi and swaziland. *Per Linguam: a Journal of Language Learning= Per Linguam: Tydskrif vir Taalaanleer*, 19(1_2):39–54.

Donald E. Knuth. 1998. *The Art of Computer Programming—Sorting and Searching*, 2nd edition, volume 3. Addison-Wesley Publishing Company, Reading, MA, USA.

Peter Kolb. 2009. Experiments on the difference between semantic similarity and relatedness. In *Proceedings of the 17th Nordic conference of computational linguistics (NODALIDA 2009)*, pages 81–88.

Nancy Kula and Lutz Marten. 2008. *Central, East and Southern African Languages*, chapter 4. University of California Press.

DP Kunene. 1963. Southern Sotho words of English and Afrikaans origin. *Word*, 19(3):347–375.

Yuhua Li, Zuhair A Bandar, and David McLean. 2003. An approach for measuring semantic similarity between words using multiple information sources. *IEEE Transactions on knowledge and data engineering*, 15(4):871–882.

Yanga L. P. Majola and Papi Lemeko. 2024. The influence of afrikaans on naming among the basotho of south africa. *Southern African Linguistics and Applied Language Studies*, 42(3):357–365.

Dattatreya Majumdar. 2025. Text analysis and distant reading using r. Https://ladal.edu.au/tutorials/lexsim/lexsim.html.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Infomation Retrieval*. Cambridge University Press, New York.

Yaron Matras. 2020. *Language contact*. Cambridge University Press.

Kaushal Kumar Maurya, Rahul Kejriwal, Maunendra Sankar Desarkar, and Anoop Kunchukuttan. 2023. Charspan: Utilizing lexical similarity to enable zero-shot machine translation for extremely low-resource languages. *arXiv preprint arXiv:2305.05214*.

Aaron Mnguni. 2021. Dreams and realities for south africa: Use of official languages act, 2012. *Studies in Media and Communication*, 9:1.

Francina L Moloi and Madira L Thetso. 2014. The morphology of the sesotho form/bo-: An exploratory study. *Journal of Linguistics and Language in Education*, 8(1):65–79.

BF Momin, PJ Kulkarni, and Amol Chaudhari. 2006. Web document clustering using document index graph. In *2006 International Conference on Advanced Computing and Communications*, pages 32–37. IEEE.

John Nerbonne and William Kretzschmar. 2003. Introducing computational techniques in dialectometry. *Computers and the Humanities*, 37(3):245–255.

John Nerbonne and William A Kretzschmar Jr. 2013. Dialectometry++. *Literary and Linguistic Computing*, 28(1):2–12.

J. A. K. Olivier. 2016. Bukantswe Sesotho-English bilingual dictionary. SADiLaR Language Resource Repository, License: Creative Commons Attribution 3.0 South Africa (CC BY 3.0 ZA): https://creativecommons.org/licenses/by/3.0/za/.

Claudia Maria Riehl. 2025. *Germanic Languages in Contact in Africa, Asia, and Oceania*. Oxford University Press.

Penny Silva. 1997. South African English: oppressor or liberator. *The major varieties of English*, 97:1–8.

Rakgogo J Tebogo and Itani P Mandende. 2023. Lexical similarities between Khelobedu dialect and Tshivenda and sepedi languages. *Literator-Journal of Literary Criticism, Comparative Linguistics and Literary Studies*, 44(1):1910.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37:141–188.

Robert A. Wagner and Michael J. Fischer. 1974. The string-to-string correction problem. *Journal of the Association for Computing Machinery*, 21(1):168–173.

Martijn Wieling and John Nerbonne. 2015. Advances in dialectometry. *Annu. Rev. Linguist.*, 1(1):243–264.

William E Winkler. 1990. String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage. In *Proceedings of the Section on Survey Research Methods. American Statistical Association*, pages 354—359.

Peleira Nicholas Zulu. 2013. Classification of south african languages using text and acoustic based methods: A case of six selected languages. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 280–287.