

# **Orthographic measures of language distances between the official South African languages**

P. N. Zulu, G. Botha and E. Barnard

Human Language Technologies Research Group, CSIR and  
Department of Electrical and Computer Engineering  
University of Pretoria  
Pretoria 0001, South Africa.

Tel: 012 841 4627 Fax: 012 841 4720

[pzulu@csir.co.za](mailto:pzulu@csir.co.za), [gbotha@csir.co.za](mailto:gbotha@csir.co.za), [ebarnard@csir.co.za](mailto:ebarnard@csir.co.za)

Word count: 5179

# **Orthographic measures of language distances between the official South African languages**

## **Abstract**

Two methods for objectively measuring similarities and dissimilarities between the 11 official languages of South Africa are described. The first concerns the use of  $n$ -grams. The confusions between different languages in a text-based language identification system can be used to derive information on the relationships between the languages. Our classifier calculates  $n$ -gram statistics from text documents and then uses these statistics as features in classification. We show that the classification results of a validation test can be used as a similarity measure of the relationship between languages. Using the similarity measures, we were able to represent the relationships graphically.

We also apply the Levenshtein distance measure to the orthographic word transcriptions from the 11 South African languages under investigation. Hierarchical clustering of the distances between the different languages shows the relationships between the languages in terms of regional groupings and closeness. Both multidimensional scaling and dendrogram analysis reveal results similar to well-known language groupings, and also suggest a finer level of detail on these relationships.

## **Ortografiese maatstawwe van taalafstande tussen die amptelike Suid Afrikaanse tale**

### **Opsomming**

Twee metodes vir die bepaling van verwantskappe tussen die 11 amptelike tale van Suid Afrika word beskryf. Die eerste metode maak gebruik van  $n$ -gramme. Die verwarrings wat plaasvind in 'n taalherkenningstelsel verskaf inligting oor die verhouding tussen die tale.  $N$ -gram statistieke word bepaal vanaf teksdokumente en word dan gebruik as kenmerke vir klassifikasie. Ons wys dat die uitsette van 'n bevestigingstoets gebruik kan word om te bepaal hoe naby tale aan mekaar lê. Vanuit hierdie metings het ons 'n sigbare voorstelling van die verhouding tussen tale afgelei.

Verder het ons die Levenshtein-metode gebruik om die afstand tussen die ortografiese transkripsies van woorde te bepaal, toepespits op die 11 amptelike tale van Suid Afrika. 'n Grafiese groepering volgens die afstande tussen die verskillende tale toon weer die verhoudings tussen die tale en ook familie-groepe. Met beide dendrogramme en multidimensionele skalering word bepaalde familie-groepe aangedui, en selfs die fynere verwantskappe binne hierdie familie-groepe.

## **1. Introduction**

The development of objective metrics to assess the distances between different languages is of great theoretical and practical importance. Currently, subjective measures have generally been employed to assess the degree of similarity or dissimilarity between different languages (Gooskens and Heeringa, 2004, Van-Hout and Münstermann, 1981, Van-Bezooijen and Heeringa, 2006), and those subjective decisions are, for example, the basis for classifying separate languages, and certain groups of language variants as dialects of one another. It is without doubt that languages are complex; they differ in vocabulary, grammar, writing format, syntax and many other characteristics. This presents levels of difficulty in the construction of objective comparative measures between languages. Even if one intuitively knows for example, that English is closer to French than it is to Chinese, by how much is it closer? Also, what are the objective factors that allow one to assess these levels of distance?

These questions bear substantial similarities to the analogous questions that have been asked about the relationships between different species in the science of cladistics. As in cladistics, the most satisfactory answer would be a direct measure of the amount of time that has elapsed since the languages' first split from their most recent common ancestor. Also, as in cladistics, it is hard to measure this from the available evidence, and various approximate measures have to be employed instead. In the biological case, recent decades have seen tremendous improvements in the accuracy of biological measurements as it has become possible to measure differences between DNA sequences. In linguistics, the analogue of DNA measurements is historical information on the evolution of languages, and the more easily measured, though

indirect measurements (akin to the biological phenotype) are either the *textual* or *acoustic* representations of the languages in question.

In the current article, we focus on distance measures derived from text; we apply two different techniques, namely language confusability based on  $n$ -gram statistics and the Levenshtein distance between orthographic word transcriptions, in order to obtain measures of dissimilarity amongst a set of languages. These methods are used to obtain language groupings, which are represented graphically using two standard statistical techniques (dendrograms and multi-dimensional scaling). This allows us to assess the methods relative to known linguistic facts in order to assess their relative reliability.

Our evaluation is based on the 11 official languages of South Africa. These languages fall into two distinct groups, namely the Germanic group (represented by English and Afrikaans) and the South African Bantu languages, which belong to the South Eastern Bantu group. The South African Bantu languages can further be classified in terms of different sub-groupings: Nguni (consisting of Zulu, Xhosa, Ndebele and Swati), Sotho (consisting of Southern Sotho, Northern Sotho and Tswana), and a pair that falls outside these sub-families (Tsonga and Venda).

We believe that an understanding of these language distances is of inherent interest, but also of great practical importance. For purposes such as language learning, the selection of target languages for various resources, and the development of Human Language Technologies, reliable knowledge of language distances would be of great value. Consider, for example, the common situation of an organization that wishes to publish information relevant to a particular multi-lingual community, but with insufficient funding to do so in all the languages of that community. Such an organization can be guided by knowledge of language distances to make an appropriate choice of publication languages.

The following sections describe in more detail  $n$ -grams and Levenshtein distance. Thereafter we present an evaluation on the 11 official languages of South Africa, highlighting language groupings and proximity patterns. We close with a discussion of the results, interesting directions and a brief summary.

## 2. Theoretical background

Orthographic transcriptions are one of the most basic types of annotation used for speech transcription. Orthographic transcriptions of speech are important in most fields of research concerned with spoken language. The orthography of a language refers to the set symbols used to write a language and includes the writing system of a language. English, for example, has an alphabet of 26 letters for both consonants and vowels. However, each English letter may represent more than one phoneme, and each phoneme may be represented by more than one letter. In the current research, we investigate two different ways to use orthographic distances for the assessment of language similarities.

### 2.1 Language identification using $n$ -grams

Text-based language identification (LID) is of great practical importance, as there is a widespread need to automatically identify the language in which documents are written. A typical application is Web searching, where knowledge of the language of a document or Web page is valuable information for presentation to a user, or for further processing. The general topic of text-based LID has consequently been studied extensively, and a spectrum of approaches has been proposed with the most important distinguishing factor being the depth of linguistic processing that is utilized.

Here we attempt to identify the languages by using simple statistical measures of the text under consideration. For example, statistics can be gathered from:

- letter sequences (Murthy and Kumar, 2006);
- presence of certain keywords (Giguet, 1995);
- frequencies of short words (Grefenstette, 1995); or
- unique or highly distinctive letters or short character strings (Souter et al., 1994).

Conventional algorithms from pattern recognition are then used to perform text-based LID based on these statistics.

$N$ -gram statistics is a well known choice for building statistical models (Cavnar and Trenkle, 1994, Beesley, 1998, Padro and Padro, 2004, Kruengkrai et al., 2005, Dunning, 1994). An  $n$ -gram is a sequence of  $n$  consecutive letters. The  $n$ -grams of a

string are gathered by extracting adjacent groups of  $n$  letters. The  $n$ -gram combinations in the string “example” are:

<b>bi-grams</b>	:	<b>ex</b>	<b>xa</b>	<b>am</b>	<b>mp</b>	<b>pl</b>	<b>le</b>
<b>tri-grams</b>	:	<b>exa</b>	<b>xam</b>	<b>amp</b>	<b>mpl</b>	<b>ple</b>	
<b>quad-grams</b>	:	<b>exam</b>	<b>xamp</b>	<b>ampl</b>	<b>mple</b>		

In  $n$ -gram based methods for text-based LID, frequency statistics of  $n$ -gram occurrences are used as features in classification. The advantage is that no linguistic knowledge needs to be gathered to construct a classifier. The  $n$ -grams are also extremely simple to compute for any given text, which allows a straightforward trade-off between accuracy and complexity (through the adjustment of  $n$ ) and have been shown to perform well in text-based LID and related tasks in several languages.

We have shown elsewhere (Botha and Barnard, 2007) that several factors influence the accuracy of LID using  $n$ -gram statistics, and those factors are undoubtedly important in the current application as well. For the current research we have not searched for the optimal configuration to assess the relationships between languages; rather, as we report below, a reasonable configuration was selected and employed consistently.

## 2.2 Levenshtein distance

There are several ways in which phoneticians have tried to measure the distance between two linguistic entities, most of which are based on the description of sounds via various representations. This section introduces one of the more popular sequence-based distance measures, the Levenshtein distance measure. In 1995 Kessler introduced the use of the Levenshtein distance as a tool for measuring linguistic distances between dialects (Kessler, 1995). The basic idea behind the Levenshtein distance is to imagine that one is rewriting or transforming one string into another. Kessler successfully applied the Levenshtein algorithm to the comparison of Irish dialects. In this case the strings are transcriptions of word pronunciations. The

rewriting is effected by basic operations, each of which is associated with a cost, as illustrated in Table 2.1 in the transformation of the string “*mošemane*” to the string “*umfana*”, which both are orthographic translations of the word boy in Northern Sotho and Zulu respectively.

Table 2.1: Levenshtein distance between two strings.

	Operation	Cost
mošemane	delete m	1
ošemane	delete š	1
oemane	delete e	1
omane	insert f	1
omfane	substitute o/u	2
umfane	substitute e/a	2
umfana		
Total cost		8

The Levenshtein distance between two strings can be defined as the least costly sum of costs needed to transform one string into another. In Table 2.1, the transformations shown are associated with costs derived from operations performed on the strings. The operations used were: (i) the deletion of a single symbol, (ii) the insertion of a single symbol, and (iii) the substitution of one symbol for another (Kruskal, 1999). The edit distance method was also taken up by (Nerbonne et al., 1996) who applied it to Dutch dialects. Whereas Kruskal (1999) and Nerbonne *et al.* (1996) applied this method to phonetic transcriptions in which the symbols represented sounds, here the symbols are associated with alphabetic letters.

Gooskens and Heeringa (2004) calculated Levenshtein distances between 15 Norwegian dialects and compared them to the distances as perceived by Norwegian listeners. This comparison showed a high correlation between the Levenshtein distances and the perceptual distances.

### 2.2.1 Language grouping

In using the Levenshtein distance measure, the distance between two languages is equal to the average of a sample of Levenshtein distances of corresponding word pairs. When we have  $n$  languages, then the average Levenshtein distance is calculated for each possible pair of languages. For  $n$  languages  $n \times n$  pairs can be formed. The

corresponding distances are arranged in an  $n \times n$  matrix. The distance of each language with respect to itself is found in the distance matrix on the diagonal from the upper left to the lower right. As this is a dissimilarity matrix, these values are always zero and therefore give no real information, so that only  $n \times (n - 1)$  distances are relevant. Furthermore, the Levenshtein distance is symmetric, implying that the distance between word  $X$  and word  $Y$  is equal to the distance between word  $Y$  and word  $X$ . This further implies that the distance between language  $X$  and  $Y$  is equal to the distance between language  $Y$  and  $X$  as well. Therefore, the distance matrix is symmetric. We need to use only one half which contains the distances of  $(n \times (n - 1))/2$  language pairs. Given the distance matrix, groups of larger sizes are investigated. Hierarchical Clustering methods are employed to classify the languages into related language groups using the distance matrix.

Data clustering is a common technique for statistical data analysis, which is used in many fields, including machine learning, bioinformatics, image analysis, data mining and pattern recognition. Clustering is the classification of similar objects into different groups, or more precisely, the partitioning of a data set into subsets, so that the data in each subset share some common trait according to a defined distance measure. The result of this grouping is usually illustrated as a dendrogram, a tree diagram used to illustrate the arrangement of the groups produced by a clustering algorithm (Heeringa and Gooskens, 2003).

### **3. Evaluation**

This evaluation aims to present language groups of the 11 official languages of South Africa generated from similarity and dissimilarity matrices of the languages. These matrices are the results of  $n$ -gram language identification and Levenshtein distance measurements respectively. The diagrams provide visual representations of the pattern of similarities and dissimilarities between the languages.

#### **3.1 Language grouping with text-based LID**

##### **3.1.1 LID text data**

Texts from various domains in all 11 South African languages were obtained from Professor D.J. Prinsloo of the University of Pretoria and using a web crawler (Botha and Barnard, 2005). The data included text from various sources (such as newspapers,

periodicals, books, the Bible and government documents) and therefore, the corpus spans several domains.

### **3.1.2 Classification features**

For either a fixed-length sample or an unbounded amount of text, the frequency counts of all  $n$ -grams were calculated. The characters that can be included in  $n$ -gram combinations were a space, the 26 letters of the Roman alphabet, the other 14 special characters found in Afrikaans, Northern Sotho and Tswana, and the unique combination 'n', which functions as a single character in Afrikaans. No distinction was made between upper and lower case characters.

### **3.1.3 Support vector machine**

The support vector machine (SVM) is a non-linear discriminant function that is able to generalize well, even in high-dimensional spaces. The classifier maps input vectors to a higher dimensional space where a separating hyper-plane is constructed. The hyper-plane maximizes the margin between the two datasets (Burges, 1998). In real-world problems data can be noisy and the classifier would usually over-fit the data. For such data, constraints on the classifiers are relaxed by introducing slack variables. This improves overall generalization (Cristianini and Shawe-Taylor, 2005).

The LIBSVM (Chang and Lin, 2001) library provides a full implementation of several SVMs. The size of the feature space grows exponentially with  $n$ , which leads to long training times and extensive resource usage as  $n$  becomes large; we therefore limited our classification features to only 3-gram combinations. Thus the feature dimension of the SVM is equal to the number of 3-gram combinations. Two language models were built. The one model was built with samples of 15 characters from a training set of 200 000 characters per language. The other model was built with samples of 300 characters using the same training set. For the 15 character language models a sample contained the frequency count of each 3-gram combination in the sample string of 15 characters. The 300 character model a sample similarly contains the frequency count of each 3-gram combination in the sample string of 300 characters. Samples of the testing set are created using the same character window (namely 15 characters or 300 characters) as used to build the language model. After training the SVM language model the test samples can be classified according to language.

The SVM used an RBF kernel, and overlap penalties (Botha and Barnard, 2005) were employed to allow for non-separable data in the projected high-dimensional feature space. Sensible values for the two free parameters (kernel width ( $h=1$ ) and margin-overlap trade-off ( $C=180$ , a large penalty for outliers)) were found on a small set of data. These “reasonable” parameters were employed throughout our experiments. Classification is done in a “one-against-one” approach in which  $k(k-1)/2$  classifiers are constructed (in our case 55 classifiers are created) and each one trains from data of two different classes. Classification is done by a voting strategy. Each binary classification is considered to be a vote for the winning class. All the votes are tallied, and the test sample is assigned to the class with the largest number of votes.

### 3.1.4 Confusion matrix

In the confusion matrix below (Table 3.1), each row represents the correct language of a set of samples. The columns indicate the languages selected by the classifier. Thus, more samples on the diagonal axis of the matrix indicate better overall accuracy of the classifier, consequently generating a similarity matrix. It is clear that the higher values in the matrix reflect high levels of similarity between the paired languages.

Table 3.1: Confusion matrices for SVM classifier. (a) 300 character text fragments classified using 3-gram feature statistics. (b) 15 character text fragments classified using 3-gram feature statistics.

	S. Sot	N. Sot	Tsw	Xho	Zul	Nde	Swa	Ven	Tso	Afri	Eng
S. Sot	646	0	1	0	2	0	0	0	0	0	1
N. Sot	0	648	2	0	3	0	0	0	0	0	0
Tsw	2	6	643	0	0	0	0	0	0	0	0
Xho	0	0	0	610	25	16	0	0	0	0	1
Zul	0	2	0	43	589	15	0	0	0	0	1
Nde	0	0	0	23	50	585	0	0	0	0	1
Swa	0	0	0	0	1	0	650	0	0	0	3
Ven	0	0	0	0	0	0	0	657	0	0	0
Tso	0	0	0	0	1	0	0	0	655	0	0
Afr	0	0	0	0	0	0	0	0	0	660	0
Eng	0	0	0	0	0	0	0	0	0	0	650

(a)

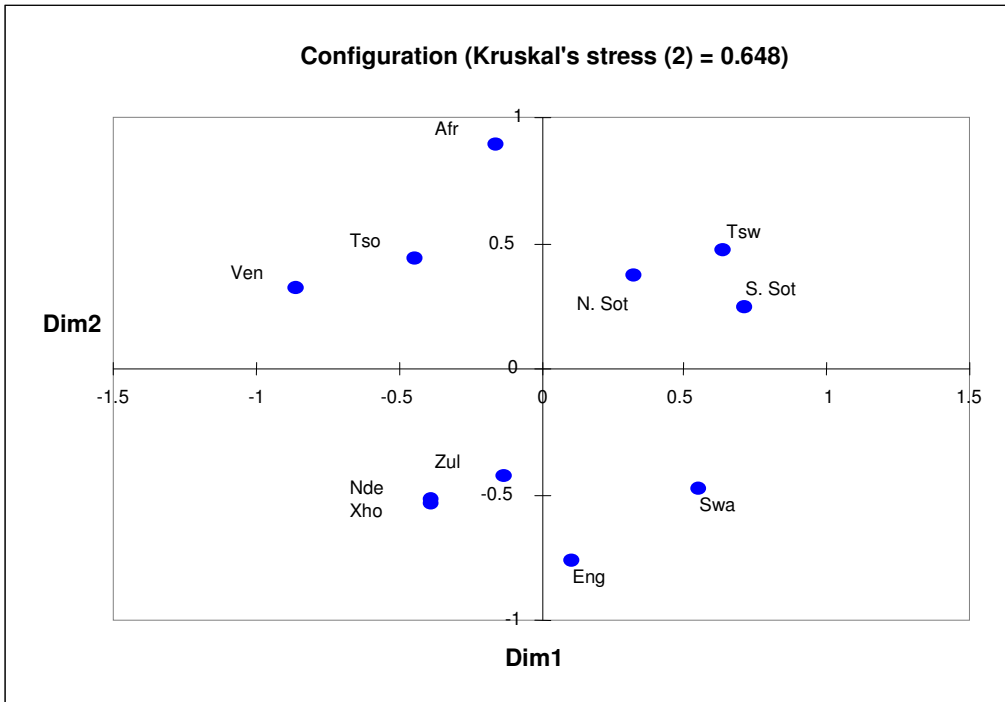
	S. Sot	N. Sot	Tsw	Xho	Zul	Nde	Swa	Ven	Tso	Afr	Eng
S. Sot	9743	1370	1589	36	50	41	32	75	75	28	68
N. Sot	1698	9237	1906	34	50	41	14	49	75	15	53
Tsw	1991	1994	8843	25	23	45	32	36	58	29	36
Xho	72	32	15	8123	2411	1821	434	44	69	41	50
Zul	52	42	16	2769	7177	2192	663	54	69	30	83
Nde	82	59	42	2343	2692	7157	594	98	115	12	47
Swa	70	26	33	600	851	647	10622	41	122	24	137
Ven	142	80	67	139	90	158	53	12158	270	15	46
Tso	138	124	77	124	87	106	161	250	12028	22	78
Afr	27	14	17	30	25	12	25	9	11	12876	232
Eng	44	38	9	34	53	27	51	21	45	177	12608

(b)

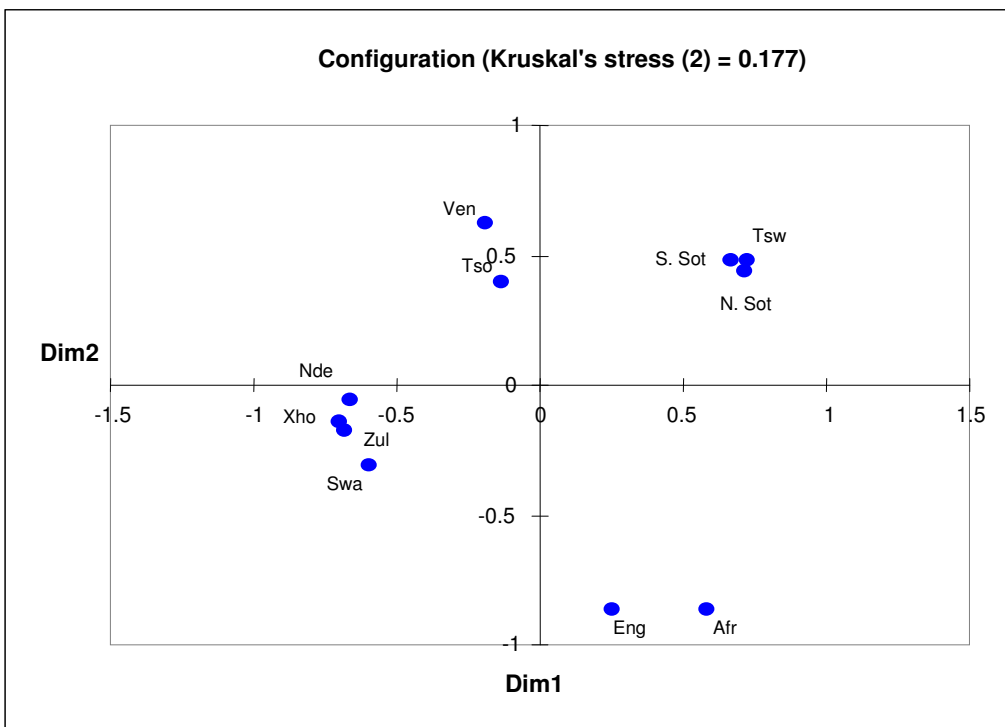
### 3.1.5 A graphical representation of language distances

The confusion matrices provide a clear indication of the ways the languages group into families. These relationships can be represented visually using graphical techniques. Multidimensional scaling (MDS) is a technique used in data visualization for exploring the properties of data in high-dimensional spaces. The algorithm uses a matrix of similarities between items and then assigns each item a location in a low dimensional space to match those distances as closely as possible. We used the confusion matrix to serve as similarity measure between languages, using the statistical package XLSTAT (XLSTAT, 2007). The confusion matrix was processed into a matrix of distances using the Pearson correlation coefficients between the rows, and input into the multidimensional scaling algorithm which mapped the language similarities in a 2-dimensional space.

Figure 3.1 shows the mapping that was created using the confusion matrix in Table 3.1; we can see that the languages from the same subfamilies group together. The mapping using the 15 character text fragment shows a more definite grouping of the families than the mapping that uses the 300 character text fragment. In the 15 character mapping the Nguni and Sotho languages are more closely related internally than the pair of Germanic languages and within the Nguni languages Swati is somewhat distant from the other three languages. As expected, Venda and Tsonga are consistently separated from the other nine languages.



(a)

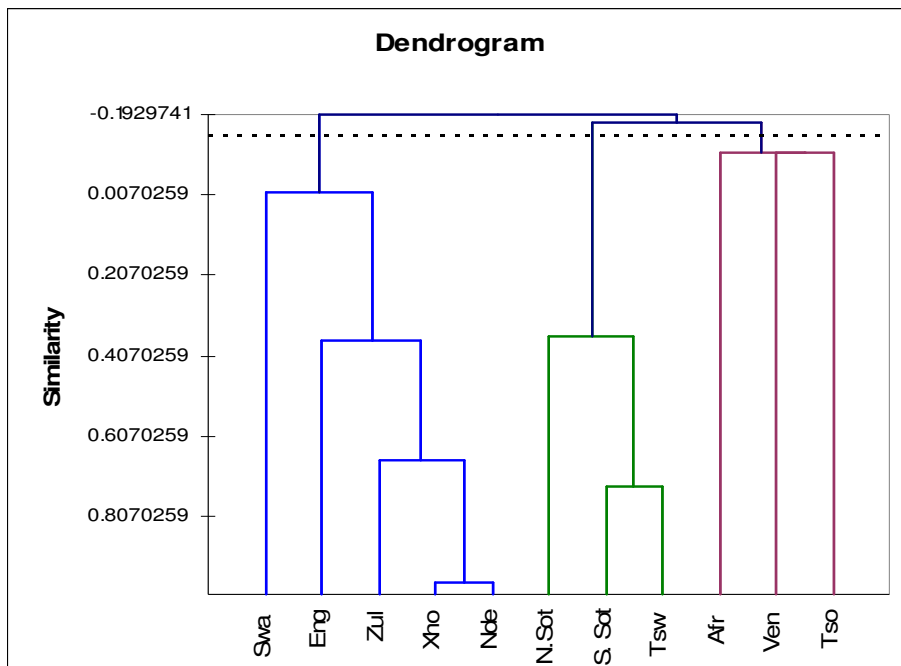


(b)

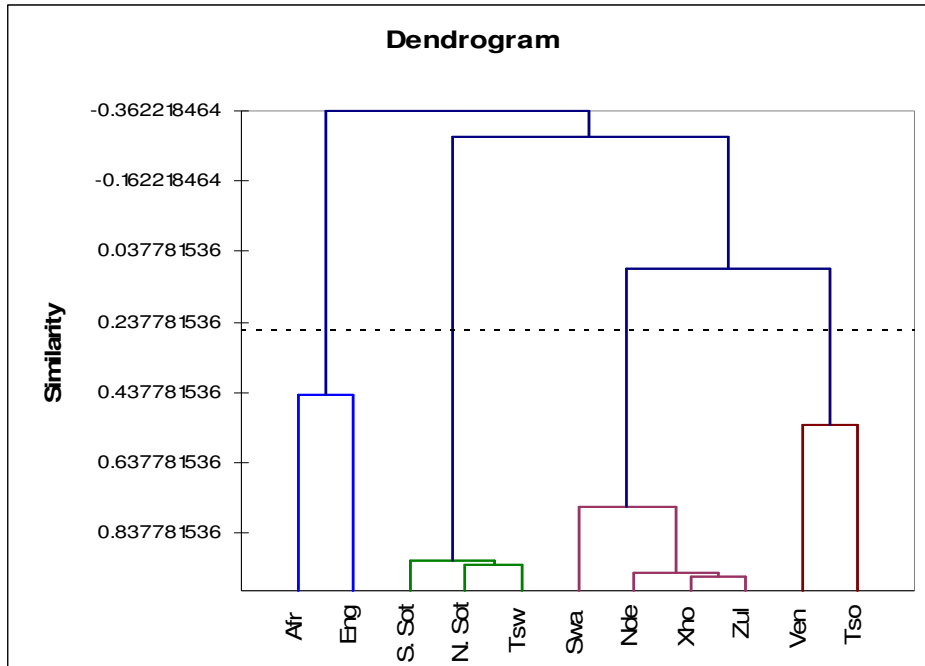
Figure 3.1: Multi-dimensional scale to represent similarities between languages calculated from the confusion matrices in Table 3.1. (a) 300 character text fragments, and (b) 15 character text fragments.

In conjunction with multidimensional scaling, dendrograms also provide a visual representation of the pattern of similarities or dissimilarities among a set of objects. We again used the confusion matrix, processed into a matrix of distances using the Pearson correlation coefficients to serve as similarity measure between languages, using the statistical package XLSTAT (XLSTAT, 2007).

Figure 3.2 illustrates the dendrograms derived from clustering the similarities between the languages as depicted by the confusion matrices in Table 3.1. The dendrogram using the 15 character text fragment shows four classes representing the previously defined language groupings, Nguni, Sotho, Venda and Tsonga and English and Afrikaans. This dendrogram closely relates to the language groupings described in Heine and Nurse, (2000).



(a)



(b)

Figure 3.2: Dendrogram calculated from the confusion matrices of Table 3.1. (a) 300 character text fragments, and (b) 15 character text fragments.

### 3.2 Language grouping using Levenshtein distance

#### 3.2.1 Levenshtein distance data

Levenshtein distances were calculated using existing parallel orthographic word transcriptions of sets of 50 and 144 words from each of the 11 official languages of South Africa. The data was manually collected from various multilingual dictionaries and online resources. Initially, 200 common English words, mostly common nouns easily translated into the other 10 languages, were chosen. From this set, those words having unique translations into each of the other 10 languages were selected, resulting in 144 words (and also a subset of 50 from the 144 words) that were used in the evaluations.

#### 3.2.3 Distance matrix

Table 3.2 represents distance matrices, containing the distances, taken pair-wise, between the different languages as calculated from the summed Levenshtein distance between the 50 and 144 target words. In contrast to the confusion matrices, lower numbers in the matrices reflect less dissimilarity between the selected pair of

languages. The distance matrices again contain  $n \times (n - 1)/2$  independent elements in light of the symmetry of the distance measure.

Table 3.2: Distance matrices calculated from Levenshtein distance between (a) 50 words, and (b) 144 words.

	Afr	Eng	Nde	Xho	Zul	N. Sot	S. Sot	Tsw	Swa	Ven	Tso
Afr	0	157	438	443	451	279	452	390	462	352	390
Eng	157	0	437	437	444	276	438	382	450	355	389
Nde	438	437	0	279	232	389	440	427	257	403	390
Xho	443	437	279	0	276	375	403	418	306	396	395
Zul	451	444	232	276	0	384	430	426	194	395	399
N. Sot	279	276	389	375	384	0	271	186	384	317	363
S. Sot	452	438	440	403	430	271	0	292	410	446	448
Tsw	390	382	427	418	426	186	292	0	416	364	382
Swa	462	450	257	306	194	384	410	416	0	395	410
Ven	352	355	403	396	395	317	446	364	395	0	350
Tso	390	389	390	395	399	363	448	382	410	350	0

(a)

	Afr	Eng	Nde	Xho	Zul	N. Sot	S. Sot	Tsw	Swa	Ven	Tso
Afr	0	443	1025	984	1014	829	931	887	1049	874	898
Eng	443	0	1018	981	1002	820	920	881	1044	865	896
Nde	1025	1018	0	519	328	900	954	956	472	889	798
Xho	984	981	519	0	502	867	887	922	597	873	819
Zul	1014	1002	328	502	0	881	925	945	348	870	759
N. Sot	829	820	900	867	881	0	349	315	883	727	762
S. Sot	931	920	954	887	925	349	0	480	912	851	855
Tsw	887	881	956	922	945	315	480	0	943	808	825
Swa	1049	1044	472	597	348	883	912	943	0	892	785
Ven	874	865	889	873	870	727	851	808	892	0	722
Tso	898	896	798	819	759	762	855	825	785	722	0

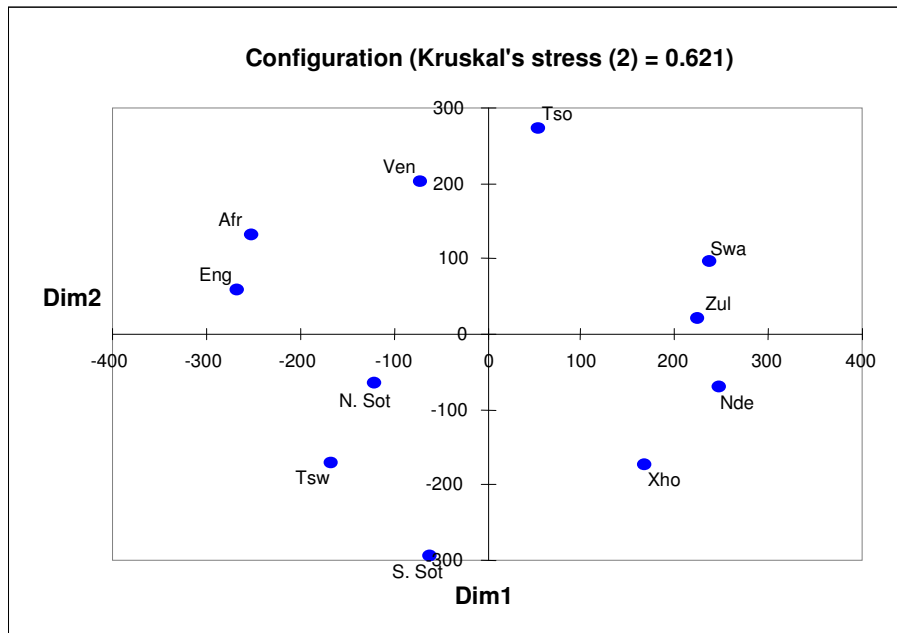
(b)

### 3.2.4 Visual representation

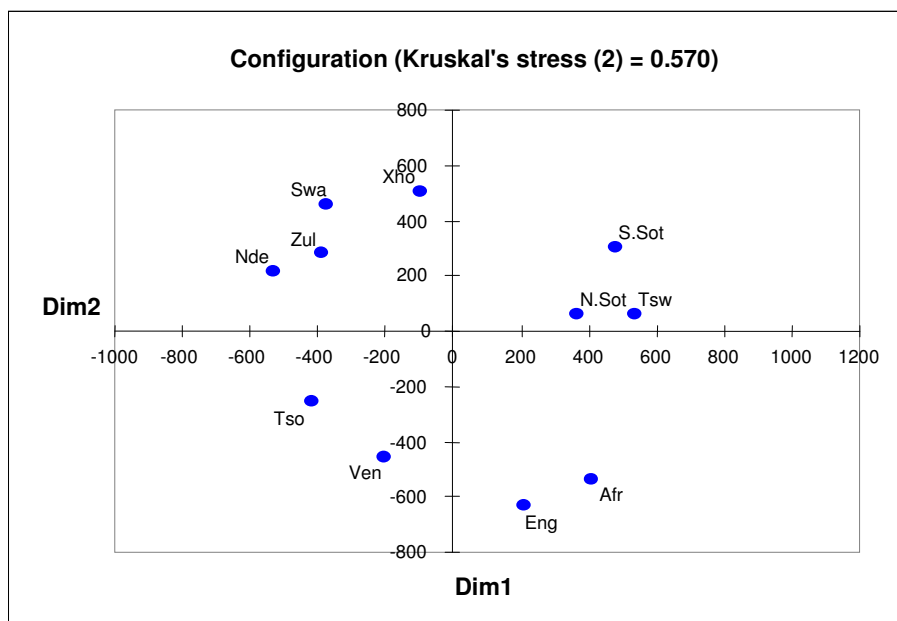
As above, the relationships between the languages for the matrices derived from the Levenshtein distance are represented visually in Figures 3.3 and 3.4 using graphical techniques. Again, multidimensional scaling is used. However, in this case the algorithm uses distance matrices of dissimilarities as opposed to the confusion matrices of similarities. The language dissimilarities are mapped onto a 2-dimensional space (Figure 3.3).

Figure 3.3 shows the mappings generated using the distance matrices in Table 3.2. Here also, though in different quadrants, the languages from the same subfamilies group together. The relative closeness within the Nguni and Sotho sub-families is not as clearly indicated in Figure 3.3 (a) as in Figure 3.3 (b) or Figure 3.1 (b), and the

individual languages appear more spaced out in the quadrants. As before, Venda and Tsonga are consistently separated from the other nine languages.



(a)

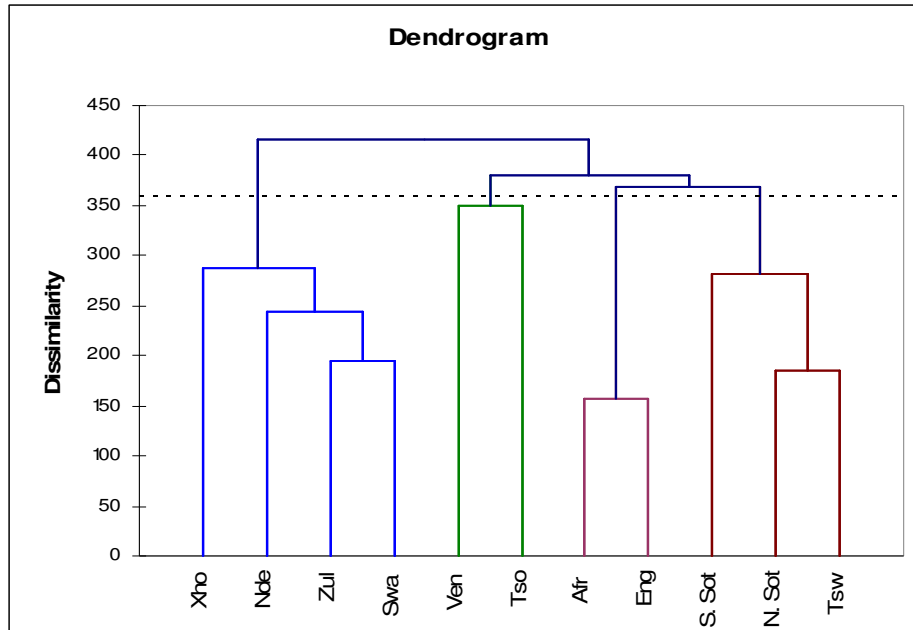


(b)

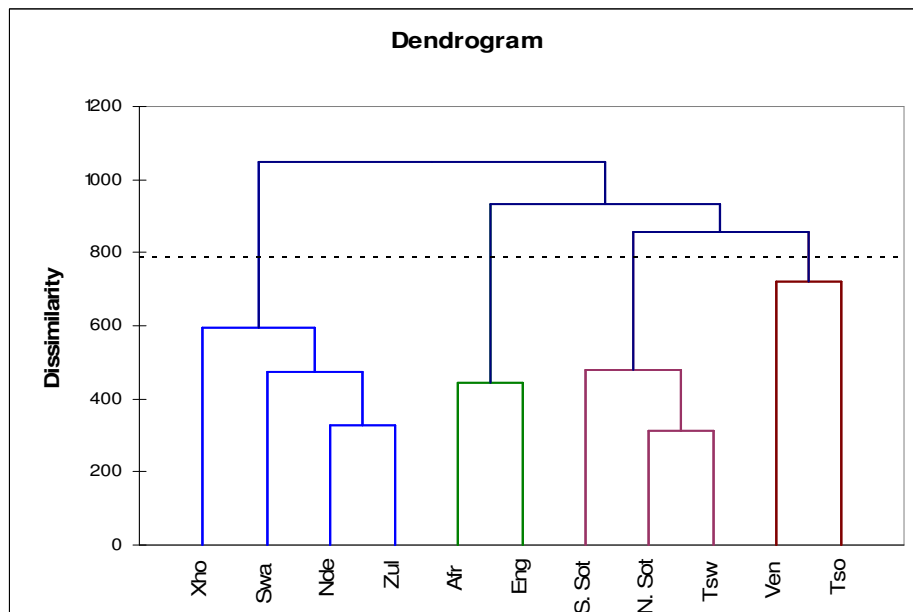
Figure 3.3: Multi-dimensional scale to represent dissimilarities between languages calculated from the distance matrix in Table 3.2. (a) 50 words, and (b) 144 words.

Figure 3.4 shows dendrograms generated from the dissimilarities matrices of Table 3.2. As in Figure 3.2 (b), here too the dendrograms show four classes representing the

previously defined language groupings. In the Nguni class of Figure 3.4 (b), the relative spacing of the languages differs from that of Figure 3.2 (b). For example, in Figure 3.4 (b), Zulu appears closer to Ndebele whereas in Figure 3.2, Zulu is closer to Xhosa. We note also that Figure 3.4 (a) depicts a more refined grouping of the languages than Figure 3.2 (a).



(a)



(b)

Figure 3.4: Dendrogram calculated from the distance matrix of Table 3.2. (a) 50 words, and (b) 144 words.

#### 4. Conclusions

We have seen that both confusion matrices between languages resulting from text-based language identification and Levenshtein distance matrices can be effectively combined with MDS and dendrograms to represent language relationships. Both methods reflect the known family relationships between the languages being studied. The main conclusion of this research is therefore that statistical methods, based on only orthographic transcriptions, are able to provide useful *objective* measures of language similarities. It is clear that these methods can be refined further using other inputs such as phonetic transcriptions or acoustic measurements; such refinements are likely to be important when, for example, fine distinctions between dialects are required.

Each approach has its advantages and disadvantages. Levenshtein distance measures do not require much data to perform a reasonable classification of the data. With as few as 50 words per language, reasonable classification is possible. Also, the process of generating the distance matrix is not computationally taxing. However, this method is seen to be less discriminating in assessing language similarities – from the historical record (Heine and Nurse, 2000), it is clear, for example, that the tighter internal grouping of the Sotho and Nguni languages (as found with the LID-based approach) is more accurate. Similarly, the slightly larger separation of Swati from the other Nguni languages agrees with the anecdotal evidence on mutual intelligibility.

In a text-based LID system, high classification accuracy is a central goal. The size of the text fragment to be identified plays an important role in the accuracy achieved, since a larger text fragment can generally be identified more accurately. Hence, LID systems tend to use the longest text fragments available. However, for measuring language similarities, shorter text fragments may actually be preferable: In our experiments we found that the lower classification accuracy achieved on a smaller text fragment enables us to cluster the languages in a more discriminative fashion.

It would be most interesting to see whether closer agreement between these methods can be achieved by measuring Levenshtein distances between larger text collections – perhaps even parallel corpora rather than translations of word lists. Comparing these distance measures with measures derived from acoustic data is another pressing

concern. Finally, it would be very valuable to compare various distance measures against other criteria for language similarity (e.g. historical separation or mutual intelligibility) in a rigorous fashion.

## List of references

- BEESLEY, K. R. (1998) Language identifier: A computer program for automatic natural language identification of online text. *Language at Crossroads: Proceedings of the 29th Annual Conference of the American Translators Association*, pp. 47-54.
- BOTHA, G. & BARNARD, E. (2005) Two approaches to gathering text corpora from the World Wide Web. *Proceedings of the 16th Annual Symposium of the Pattern Recognition Association of South Africa*, pp. 194.
- BOTHA, G. & BARNARD, E. (2007) Factors that affect the accuracy of text-based language identification. *The 18th Annual Symposium of the Pattern Recognition Association of South Africa 2007*, pp. 7-12.
- BURGES, C. J. C. (1998) A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, vol. 2, pp. 121-167.
- CAVNAR, W. B. & TRENKLE, J. M. (1994) N-gram-based text categorization. *Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval*, pp. 161-169.
- CHANG, C. & LIN, C. (2001) LIBSVM: a library for support vector machines. <http://www.csie.ntu.edu.tw/~cjlin/libsvm>. Date of access: 30 Jul. 2007.
- CRISTIANINI, N. & SHAWE-TAYLOR, J. (2005) An introduction to support vector machines and other kernel-based learning methods. Cambridge: Cambridge University Press.
- DUNNING, T. (1994) Statistical identification of language. *Computing Research Lab, New Mexico State University, Technical Report CRL MCCS-94-273*.
- GIGUET, E. (1995) Categorization according to language: A step toward combining linguistic knowledge and statistical learning. *Proceedings of the 4th International Workshop on Parsing Technologies*.
- GOOSKENS, C. & HEERINGA, W. (2004) Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data. *Language Variation and Change*, vol. 16, pp. 189-207.
- GREFENSTETTE, G. (1995) Comparing two language identification schemes. *Third International Conference on Statistical Analysis of Textual Data*. Rome.
- HEERINGA, W. & GOOSKENS, C. (2003) Norwegian dialects examined perceptually and acoustically. *Computers and the Humanities*, vol. 37, pp. 293-315.
- HEINE, B. & NURSE, D. (2000) African languages: An introduction. Cambridge University Press.
- KESSLER, B. (1995) Computational dialectology in Irish Gaelic. *The 7th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 60-67.

- KRUENKRAI, C., SRICHAIVATTANA, P., SORLERTLAMVANICH, V. & ISAHARA, H. (2005) Language identification based on string kernels. *IEEE International Symposium on Communications and Information Technology*, vol. 2, pp. 926-929.
- KRUSKAL, J. B. (1999) An overview of sequence comparison. Stanford.
- MURTHY, K. N. & KUMAR, G. B. (2006) Language identification from small text samples. *The Journal of Quantitative Linguistics*, vol. 13, pp. 57-80.
- NERBONNE, J., HEERINGA, W., HOUT, E. V. D., KOOI, P. V. D., OTTEN, S. & VIS, W. V. D. (1996) Phonetic distance between Dutch dialects. *Sixth CLIN Meeting*, pp. 185-202.
- PADRO, M. & PADRO, L. (2004) Comparing methods for language identification. *Proceedings of the XX Congreso de la Sociedad Espanola para el Procesamiento del Lenguaje Natural*, pp. 155-162.
- SOUTER, C., CHURCHER, G., HAYES, J., HUGHES, J. & JOHNSON, S. (1994) Natural language identification using corpus-based models. *Hermes Journal of Linguistics*, vol. 13, pp. 183-203.
- VAN-BEZOOIJEN, R. & HEERINGA, W. (2006) Intuitions on linguistic distance: geographically or linguistically based? In: Tom Koole, Jacomine Northier and Bert Tahitu (eds). *Artikelen van de Vijfde sociolinguistische conferentie*, pp. 77-87.
- VAN-HOUT, R. & MÜNSTERMANN, H. (1981) Linguistic distance, dialect and attitude. *Gramma 5*, pp. 101-123.
- XLSTAT (2007) XLSTAT. <http://www.xlstat.com/en/download/>. Date of access: 20 Aug. 2007.

**Key concepts:**

Language distances

*n*-gram

language identification

Levenshtein distance

clustering

**Kernbegrippe:**

Taalafstande

*n*-gram

taalherkenning

Levenshtein-afstand

groepering