Classification of South African languages using text and acoustic based methods: A case of six selected languages

Peleira Nicholas Zulu
University of KwaZulu-Natal
Electrical, Electronic and
Computer Engineering
Durban, 4041, South Africa
zulup1@ukzn.ac.za

Abstract

Language variations are generally known to have a severe impact on the performance of Human Language Technology Systems. In order to predict or improve system performance, a thorough investigation into these variations, similarities and dissimilarities, is required. Distance measures have been used in several applications of speech processing to analyze different varying speech attributes. However, not much work has been done on language distance measures, and even less work has been done involving South African languages. This study explores two methods for measuring the linguistic distance of six South African languages. It concerns a text based method, (the Levenshtein Distance), and an acoustic approach using extracted mean pitch values. The Levenshtein distance uses parallel word transcriptions from all six languages with as little as 144 words, whereas the pitch method is text-independent and compares mean language pitch differences. Cluster analysis resulting from the distance matrices from both methods correlates closely with human perceptual distances and existing literature about the six languages.

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 & Heeringa, 2004; Van-Bezooijen & Heeringa, 2006; Van-Hout & Münstermann, 1981), 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 well known 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, what are the objective factors that allow one to assess the levels of distance?

This bears 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.

Proceedings of NAACL-HLT 2013, pages 280–287,
Atlanta, Georgia, 9–14 June 2013. ©2013 Association for Computational Linguistics

In the current article, we focus on language distance measures derived from both text and acoustic formats; we apply two different techniques, namely Levenshtein distance between orthographic word transcriptions, and distances between language pitch means 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 multidimensional scaling and dendrograms—two standard statistical techniques. This allows us to visualize and assess the methods relative to known linguistic facts in order to judge their relative reliability(Zulu, Botha, & Barnard, 2008).

Our evaluation is based on six of the eleven official languages of South Africa.[1] The eleven official 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). The six languages chosen for our evaluation are English, Afrikaans, Zulu, Xhosa, Northern Sotho (also known as Sepedi) and Tswana, which equally represent the three groups; Germanic, Nguni and Sotho.

We believe that an understanding of these language distances is not only 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 all languages in a particular multi-lingual community, but has insufficient funding to do so. Such an organization can be guided by knowledge of language distances and mutual intelligibility between languages to make an appropriate choice of publication languages.

The following sections describe the Levenshtein distance and pitch characteristics in detail. Thereafter, the paper will present an evaluation on the six languages of South Africa, highlighting language groupings and proximity patterns. In conclusion, the paper discusses the results.

2 Theoretical Background

Orthographic transcriptions are one of the most basic types of annotation used for speech transcription, and are particularly 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 its writing system. English, for example, has an alphabet of 26 letters which includes 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 the use of Levenshtein distance on orthographic transcriptions for the assessment of language similarities.

On the other hand, speech has been and still very much is the most natural form of communication. Prosodic characteristics such as rhythm, stress and intonation in speech convey important information regarding the identity of a spoken language. Results of perception studies on spoken language identification confirm that prosodic information, specifically pitch and intensity—which represent intonation and stress respectively—are useful for language identification (Kometsu, Mori, Arai, & Murahara, 2001; Mori et al., 1999). This paper presents a preliminary investigation of pitch and its role in determining acoustic based language distances.

2.1 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 the Levenshtein Distance Measure, one of the more popular sequence-based distance measures. 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 Distance

[1] Data for all eleven languages is available on the Lwazi website: (http://www.meraka.org.za/lwazi/index.php).

measure to the comparison of Irish dialects. In his work, the strings were transcriptions of word pronunciations. In general, rewriting is effected by basic operations, each of which is associated with a cost, as illustrated in Table 1 in the transformation of the string "*mošemane*" to the string "*umfana*", which are both orthographic translations of the word boy in Northern Sotho and Zulu respectively.

	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

Table 1. Levenshtein Distance between two strings.

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 1, the transformations shown are associated with costs derived from operations performed on the strings. The operations used are: (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 were associated with alphabetic letters.

Similarly, 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 Language pitch distance

Speech is primarily intended to convey some message through a sequence of legal sound units in a language. However, speech cannot merely be characterized as a sequence of sound units. There are some characteristics that lend naturalness to speech, such as the variation of pitch, which provides some recognizable melodic properties to spoken language. This controlled modulation of pitch is referred to as *intonation*. The sound units are shortened or lengthened in accordance to some underlying pattern giving rhythmic properties to speech. The information attained from these rhythmic patterns increases the intelligibility of spoken languages, enabling the listener to segment continuous speech into phrases and words with ease (Shriberg, Stolcke, Hakkani-Tur, & Tur, 2000). The characteristics that make us perceive this and other information such as stress, accent and emotion are collectively referred to as prosody. Comparisons have shown that languages differ greatly in their prosodic features (Hirst & Cristo, 1998), therefore providing a basis for objective comparison between languages. Further, pitch is a perceptual attribute of sound, the physical correlate of which is fundamental frequency (F_0), which represents vibration of the vocal folds.

This paper extracts pitch contours from six different languages, and uses the mean fundamental frequency values for each language to calculate the differences in pitch amongst them. From this we derive a distance matrix of F_0 dissimilarities (differences) which in turn is used to obtain language groupings.

2.3 Language Clustering

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. With pitch, the distance between two languages is merely the difference between the mean fundamental frequencies of the two languages. When we have n languages, then these distances are calculated for each possible pair of languages. For n languages n x n distances can be calculated. The corresponding distances are arranged in an n x 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 x (n - 1) distances are relevant. Furthermore, both the Levenshtein and pitch distances are symmetric, implying that the distance between language X and Y is equal to the distance between language Y and X. Therefore, the distance matrix is symmetric. We need to use only one half which contains (n x (n - 1))/2 distances. 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 & Gooskens, 2003), whereas multidimensional scaling adds to illustrate the visualization of the language proximities in a 2-dimensional space.

3 Evaluation

This evaluation aims to present language groups of the six chosen languages of South Africa generated from dissimilarity matrices of the languages. These matrices are the results of Levenshtein distance and average pitch distance measurements. The diagrams provide visual representations of the pattern of similarities and dissimilarities between the languages.

3.1 Language grouping using Levenshtein distance

Levenshtein distances were calculated using existing parallel orthographic word transcriptions of 144 words from each of the six languages. 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 five languages, were chosen. From this set, those words having unique translations into each of the other five languages were selected, resulting in 144 words that were used in the evaluations. Examples of four word translations in all six languages are shown in Table 2.

Eng	Afr	Xho	Zul	N.Sot	Tsw
fish	vis	intlanzi	inhlanzi	hlapi	tlhapi
house	huis	indlu	indlu	ntlo	ntlo
mother	ma	uma	umama	mma	mme
school	skool	isikolo	isikole	sekolo	sekole

Table 2. Example translations of four common words.

Distance matrix

Table 3 represents the distance matrix, containing the distances, taken pair-wise, between the different languages as calculated from the summed Levenshtein Distances between the 144 target words. The zero values along the diagonal axis of the matrix indicate no dissimilarity, making it clear that higher values reveal high levels of dissimilarity between the paired languages. The distance matrix contains *n* x *(n – 1)/2* independent elements in light of the symmetry of the distance measure.

	Afr	Eng	Xho	Zul	N. Sot	Tsw
Afr	0	443	984	1014	829	887
Eng	443	0	981	1002	820	881
Xho	984	981	0	502	867	922
Zul	1014	1002	502	0	881	945
N. Sot	829	820	867	881	0	315
Tsw	887	881	922	945	315	0

Table 3. Distance matrices calculated from Levenshtein Distance between 144 words.

Graphical representation

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 dissimilarities between items and then assigns each item a location in a low dimensional space to match those distances as closely as possible. The study used the dissimilarity matrix to serve as a measure between languages, and used the statistical package XLSTAT (XLSTAT, 2012). The dissimilarity matrix was input into the multidimensional scaling algorithm which mapped the language dissimilarities in a 2-dimensional space.

Figure 1 shows the mapping that was created using the dissimilarity matrix in Table 3; we can see that the languages from the same subfamilies group together. The mapping using just 144 words shows a definite grouping of the families. In the mapping the Sotho languages are more closely related internally than both the Nguni and Germanic languages as expected — from the historical record (Heine & Nurse, 2000), it is clear that a tighter internal grouping of the Sotho and Nguni languages is accurate.

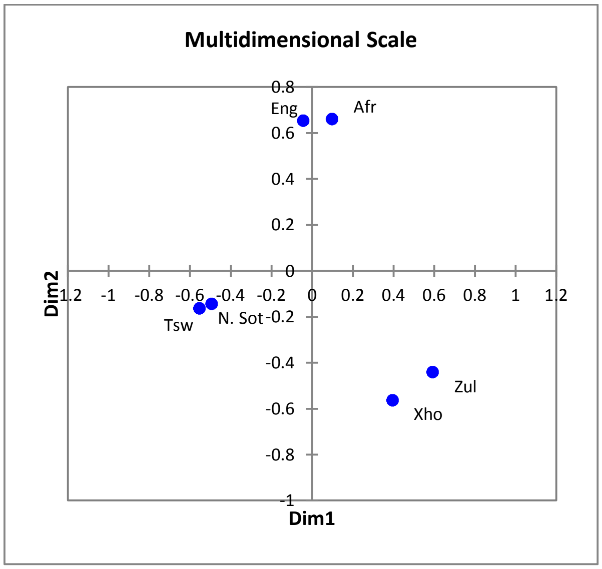

Figure 1. Multidimensional scale to represent dissimilarities between languages calculated from the dissimilarity matrix in Table 3.

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 dissimilarity matrix in Table 3 with the statistical package XLSTAT.

Figure 2 illustrates the dendrogram derived from clustering the dissimilarities between the languages as depicted by the dissimilarity matrix in Table 3. The dendrogram shows three classes representing the previously defined language groupings, Nguni, Sotho and Germanic. This dendrogram closely relates to the language groupings described in (Heine & Nurse, 2000).

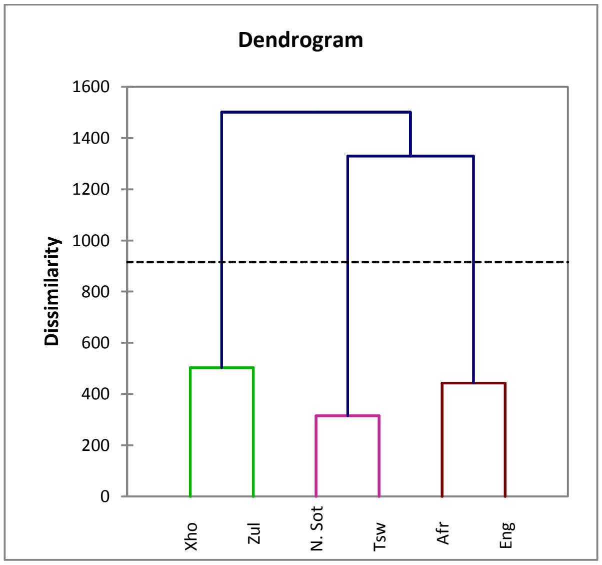

Figure 2. Dendrogram calculated from the dissimilarity matrix of Table 3.

3.2 Pitch Extraction and language grouping

The extraction of pitch contours was carried out with *Praat* (Boersma & Weenink, 2011), a free scientific software program for the analysis of speech and phonetics. The use of Praat is advantageous in that it is fairly easy to use, has high processing speed, is accurate and allows scripting, which is very useful in processing large numbers of files (in our case, speech recordings).

A Praat script was written specifying two main parameters; the expected minimum and maximum pitch values in Hertz, which were selected to be 75Hz and 600Hz respectively. The extraction of pitch contours is based on the detection of periodicities. The Praat command *To PointProcess (periodic, peaks)...* analyses the selected speech file and creates a sequence of points in time. The acoustic periodicity detection is performed on the basis of an accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio working in the autocorrelation domain as described by Boersma (Boersma, 1993). This method was able to achieve more accurate and noise-resistant results when compared to combs or cepstrum based methods (Pokorny, 2011). The extracted acoustic periodicity contour is interpreted as being the frequency of an underlying sequence of glottal closures in vocal fold vibrations. For each speech file—for every voiced interval—a number of points representing glottal pulses are found and their points in time are saved, forming the pitch contour for that particular speech file (Pokorny, 2011). Pitch contours were extracted from 5000 speech files per language for each of the six languages, with each language having approximately 200 different speakers (25 recordings per speaker) with a relatively equal distribution of males and females, all aged between 18 and 65 years.

The extracted pitch frequency points for all 5000 files were collected and placed in a single array for each language. Each array represents the pitch distribution for the specific language, and the mean frequency for each language was used to model the respective language. The dissimilarity matrix was then derived from the differences of these means for each pair of languages. Figure 3 illustrates the distribution of pitch frequencies for the selected languages. It clearly shows the relative pitch content variations of the different languages, which is key to determining the dissimilarity amongst the languages. Also of note in Figure 3 are the peak positions representing approximate positions of male and female fundamental frequencies—in the range of 85 to 180Hz for males and 165 to 255 Hz for females.

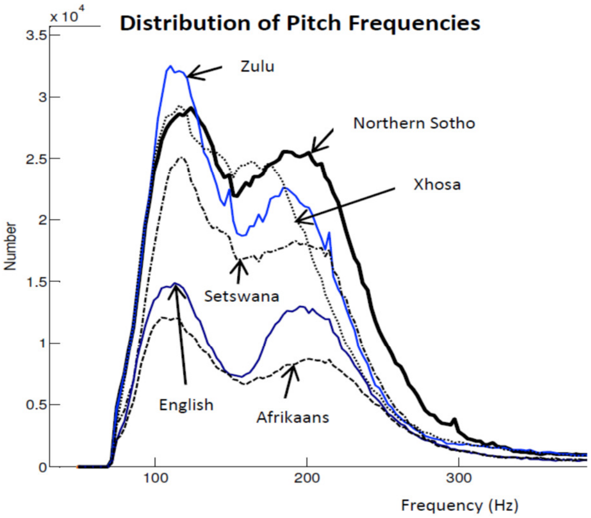

Figure 3. Distribution of pitch frequencies extracted from 6 South African languages.

Distance matrix

Table 4 represents the distance matrix—containing the distances taken pair-wise—between the different languages as calculated from the mean pitch frequencies of the six languages. Again, higher numbers in the matrix reflect high dissimilarity between the selected pair of languages.

	Afr	Eng	Xho	Zul	N. Sot	Tsw
Afr	0	5.1	16.09	17.11	9.66	12.61
Eng	5.1	0	10.99	12.01	4.56	7.51
Xho	16.09	10.99	0	1.02	6.43	3.48
Zul	17.11	12.01	1.02	0	7.45	4.5
N. Sot	9.66	4.56	6.43	7.45	0	2.95
Tsw	12.61	7.51	3.48	4.5	2.95	0

Table 4. Distance matrix calculated from mean pitch frequencies of six South African languages.

Graphical representation

As with the Levenshtein Distance, the relationships between the languages are represented visually in Figures 4 and 5 using graphical techniques and multidimensional scaling. The language dissimilarities are mapped on to a 2-dimensional space shown in Figure 4. Here also, the languages from the same sub-families are grouped together. The relative closeness within the three sub-families is not as clearly indicated in Figure 4 as in Figure 1, but the distinction is clearly visible.

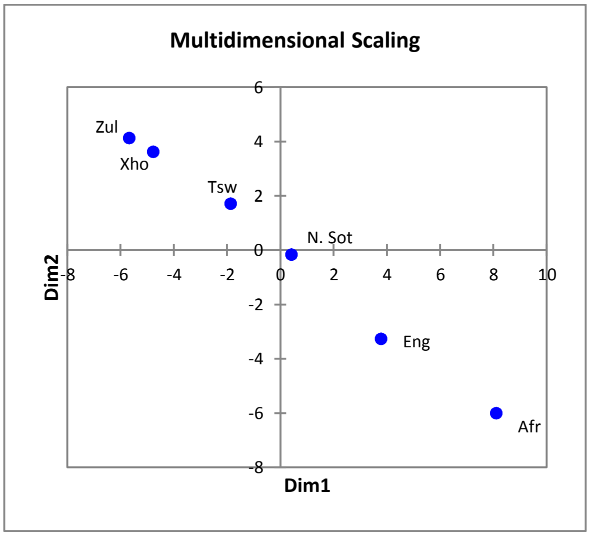

Figure 4. Multi-dimensional scale calculated from the pitch-based matrix of Table 4.

Figure 5 shows the dendrogram generated from the dissimilarities matrix of Table 4. As in Figure 2, the dendrogram shows three classes representing the previously defined language sub-families. Figure 5 differs from Figure 2 in the branching of the three sub-families, where Figure 2 shows the Germanic languages branching from the same parent as the Sotho sub-family. Figure 5 offers a more accurate account by separating the Germanic subgroup from the Bantu languages. Thus, Figure 5 depicts a more refined grouping of the languages than Figure 2.

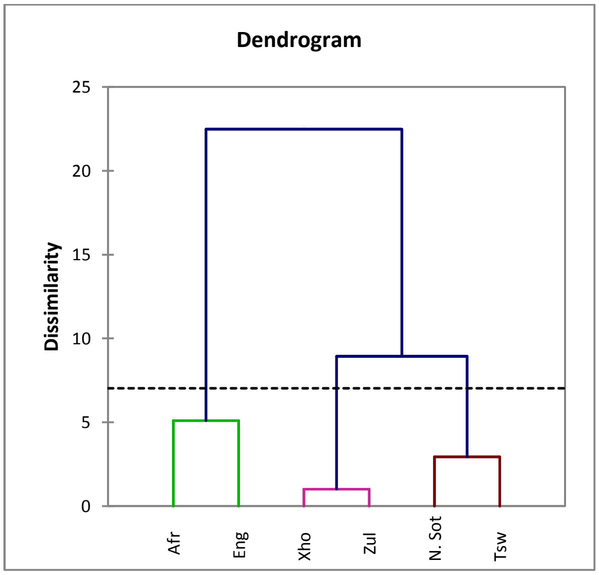

Figure 5. Dendrogram calculated from the pitch-based distance matrix of Table 4.

Conclusion

Both dissimilarity matrices resulting from the text-based Levenshtein Distance and the acoustic mean pitch frequency differences can effectively be combined with multidimensional scaling and dendrograms to epitomize 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, used with both text-based and acoustic-based methods and data, are able to provide useful objective measures of language similarities or dissimilarities. It is clear that these methods can be refined further using other inputs such as phonetic transcriptions or further acoustic measurements; such refinements are likely to be important when, for example, fine distinctions between dialects are required.

However, 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 less discriminating in assessing languages with different writing styles, for example Chinese and English. Using pitch bares the advantage of using language data in its most natural form, but has its disadvantages in being computationally taxing when dealing with large amounts of data—which is generally required in order to produce good results.

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 different acoustic parameters, or a combination of parameters, is another pressing concern. Finally, it would be valuable to compare various distance measures against other criteria for language similarity (e.g. historical separation or mutual intelligibility) in a rigorous fashion.

References

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *Institute of Phonetic Sciences, vol 17*, pp 97-110.

Boersma, P, & Weenink, D. (2011). Praat Version 5.3 2011. http://www.fon.hum.uva.nl/praat/ *Date of last access: 27 July. 2012*

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.

Heeringa, W, & Gooskens, C. (2003). Norwegian dialects examined perceptually and acoustically. *Computers and the Humanities, 37*, pp. 293-315.

Heine, B, & Nurse, D. (2000). African languages: An introduction. Cambridge University Press.

Hirst, D, & Cristo, A Di. (1998). Intonation systems: A survey of twenty languages. *Cambridge University Press, Cambridge.*

Kessler, B. (1995). Computational dialectology in Irish Gaelic. *The 7th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 60-67.

Kometsu, M, Mori, K, Arai, T, & Murahara, Y. (2001). *Human language identification with reduced segmental information: comparison between Monolinguals and Bilinguals.* Paper presented at the EUROSPEECH, Scandanavia. pp 149-152.

Kruskal, J B. (1999). An overview of sequence comparison. Stanford. .

Mori, K, Toba, N, Harada, T, Arai, T, Kometsu, M, Aoyagi, M, & Murahara, Y. (1999). *Human language identification with reduced spectral information.* Paper presented at the EUROSPEECH, Budapest, Hungary. pp. 391-394.

Nerbonne, J, Heeringa, W, Hout, E Van den, Kooi, P Van der, Otten, S, & Vis, W Van de. (1996). Phonetic distance between Dutch dialects. *Sixth CLIN Meeting*, pp. 185-202.

Pokorny, F. (2011). Extraction of prosodic features from speech signals. Graz, Austria: Institute of Electronic Music and Acoustics, University of Music and Performing Arts.

Shriberg, E, Stolcke, A, Hakkani-Tur, D, & Tur, G. (2000). Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication. pp. 127-154*(32).

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 sociolinguistiche conferentie*, pp. 77-87.

Van-Hout, R, & Münstermann, H. (1981). Linguistic distance, dialect and attitude. *Gramma 5*, pp. 101-123.

XLSTAT. (2012). XLSTAT. http://www.xlstat.com/en/download/. Date of access: 27 July. 2012

Zulu, P N, Botha, G, & Barnard, E. (2008). Orthographic measures of language distances between the official South African languages *Literator: Journal of Literary Criticism, Comparative Linguistics and Literaty Studies 29*(1), 185.