Dialect Distances Based on Orthographic and Phonetic Transcriptions

N. Zulu and E. Barnard

Human Language Technologies Research Group
Meraka Institute, Pretoria, 0001, South Africa.

pzulu@csir.co.za, ebarnard@csir.co.za

Abstract – This paper describes ongoing work in which the main objective is to quantitatively determine the linguistic distances between languages and dialects. Here, we apply the Levenshtein distance measure to orthographic and phonetic transcriptions of words from 15 Norwegian dialects. Clustering of the distances between the different dialects shows the relationships between the dialects in terms of regional groupings and closeness. Although orthographic transcriptions generate distinctive north and south groupings, the more detailed phonetic transcriptions group the dialects more decisively into their regional groups. When the phonetic transcriptions are employed, the dendrogram of distances between regions is very similar to that computed from perceptual assessment of dialect distances.

1. Introduction

The development of objective metrics to assess the distances between different dialects and languages is of great theoretical and practical importance. Currently, subjective measures are generally employed to assess the degree of similarity between different languages, and those subjective decisions are, for example, the basis for classifying certain groups of language variants as dialects of one another, whereas others are considered separate languages. This has practical implications, since the distance between a pair of languages / dialects should serve as a useful indicator of how useful resources should be in developing language technologies for the other. For example, an understanding of language distances would enable researchers to use the data of a source language, to train initial speech-recognition models of a sufficiently close target language for which little data is available.

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 distance 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? While it might be easy to rank French as closer to English than Chinese, other rankings of closeness between other languages or dialects may be more difficult. The distance between languages may also depend on whether the languages are in text or acoustic form. For example, the written forms of Chinese do not vary greatly among the regions of China, but the spoken languages differ sharply.

This paper applies the *Levenshtein distance* to orthographic and phonetic transcriptions in order to obtain a measure of the distance between different dialects. These distances are employed in a hierarchical clustering of dialects, and compared to perceptual distance classification.

2. Levenshtein Distance

There are several ways in which phoneticians have tried to measure the distance between two basic sounds, 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.

One of the standard techniques for the computational comparison of sequences is known as the Levenshtein distance, also known as *string distance* or *edit distance*. In 1995 Kessler introduced the use of the Levenshtein distance as a tool for measuring linguistic distances between dialects [1]. 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 by the example in Table 1, in the transformation of the string '*æəftənʉn*' to the string '*æftərnun*' [2].

Table 1: Levenshtein distance between two strings.

	Operation	Cost
æəftənʉn		
æftənʉn	delete ə	1
æəftərnʉn	insert r	1
æftərnun	replace [ʉ] with [u]	2
	Total	4

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 phoneticians' work on the distance between

individual phonetic sounds. The operations used were: (i) the deletion of a single sound, (ii) the insertion of a single sound, and (iii) the substitution of one sound for another [3]. In our research, substitutions were allocated a cost of two while insertions and deletions each had a cost of one. In effect, a substitution is always equal to the combination of a deletion and an insertion and thus counts as the sum of these two operations separately.

The edit distance method was also taken up by Nerbonne *et al* [4] who applied it to Dutch dialects. In both cases the use of the Levenshtein distance was based on phonetic transcriptions, where transcription segments were compared using the algorithm.

In 2003 Gooskens and Heeringa [5] 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. This investigation was based on existing recordings and corresponding phonetic transcriptions of the same text read aloud in 15 Norwegian dialects. Here too, the Levenshtein distance measurements used were based on phonetic transcriptions.

3. Data representation

In this research we aim to compare languages and dialects based on their orthographic and phonetic transcriptions. Despite years of research, the field of multilingual speech processing has suffered from the lack of common public-domain multilingual speech data-sets that could be used to evaluate different approaches to the problem. It is therefore one of the greater future goals of this research to gather data for the target South African languages and dialects to be investigated. In order to carry out our investigations we need suitable data. We need to have access to recordings of the same text in a fair number of languages, and recordings of dialects from individual languages. At the same time digitized orthographic and phonetic level transcriptions of the data are required for calculating Levenshtein distances. This investigation used existing recordings and corresponding orthographic and phonetic transcriptions of text in 15 Norwegian dialects[1]. The recordings comprise 4 male and 11 female speakers reading the Norwegian version of the fable '*The North Wind and the Sun*' translated into their 15 respective dialects. Figure 1 shows the geographical distribution of the dialects. On this map six dialect areas (or groups) are represented. These groups are: Nordlandsk (No), Sørvestlandsk (Sv), Nordvestlandsk (Nv), Midlandsk (Mi), Austlandsk (Au) and Trøndsk (Tr).

[1] Data from the Department of Linguistics, University of Trondheim. Available at http://www.ling.hf.ntnu.no.nos.

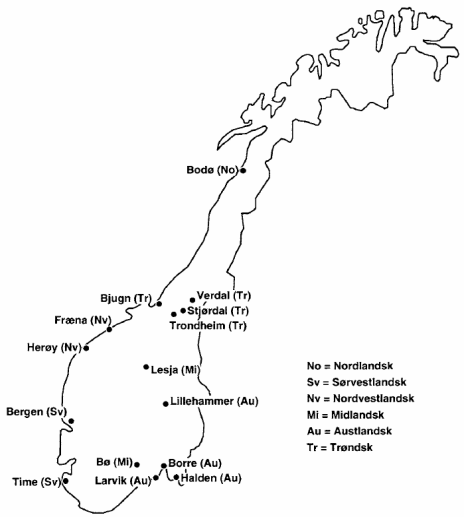

Figure 1: Map of Norway showing 15 dialects. The abbreviation after the name of each dialect indicates the dialect region to which it belongs [6].

Orthographic Transcriptions

This is 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 [7]. However, each English letter may represent more than one phoneme, and each phoneme may be represented by more than one letter.

Phonetic Transcriptions

There are over a hundred different phones recognized as distinctive by the *International Phonetic Association (IPA)* and transcribed in their International Phonetic Alphabet.

The Speech Assessment Methods Phonetic Alphabet (SAMPA) is a machine-readable phonetic alphabet developed by an international group of phoneticians in the late 1980's. SAMPA basically consists of a mapping of symbols of the International Phonetic Alphabet onto ASCII codes. In its basic form, SAMPA was seen as catering essentially for segmental transcriptions, particularly of a traditional phonemic kind. Prosodic notation was not adequately developed. Our investigation uses an extended version of the segmental alphabet, X-SAMPA, which extends the basic agreed conventions so as to make provision for every symbol on the IPA chart, including all diacritics and tone marks. In principle this makes it possible to produce a machine-readable phonetic transcription for every known human language. Table 2 shows an example of the differences in transcriptions of two words from two different dialects.

Table 2: Different representations of two words from two different dialects, Bergen and Bjugn.

Orthographic		**X-SAMPA**	
Bergen	*Bjugn*	*Bergen*	*Bjugn*
noravinn	nolavinnj	""nu:M\A%Pin:\	""nu:r`A%PiJ:\
kranglet	krangla	""kM\ANlet	""k4ANr`A

The Levenshtein distances in this study are based both on orthographic and phonetic transcriptions. For all 15 dialects, 50 similar words from each dialect were used.

4. Clustering Dialects

In using the Levenshtein distance measure, the distance between two dialects is equal to the average of a sample of Levenshtein distances of corresponding word pairs. When we have *n* dialects, then the average Levenshtein distance is calculated for each possible pair of dialects. For *n* dialects *n* x *n* pairs can be formed. The corresponding distances are arranged in a *n* x *n* matrix. The distance of each dialect with respect to itself is found in the distance matrix on the diagonal from the upper left to the lower right. These values are always zero and therefore give no real information, so that only *n* x (*n* - 1) distances are interesting. Furthermore, the Levenshtein distance is symmetric. This means that the distance between word *X* and word *Y* is equal to the distance between word *Y* and word *X*. The result is that distance between dialects *X* and *Y* is equal to the distance between dialects *Y* and *X* as well. Therefore, the distance matrix is symmetric. We need to use only one half which contains the distances of (*n* x (*n* - 1))/2 dialect pairs. Given the distance matrix, groups of larger sizes are investigated. *Hierarchical Clustering* methods are employed to classify the dialects into related dialect 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 (clusters), so that the data in each subset share some common trait according to a defined distance measure. The result of cluster analysis is usually illustrated as a *dendrogram*, a tree diagram used to illustrate the arrangement of the clusters produced by a clustering algorithm. Figure 2 illustrates the dendrogram derived from the clustering of perceptual distances as perceived by Norwegian listeners for the 15 Norwegian dialects investigated in this research [6].

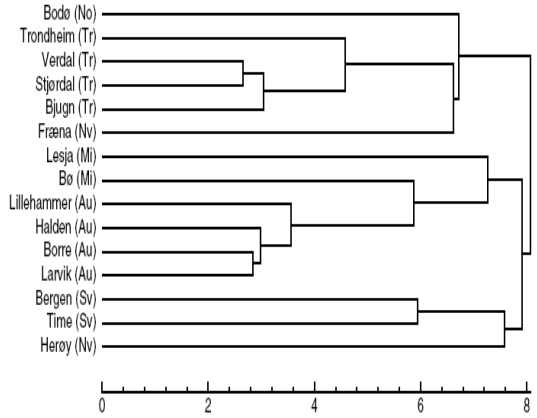

Figure 2: Dendrogram derived from the 15x15 matrix of perceptual distances showing the clustering of (groups of) Norwegian dialects [6].

Figure 3 shows the dendrogram produced by cluster analysis of the Levenshtein distances calculated from orthographic transcriptions of 50 words from the same 15 Norwegian dialects. Referring to Figure 1, the dendrogram appears to be divided into two main groups; a northern group and a southern group. The southern group consisting of the dialects; Lesja (Mi), Trondheim (Tr), Bergen (Sv), Time (Sv), Larvik (Au), Borre (Au), Lillehammer (Au) and Halden (Au), while the Northern group comprises the dialects; Bjugn (Tr), Bodø (No), Verdal (Tr), Fræna (Nv), Stjørdal (Tr), Bø (Mi) and Herøy (Nv). It is significant to note that two of the dialects have been misclassified. Bø (Mi), which is geographically located in the south has been grouped with the northern dialects, and Trondheim (Tr), which is geographically located in the north has been grouped with the southern dialects.

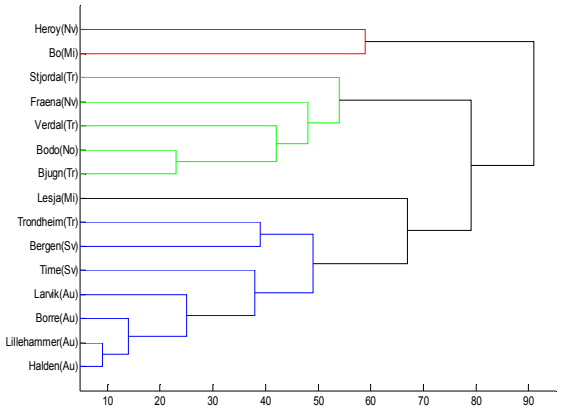

Figure 3: Dendrogram derived from Levenshtein distances of orthographic transcriptions of 50 words from 15 Norwegian dialects.

Figure 4 shows the dendrogram obtained on the basis of phonetic representation. Again the dendrogram shows a division into a northern and southern group, with the exception of Lesja (Mi) being misclassified. It is also important to note that the dialects are more group oriented than in Figure 3. South-western and south-eastern groups are more clearly defined as well as north-western and north-eastern groups. Similar to the orthographic transcriptions, the phonetic transcriptions based results do not show as great a distinction between dialect groups as the perceptual results do (Figure 2), although the clustering based on phonetic transcriptions certainly is more comparable to the perceptual results.

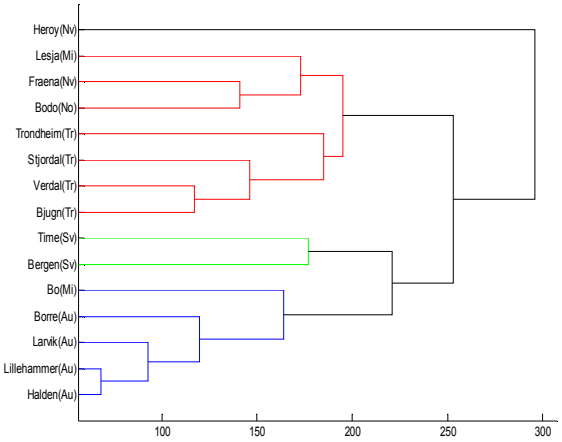

Figure 4: Dendrogram derived from Levenshtein distances of phonetic transcriptions of 50 words from 15 Norwegian dialects.

5. Conclusions

The results of this paper primarily reinforce the previous work of researchers in trying to classify languages and dialects based on linguistic distances. Here we investigated the distance between dialects based on both their orthographic and phonetic transcriptions. The results show that of the two transcription methods, the Levenshtein distances based on the phonetic transcriptions more closely match the perceptual distances and group the dialects into their regional groups more closely. Thus, when time and human resources are limited for perceptual evaluations to be carried out, phonetic transcriptions can be used as a substitute to classify dialects and languages.

6. Future Directions

This paper presented a primary investigation into the much broader topic of language distance. We would like to extend this work in several directions.

We aim to find acoustic distance measures between languages and dialects which approximate perceptual distance measures as perceived by humans. This in itself encompasses a number of objectives. One of the objectives is to provide a comprehensive review of existing language distance literature with particular emphasis on data representation and classification algorithms that have previously been used.

The second and main objective will be to design and implement a comprehensive language and dialect acoustic distance measure incorporating both old and new techniques and algorithms. Perceptual measures and Levenshtein distances calculated on the basis of transcriptions will form a baseline for the comparison of results obtained in our research. This will be coupled with measures of statistical relevance of the various measures employed. The baseline will act as the point of reference to which all developed and other commonly used algorithms will be compared. The main evaluation of the system will be based for the most part on its performance on South African languages. Experiments to evaluate the effectiveness of work in this research will be performed on suitable existing speech data and other data collected in the region of South Africa. We will use and compare different representations of the acoustic signals, different distance measures and classification algorithms.

It is not the aim of this project to present a complete solution to the problem but simply to contribute to the immense ongoing research on language distance that already exists.

Acknowledgments

We are grateful to Laurens Cloete, who first suggested the task of the objective measurement of language distances in the South African context to us.

References

[1] B. Kessler, "Computational Dialectology in Irish Gaelic," presented at The 7th Conference of the European Chapter of the Association for Computational Linguistics, Dublin, 1995.

[2] J. Nerbonne, "Computational Contributions to the Humanities," presented at The Joint Conference of The Association for Literary and Linguistic Computing and The Association for Computers and the Humanities at the University of Gothenburg., Gothenburg, Sweden, 2004.

[3] J. B. Kruskal, *An Overview of Sequence Comparison*, 2nd ed: Stanford, 1999.

[4] J. Nerbonne, W. Heeringa, E. V. d. Hout, P. V. d. Kooi, S. Otten, and W. V. d. Vis, "Phonetic Distance Between Dutch Dialects," presented at Sixth CLIN Meeting, University of Antwerp,

Center for Dutch Language and Speech, Antwerp, 1996.

[5] C. Gooskens and W. Heeringa, "Perceptive Evaluation of Levenshtein Dialect Distance Measurements Using Norwegian Dialect Data," *Language Variation and Change*, vol. 16, pp. 189-207, 2004.

[6] W. Heeringa and C. Gooskens, "Norwegian Dialects Examined Perceptually and Acoustically," *Computers and the Humanities*, pp. 293-315, 2003.

[7] Wikipedia, "Orthography," http://en.wikipedia.org/wiki/Orthography, [Last Accessed: 20/10/2006], 2006.