Stellenbosch Papers in Linguistics Plus, Vol. 47, 2015, 1-18
doi: 10.5842/47-0-649

Afrikaans and Dutch as closely-related languages: A comparison to West Germanic languages and Dutch dialects

Wilbert Heeringa

Institut für Germanistik, Fakultät III – Sprach- und Kulturwissenschaften, Carl von Ossietzky Universität, Oldenburg, Germany
Email: wjheeringa@gmail.com

Febe de Wet

Human Language Technology Research Group, CSIR Meraka Institute, Pretoria, South Africa | Department of Electrical and Electronic Engineering, Stellenbosch University, South Africa
Email: fdwet@csir.co.za

Gerhard B. van Huyssteen

Centre for Text Technology (CTexT), North-West University, Potchefstroom, South Africa
Email: gerhard.vanhuyssteen@nwu.ac.za

Abstract

Following Den Besten’s (2009) desiderata for historical linguistics of Afrikaans, this article aims to contribute some modern evidence to the debate regarding the founding dialects of Afrikaans. From an applied perspective (i.e. human language technology), we aim to determine which West Germanic language(s) and/or dialect(s) would be best suited for the purposes of recycling speech resources for the benefit of developing speech technologies for Afrikaans. Being recognised as a West Germanic language, Afrikaans is first compared to Standard Dutch, Standard Frisian and Standard German. Pronunciation distances are measured by means of Levenshtein distances. Afrikaans is found to be closest to Standard Dutch. Secondly, Afrikaans is compared to 361 Dutch dialectal varieties in the Netherlands and North-Belgium, using material from the *Reeks Nederlandse Dialectatlassen*, a series of dialect atlases compiled by Blancquaert and Pée in the period 1925-1982 which cover the Dutch dialect area. Afrikaans is found to be closest to the South-Holland dialectal variety of Zoetermeer; this largely agrees with the findings of Kloeke (1950). No speech resources are available for Zoetermeer, but such resources are available for Standard Dutch. Although the dialect of Zoetermeer is significantly closer to Afrikaans than Standard Dutch is, Standard Dutch speech resources might be a good substitute.

Keywords: human language technologies, speech resources, Afrikaans, Dutch, acoustic distance

1. Introduction

The development of language resources for use in human language technologies (HLTs) is time-consuming, tedious and expensive, both in terms of human- and other resources. Development can be accelerated if existing resources from closely-related languages can be used in one way or another. A popular theme in the fields of speech and language processing is therefore to find innovative ways to expedite this process as cost effectively as possible, especially for so-called “resource scarce” languages (i.e. languages without sufficient annotated electronic data that would enable one to use statistical approaches to speech and language processing). Because HLT is still a relatively new field in South Africa, most of the South African languages are severely under-resourced in terms of the data and software required to develop HLT applications, such as automatic speech recognition engines, speech synthesis systems, etc.

One of the approaches to developing resources for such languages is an approach where one uses data and/or technologies from a well-resourced language (L1; for example, Dutch) to assist in the development of resources for a closely-related, under-resourced language (L2; in this case, Afrikaans). The basic hypothesis is that “[if] the languages L1 and L2 are similar enough, then it should be easier [and quicker] to recycle software applicable to L1 than to rewrite it from scratch for L2 [thereby taking care of] most of the drudgery before any human has to become involved” (Rayner, Carter, Bretan, Eklund, Wirén, Hansen, Kirchmeier-Andersen, Philp, Sørensen and Thomsen 1997: 65). One therefore “recycles” resources from one language for the benefit of another language, hence referring to this approach as a “recycling approach”.

In a research project on data and technology transfer between closely-related languages, we explore various ways of recycling Dutch resources for the benefit of Afrikaans, including both text and speech resources (see Van Huyssteen and Pilon 2009). As a point of departure, we make the basic assumption that Afrikaans and Dutch are indeed closely-related languages,[1] based on:

1. the genealogical fact that both languages originate from the colloquial Dutch of the 17th century which belongs to Low Franconian (also referred to as “Frankish”), which in turn belongs to West Germanic (Van der Merwe 1951,1968), and

2. the popular belief that Afrikaans and Dutch are by and large mutually intelligible (see, for example, entries on Afrikaans as a language on www.en.wikipedia.org or www.urbandictionary.com; compare also Gooskens and Bezooijen 2006, and Bezooijen and Gooskens 2006 for supporting research evidence).

In this article, our focus is restricted to speech resources. We are particularly interested in constructing a large vocabulary continuous speech recognition system for Afrikaans. One of the resources required to develop such a system is a large quantity of annotated audio data. Given that very little Afrikaans data is currently available, we would like to investigate the possibility of using Dutch data to accelerate the development process for Afrikaans. For example, existing acoustic models for Dutch could be used to transcribe Afrikaans data automatically, given a mapping between the two languages' phone sets and an appropriate pronunciation dictionary. Dutch data could also be used to bootstrap a first set of acoustic models for Afrikaans. These models can initially be adapted with the limited Afrikaans data that is available and may eventually be replaced by "home grown" models when an adequate amount of transcribed data has been accumulated for Afrikaans.[2]

---

[1] Hajič, Hric and Kuboň (2000) distinguish between “language variants” (considered to be one language, e.g. Hollandic and Flemish), “very close languages” (similarity in morphology, syntax and lexis, e.g. Dutch and Afrikaans), “closely-related languages” (similarity in morphology and lexis, e.g. Dutch and German) and “related languages” (shared origin and influences without necessarily sharing linguistic similarities, e.g. Dutch and Swedish). For our purposes, we consider Afrikaans and Dutch to be somewhere between “very close” and “closely-related” on the continuum, but use the term “closely-related” throughout this article.

Although the assumptions we make intuitively seem valid enough, we would like to provide at least some experimental evidence to support these claims. Specifically, the aim of this article is to answer the following sets of questions:

1. Is Dutch, acoustically speaking, indeed the closest West Germanic language to Afrikaans? Can we prove that Standard Dutch is significantly closer to Standard Afrikaans (both from the Low Franconian group) than, say, Standard German (as an example of the High German group) or Standard Frisian (as an example of the Frisian group)?[3]

2. If so, are there Dutch dialects which are closer to Afrikaans than Standard Dutch is? If this is so, which one is closest and would therefore be better suited for our purposes of technology recycling? For example, Afrikaans tourists often claim that they understand Flemish (spoken mainly in Belgium) better than Hollandic (spoken in the urban centre of the Netherlands and is mostly the basis for Standard Dutch). Hence, is there any acoustic evidence that Flemish is closer to Afrikaans than Hollandic? For that matter, which dialect of Dutch is closest to Afrikaans and would therefore be best suited to achieve our goals?

3. If dialects are found which are closer to Afrikaans than Standard Dutch, is the closest one significantly closer to Afrikaans than Standard Dutch is? This is important since language technology is usually developed for standard languages, not for dialects.

The aim of the study is therefore to provide a hypothesis regarding which West Germanic language(s) and/or dialect(s) might be best to use for the development of speech technology applications for Afrikaans, using a recycling approach. Given that we focus on acoustic data, we will attempt to quantify the relationship between the pronunciation of Afrikaans and other West Germanic languages (i.e. Standard Dutch, Standard Frisian and Standard German) and 361 Dutch dialects in terms of an acoustic distance measure. The pronunciation distances we report on here were determined using the Levenshtein distance, a string edit distance measure first used by Kessler (1995) for measuring linguistic distances.

In section 2 of this article, we provide a brief perspective on some conflicting theories regarding the origin of Afrikaans, indicating that it is recognised to be quite difficult to determine which dialect of Dutch could be considered the basis for modern-day Afrikaans. In section 3, we give a description of our methodology, focusing both on the data and algorithm we use in our research. Section 4 presents our results, while section 5 concludes and presents some directions for future research.

---

[2] The technology referred to here is envisaged for Standard Afrikaans only and currently does not include one of the other two main dialects, viz. Cape Afrikaans and Orange River Afrikaans. In the remainder of this article, the term "Afrikaans" will therefore refer to Standard Afrikaans.

[3] Within the scope of this article, we omit English, which is considered the other major language in the West Germanic group.

2. Theories about the relationship between Afrikaans and Dutch

Much has been written about the relation between Afrikaans and Dutch, both from a diachronic perspective (i.e. the history of Afrikaans) and from a synchronic perspective (i.e. similarities and differences between modern Afrikaans and Dutch). Since our research concerns developing resources for modern-day Afrikaans, our concern is more a synchronic one. For comparisons between Afrikaans and Dutch, see De Villiers (1978), Conradie (1986), Ehlers and Beek (2004) and Van Huyssteen and Pilon (2009), amongst others.

In order to contextualise our research (and some of our findings), we provide a brief perspective on some of the different theories related to the history of Afrikaans. De Kleine (1997) points out that there are generally two kinds of theories about the origin of the language: those theories that claim that Afrikaans can be traced mainly to 17th century varieties of Dutch (De Villiers 1978, Raidt 1991), and those theories that claim that a pidgin or creole was once spoken in the Cape Colony which strongly influenced the variety of Dutch that later developed into Afrikaans (Den Besten 1989). Although our research does not necessarily aim to contribute to this theoretical debate, our assumptions could be seen as belonging more to the former group of theories, although we do not deny any evidence of the complex language contact situation during the historical development of Afrikaans.

For pragmatic purposes, we assume that Afrikaans can be considered a daughter language of Dutch, diverging from the latter during the last half of the 17th century. Although there is evidence of language contact between the Dutch and the Khoi (the original inhabitants of the area that would later become known as the Cape of Good Hope) as early as the late 16th century, the formative years of Afrikaans can be set from 1652 onwards, when Jan van Riebeeck founded a refreshment station at the Cape of Good Hope on the way to the Indies, and formally introduced a variety of Dutch to this region. According to Van Reenen and Coetzee (1996), Van Riebeeck and his group of settlers came from the southern part of the Dutch province of South-Holland, and it is therefore easy to assume that the variety of Dutch that they spoke (i.e. South-Hollandic) would be the main basis for Afrikaans. The famous Dutch dialectologist G.G. Kloeke (1950: 262-263) writes in his *Herkomst en Groei van het Afrikaans* ("Origin and Growth of Afrikaans") that the old dialects of South-Holland on the one hand and "High" Dutch on the other are the chief sources of Afrikaans.

In contrast, Scholtz (1963) does not agree with Kloeke but wonders whether Afrikaans is derived from a common Hollandic language, the Hollandic norm of the second half of the 17th century. However, Van Reenen and Coetzee (1996) doubt whether a common Hollandic language already existed in that period.

Regarding these contradictory points of view, De Villiers (1978) unequivocally states that it is difficult to determine which Hollandic dialects have had the most influence on Afrikaans. Den Besten (2009) echoes this when he argues that research regarding the founding dialects of Afrikaans would not be simplistic. He continues to identify this difficult debate on the founding dialects of Afrikaans as a desideratum for historical linguistics of Afrikaans, but warns that results should be presented in a careful and nuanced way. As is clear from this discussion, this remains a difficult question to answer (especially in the absence of representative corpora from the time), but we believe that the methodology that we employ for our current research could, in addition to addressing our main goals, shed light on the relationship (i.e. closeness) between Standard Afrikaans and various Dutch dialects.

3. Methodology

3.1 Data sources

3.1.1 Dutch dialects

In order to study the relationship between Afrikaans and Dutch dialectal varieties, it would be preferable to use data from around 1652, the time period coinciding with Jan van Riebeeck's influence on the Afrikaans language. Of course, we do not have phonetic transcriptions from that time. The oldest available source containing phonetic transcriptions of a dense sample of dialect locations is the *Reeks Nederlandse Dialectatlassen* (RND), a series of Dutch dialect atlases which were edited by Blancquaert and Pée (1925-1982). The atlases cover the Dutch dialect area, i.e. the Netherlands, the northern part of Belgium, a smaller north-western part of France and the German county of Bentheim.

In the RND, the same 141 sentences are translated and transcribed in phonetic script for each dialect. Blancquaert (1939) mentions that the questionnaire was conceived as a range of sentences with words that illustrate particular sounds. The design saw to it that, for example, possible changes of Old Germanic vowels, diphthongs and consonants are represented in the questionnaire. Since digitising the phonetic texts is time-consuming, and since the material was intended to be processed by the word-based Levenshtein distance, a set of only 125 words was selected from the text (Heeringa 2001). The words were selected more or less randomly and may be considered a random sample. The transcriptions of the 125 word pronunciations were digitised for each dialect. The words represent (nearly) all vowels (monophthongs and diphthongs) and consonants. The consonant combination [sx] is also represented, which is pronounced as [sk] in some dialects and as [ʃ] in others.

The RND contains transcriptions of 1956 Dutch varieties. Since it would be very time-consuming to digitise all transcriptions, a selection of 361 dialects was made (Heeringa 2001). The dialects were selected with the aim to obtain a net of evenly scattered dialect locations. A denser sampling was used in the areas of Friesland and Groningen, and in the area in and around Bentheim. In Friesland, the Town Frisian dialect islands were added to the set of varieties which belong to the (rural) Frisian dialect continuum. In Groningen, some additional localities were added because of personal interest. In the area in and around Bentheim, additional varieties were added because of a detailed investigation in which the relationship among dialects on both sides of the border was studied. In addition, the dialects' relationship to Standard Dutch and Standard German was studied (Heeringa 2001).

In the RND, the transcriptions are noted in a predecessor of the International Phonetic Alphabet (IPA). The transcriptions were digitised using a computer phonetic alphabet which might be considered a dialect of X-SAMPA. The data is freely available at http://www.let.rug.nl/~heeringa/dialectology/atlas/rnd/.

3.1.2 Languages

In this article, Dutch dialects are compared to Afrikaans. The 125 words selected from the RND sentences were therefore translated into Afrikaans and pronounced by an older male and a young female, both native speakers of Afrikaans. Older males are known to be conservative speakers, while young females are usually innovative speakers (Hinskens, Auer and Kerswill 2005). Our measurements reflect the average of the two speakers when we compare Dutch dialects to Afrikaans. The pronunciations of the two speakers were transcribed consistently with the RND transcriptions.

Afrikaans is also compared to Standard Dutch, Standard Frisian and Standard German. Although Standard Afrikaans is not as well-defined as its European counterparts, care was taken not to use speakers with a strong regional accent in this study. To ensure consistency with the existing RND transcriptions, the Standard Dutch transcription is based on Blancquaert's (1939) *Tekstboekje*. However, words such as *komen*, *rozen* and *open* are transcribed as [koˑmə], [roːzə] and [oˑpə], respectively. In *Tekstboekje* (Blancquaert 1939), these words would end on a [n], as suggested by the spelling. For more details, see Heeringa (2001).

The RND transcription of the Frisian variety of Grouw was used as Standard Frisian, since Standard Frisian is known to be close to the Grouw variety.

The Standard German word transcriptions are based on *Wörterbuch der deutschen Aussprache* (Krech and Stötzer 1969). However, the transcriptions were adapted so that they are consistent with the RND data. In the dictionary, the <r> is always noted as [r], never as [R]. Because both realisations are allowed in German, two variants are noted for each pronunciation containing one or more <r>'s – one in which the [r] is pronounced and another in which the [R] is pronounced. More details are given in Heeringa and Nerbonne (2000). Both realizations were taken into account in the experiment reported on in this article.

3.2 Measuring pronunciation distances

As previously mentioned, pronunciation differences are measured with the Levenshtein distance which was first applied by Kessler (1995) to transcriptions of Irish Gaelic dialectal varieties. The Levenshtein distance was later applied to Dutch dialects by Nerbonne, Heeringa, Den Hout, Van der Kooi, Otten and Van de Vis (1996; more detailed results are given in Heeringa 2004), to Sardinian by Bolognesi and Heeringa (2002), to Norwegian by Gooskens and Heeringa (2004), to German by Nerbonne and Siedle (2005), to Bantu by Alewijnse, Nerbonne, Van der Veen and Manni (2007), to Bulgarian by Heeringa, Nerbonne and Osenova (2010) and to American English by Nerbonne (2015). The Levenshtein distance corresponds to the distance between the transcriptions of two pronunciations of the same concept corresponding to two different varieties. The distance is equal to the minimum number of insertions, deletions and substitutions of phonetic segments needed to transform one transcription into another. The distance between two varieties is based on several pronunciation pairs, in our case 125. The corresponding Levenshtein distances are averaged.

Pronunciation variation includes variation in sound components and morphology. The items to be compared should have the same meaning and should be cognates.

3.2.1 Algorithm

Using the Levenshtein distance, two varieties are compared by measuring the pronunciation of words in the first variety against the pronunciation of the same words in the second (Kruskal 1999). We determine how one pronunciation might be transformed into the other by inserting, deleting or substituting sounds. In this way, distances between the transcriptions of the pronunciations are calculated. Weights are assigned to these three operations; in the simplest form of the algorithm, all operations have the same cost. Assume, for example, the Standard Dutch word *hart* ('heart') is pronounced as [hɑrt] in Afrikaans and as [ærtə] in the East Flemish dialect of Nazareth (Belgium). Changing one pronunciation into the other can be done as follows:

Table 1: hɑrt → ærtə

hɑrt	delete h	1
ɑrt	replace ɑ with æ	1
ært	insert ə	1
ærtə		
		3

In fact, many string operations map [hɑrt] to [ærtə]. The power of the Levenshtein algorithm is that it always finds the least costly mapping.

To deal with syllabification in words, the Levenshtein algorithm was adapted so that it did not allow alignments of vowels with consonants (Heeringa 2004). In this way, unlikely mappings (e.g. a [p] with an [a]) were prevented. Exceptions were the semivowels [j] and [w] and their respective vowel counterparts [i] and [u], which may match with anything. Additionally, we allowed the schwa to be aligned with a sonorant (and vice versa). It is not unusual that, e.g. a [r] matches with an [ə]. For example, two possible pronunciations for the Dutch word *vier* ('four') are [fiːr] and [fiːə]. Here we wanted the ending [r] and the ending [ə] to match with each other. In our example we thus have the following alignment:

Table 2: Alignment of hɑrt → ærtə

h	ɑ	r	t	
	æ	r	t	ə
1	1			1

This corresponds to a total cost of three operations and an alignment length of 5. Aggregated distances between multiple words can also be combined to calculate the pronunciation distance between two dialects. For example, if four words are taken into consideration to calculate the distance between Afrikaans and the Nazareth dialect, the "total" pronunciation distance can be calculated, as shown in Table 3.[4]

---

[4] In order to keep the example clear, diacritics are ignored and all operation costs have a weight of 1.

Table 3: Calculation of the aggregated pronunciation distance between Afrikaans and Nazareth on the basis of four word pairs

Dutch	English	Afrikaans	Nazareth	distance	alignment length
werk	work	ʋærk	wɪrək	3	5
schip	ship	sxʏp	sxep	2	4
brood	bread	brʊt	bryət	2	5
jaar	year	jɑr	jɔr	1	3
				8	17

This result can also be expressed in terms of a percentage, i.e. 8/17 × 100 = 47%. In this article, aggregated Levenshtein distances were obtained on the basis of 125 word pairs (see section 3.2).

3.2.2 Operation weights

The simplest version of this method is based on a notion of phonetic distance in which phonetic overlap is binary; non-identical phones contribute to phonetic distance and identical ones do not. Thus the pair [i,ɒ] differs to the same degree as [i,ɪ]. The version of the Levenshtein algorithm used in this article is based on the comparison of spectrograms of the sounds. Since a spectrogram is the visual representation of the acoustic signal, the visual differences between the spectrograms are reflections of the acoustic differences.

The spectrograms were made on the basis of recordings of the IPA sounds as pronounced by John Wells and Jill House on the cassette *The Sounds of the International Phonetic Alphabet* (Wells and House 1995). The different sounds were isolated from the recordings and monotonised at the mean pitch of each of the two speakers with the program PRAAT (Boersma and Weenink 2002). Next, for each sound a spectrogram was made with PRAAT using the Bark filter, a perceptually-oriented model. A Bark filter is created from a sound by band-filtering in the frequency domain with a bank of filters. In PRAAT, the lowest band has a central frequency of 1 Bark per default, and each band has a width of 1 Bark. There are 24 bands corresponding to the first 24 critical bands of hearing as found along the basilar membrane (Zwicker and Fastl 1990). A critical band is an area within which two tones influence each other's perceptibility (Rietveld and Heuven 1997). Due to the Bark scale, the higher bands summarise a wider frequency range than the lower bands.

Segment distances were calculated based on the Bark filter representation. Inserted or deleted segments were compared to silence, and silence was represented as a spectrogram in which all intensities of all frequencies are equal to 0. The [ʔ] was found closest to silence and the [a] was found most distant. This approach is described extensively in Heeringa (2004).

In perception, small differences in pronunciation may play a relatively strong role in comparison to larger differences. Therefore, logarithmic segment distances were used. The effect of using logarithmic distances is that small distances are weighted relatively more heavily than large distances, and these weights will vary between 0 and 1. In a validation study, Heeringa (2004) found that among several alternative distances obtained with the Levenshtein distance measure, using logarithmic Bark filter segment distances gives results which most closely approximate dialect distances as perceived by the speakers themselves.

3.2.3 Vowels and consonants

In addition to calculating Levenshtein distances based on all segments (full pronunciation distance), we also calculated distances based on vowels only and consonants only. If distances were calculated solely on the basis of vowels, initially the full phonetic strings were compared to each other using the Levenshtein distance.[5] Once the optimal alignment was found, the distances were based on the alignment slots which represent vowel substitutions. Consonant substitutions were calculated *mutatis mutandis*.

3.2.4 Processing RND data

The RND transcribers used slightly different notations. In order to minimise the effect of these differences, we normalised their data. The consistency problems and the way we solved them are discussed extensively in Heeringa (2001) and Heeringa (2004). For the same reason, only a part of the diacritics found in the RND was used.

As in earlier studies, we processed diacritics for length (extra short, half long, long), syllabicity (syllabic), voice (voiced, voiceless) and nasality (nasal) (Heeringa 2004). In this study, the diacritic for rounding (rounded, partly rounded, unrounded, partly unrounded) was used. The distance between, for example, [a] and rounded [i] was calculated as the distance between [a] and [y]. The distance between [a] and partly rounded [i] is equal to the average of the distance between [a] and [i] and the distance between [a] and [y]. The diacritic for rounding is important in our analysis since [ɯ] and [ɤ] are not included in the phonetic transcription system of the RND, but transcribed as unrounded [u] and [o], respectively.

The distance between a monophthong and a diphthong was calculated as the mean of the distance between the monophthong and the first element of the diphthong and the distance between the monophthong and the second element of the diphthong. The distance between two diphthongs was calculated as the mean of the distance between the first elements and the distance between the second elements. Details are given in Heeringa (2004).

4. Results

4.1 Finding the closest West Germanic language

In this section, we will answer the first research question mentioned in section 1: Is Dutch, acoustically speaking, indeed the closest West Germanic language to Afrikaans? In the same section, we found from literature that Afrikaans belongs to the West Germanic languages. In order to answer our first research question, we compared Afrikaans to the other West Germanic languages, namely Standard Dutch, Standard Frisian and Standard German. We calculated Levenshtein distances in the manner described in section 3.2 and obtained the distances as given in Table 4.

---

[5] Consequently, in the case of separate vowel and consonant distances, [j] and [w] are also considered as vowels, and [i] and [u] are also considered as consonants.

Table 4: Afrikaans compared to the other West Germanic languages – full pronunciation distances and distances obtained on the basis of vowel substitutions or consonant substitutions only

	Full pronunciation	Vowel substitutions	Consonant substitutions
Dutch	34%	11%	11%
Frisian	43%	14%	7%
German	50%	12%	14%

When we look at the full pronunciation distances, we find that Afrikaans is most closely related to Standard Dutch. Standard Dutch is also significantly closer to Afrikaans than Standard Frisian (t=5.096, n=125, p<0.001) and Standard German (t=10.861, n=125, p<0.001). This confirms the finding as suggested by, amongst others, Kloeke (1950), Van Reenen and Coetzee (1996) and Gooskens and Bezooijen (2006).

When we look at the vowel substitution distances, Afrikaans is still closest to Standard Dutch; Standard Dutch is significantly closer to Afrikaans than Standard Frisian (t=3.381, n=125, p<0.001), but is not significantly closer than Standard German (t=1.226, n=125, p=0.112).

When we look at the consonant substitution distances, Afrikaans is closest to Standard Frisian. Standard Frisian is significantly closer to Afrikaans than both Standard Dutch (t=3.771, n=125, p<0.001) and Standard German (t=5.979, n=125, p<0.001). This result may be unexpected, but consonant features which were lost in both Standard Dutch and Dutch dialects and which are still found in Afrikaans may have been retained by Standard Frisian (and varieties of Frisian) as well. We come back to this in section 4.2.2.

4.2 Finding the closest Dutch dialect

In the previous section, we compared Afrikaans to the other West Germanic standard languages and found Standard Dutch to be the closest. In this section, we answer our second research question: Are there Dutch dialects that are closer to Afrikaans than Standard Dutch? The search for the closest West Germanic variety is continued by comparing Afrikaans to the Dutch dialects. In addition, Frisian varieties are considered as we found that Standard Frisian is closest to Afrikaans when distances are measured on the basis of consonant substitutions only. Distances between 361 Dutch and Frisian dialects and Afrikaans were measured with the Levenshtein distance. The results are shown in Figure 1.

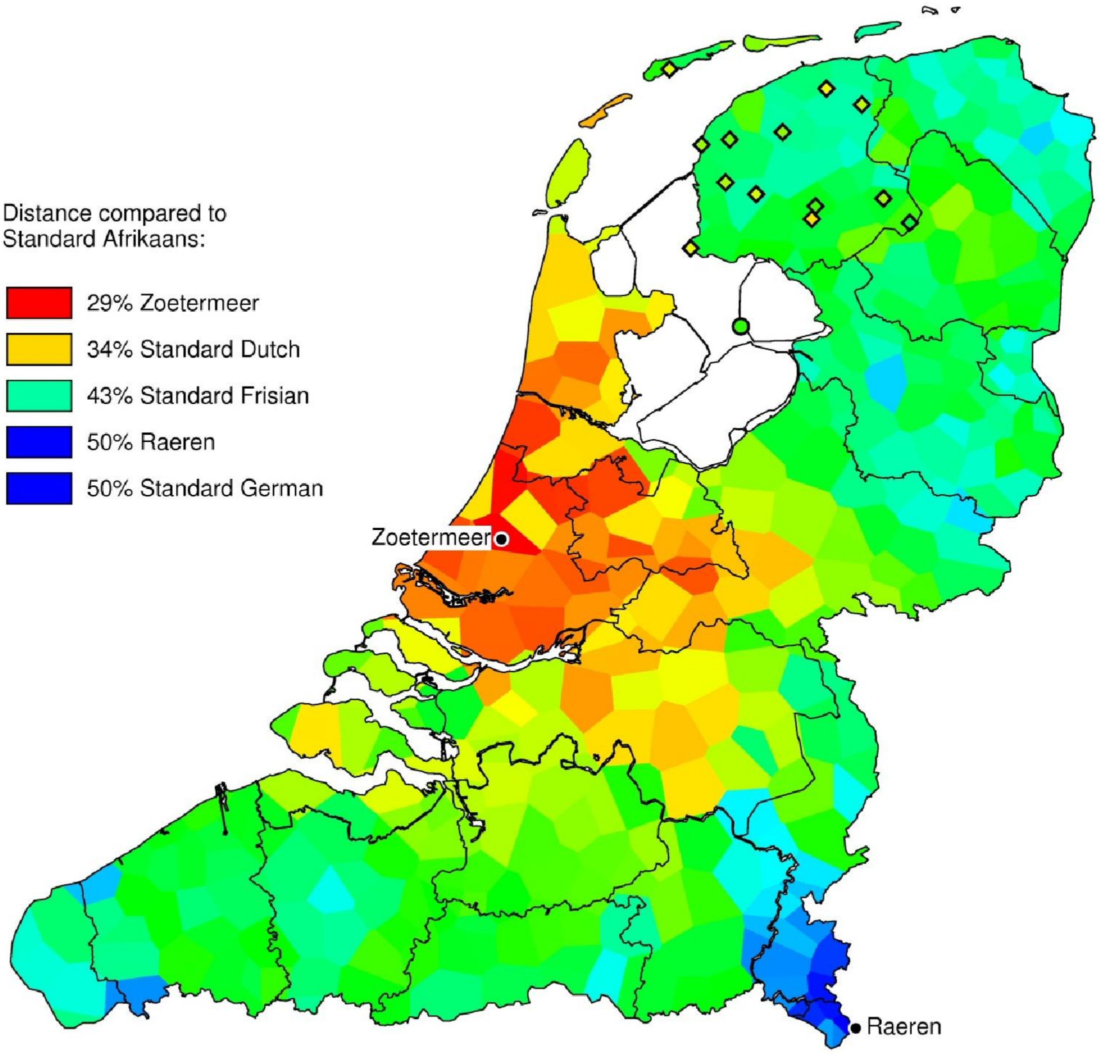

Figure 1. Distances of 361 Dutch dialectal varieties compared to Afrikaans

In this map, the varieties are represented by polygons, geographic dialect islands are represented by coloured dots, and linguistic dialect islands are represented by diamonds. Lighter polygons, dots or diamonds represent dialects which are close to Afrikaans and darker ones represent the varieties which are more distant. The distances in the legend represent the average Levenshtein distances. (The Ijsselmeer polders – Wieringermeerpolder, Noordoostpolder and Flevopolder – are not under consideration, so they are left white.)

The closest varieties were found in the province of South-Holland, with the dialect of Zoetermeer closest to Afrikaans (distance of 29%). This corresponds with Kloeke (1950) who claimed that the dialect of the first settlers was the main source of Afrikaans. These settlers came from the southern part of the Dutch province of South-Holland, the area around Rotterdam and Schiedam; Zoetermeer is slightly north of these two locations.

Some close varieties were also found in the provinces of North-Holland and Utrecht. The dialects in the southern part of Limburg were found to be most distant, where the dialect of Raeren was furthest away from Afrikaans (50%).

4.2.1 Vowels

Distances between Dutch dialects and Afrikaans based solely on vowel substitutions are shown in Figure 2.

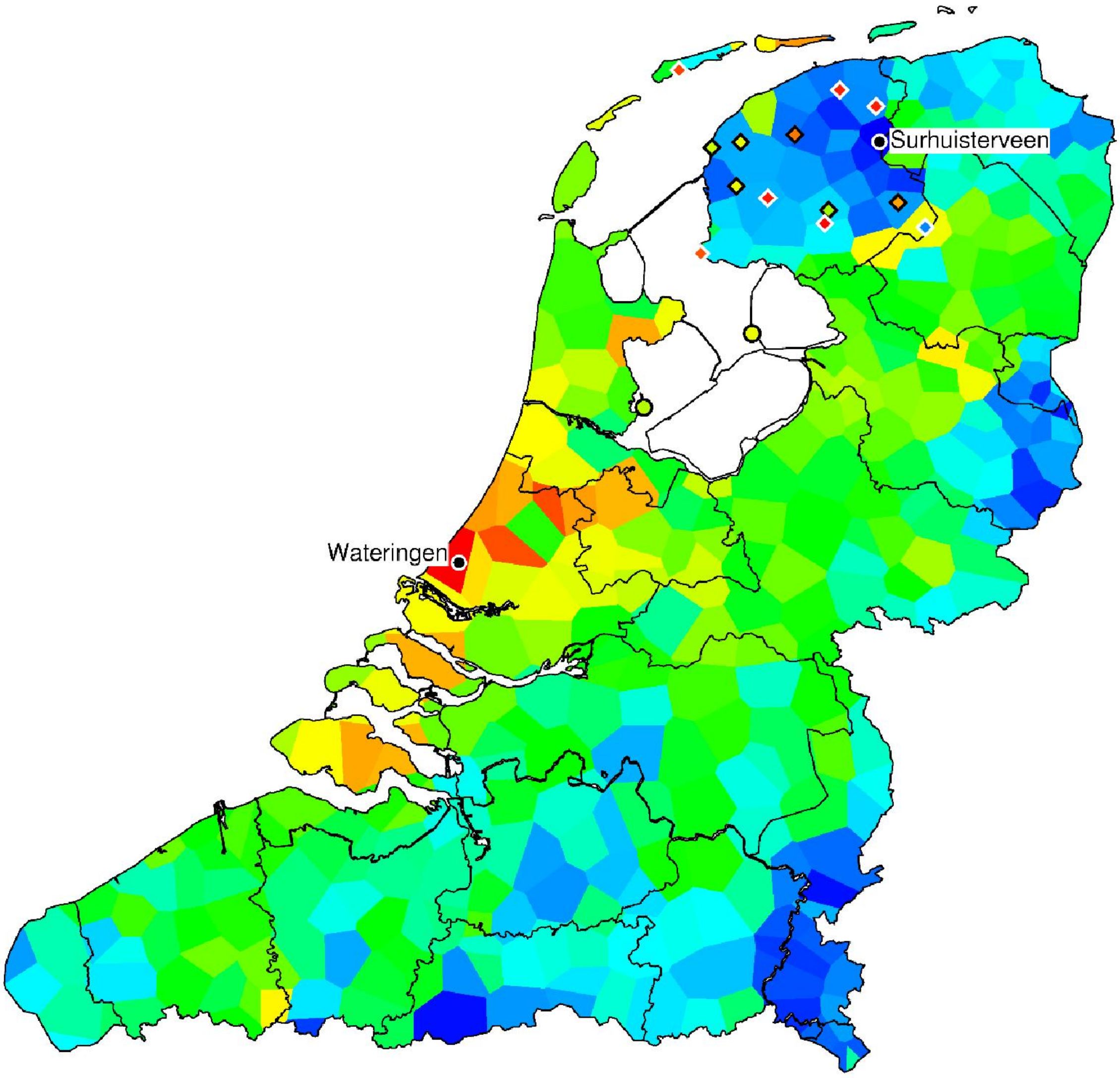

Figure 2. Vowel substitution distances between 361 Dutch dialectal varieties and Afrikaans

Again, the South-Hollandic varieties were relatively close to Afrikaans. This finding agrees with Kloeke (1950). In the summary of his book, Kloeke (1950: 262-263) writes:

> The two chief sources of Afrikaans, the old dialects of South Holland on the one hand and the "High" Dutch on the other, are reflected in the vowel system. In some respect Afrikaans is of a pronounced conservative "Holland" dialectal character, still more conservative than the dialects of Holland itself, which are gradually disappearing.

Although the Holland dialects have changed substantially since Jan van Riebeeck entered the Cape of Good Hope in 1652, the relationship to the South-Holland varieties is still found when we use the RND data.

The Frisian, Twente and Limburg varieties were found to be distant to Afrikaans. The varieties close to the Dutch/French border in the Belgian province of Brabant were also relatively distant. Most distant was the Frisian variety of Surhuisterveen (15.0%).

4.2.2 Consonants

When consonant distances between the Dutch dialects and Afrikaans were calculated, a completely different picture was obtained, as can be seen in Figure 3.

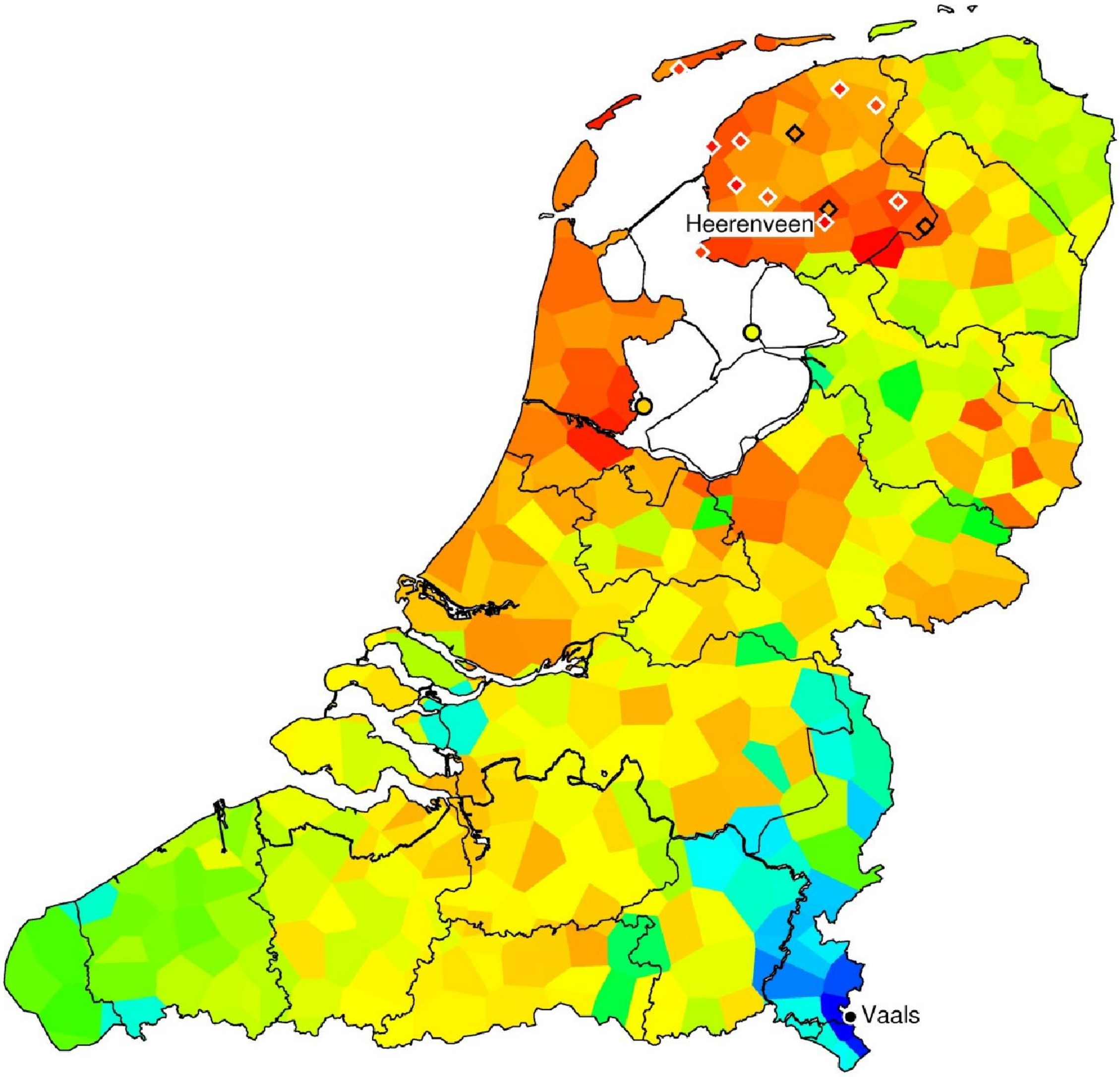

Figure 3. Consonant substitution distances between 361 Dutch dialectal varieties and Afrikaans

In terms of consonant substitutions, the Frisian varieties and the North-Holland dialects were found to be relatively close to Afrikaans. Specifically, the Town Frisian varieties were close to Afrikaans, where the dialect of Heerenveen was the closest (4.4%). Other Town Frisian varieties (Harlingen, Staveren, Bolsward, Midsland and Dokkum), the dialect of Oost-Vlieland and the dialect of Amsterdam were also among the eight closest varieties.

The strong relationship with the Town Frisian dialects may be explained by the fact that in both Afrikaans and Town Frisian the initial consonant cluster in words like *schip* (‘ship’) and *school* (‘school’) is pronounced as [sk], while most other dialects and Standard Dutch pronounce this consonant cluster as [sx]. Another shared feature is that the initial consonant in words like *vinger* (‘finger’) and *vijf* (‘five’) is a voiceless [f] and the initial consonant in words like *zee* (‘sea’) and *zes* (‘six’) is a voiceless [s]. Most other dialects and Standard Dutch have initial [v] and [z], although currently there seems to be an increasing tendency to devoice these fricatives.

The relationship of Afrikaans with Town Frisian may be an unexpected outcome at first glance. According to Kloeke (1950), Frisian did not have any significant influence on Afrikaans, but he stresses the assumption that the [sk] pronunciation was once used in the whole Dutch dialect area. Relics are presently still found in Frisia, the islands, North-Holland, Overijssel and Gelderland, and also in Noordwijk and Katwijk. Kloeke (1950: 225-226) also suggests the possibility that, in the 17th century, there may have been large relic areas in South-Holland.

As for the unvoiced fricatives, this phenomenon is partly found in the RND transcription of the South-Hollandic dialect of Zoetermeer, but not to the same extent as in the Heerenveen transcription. A similar reasoning as for the [sk] pronunciation may also apply here.

Again, the Limburg varieties are distant to Afrikaans, especially the Ripuarian varieties in the southern-most area close to the Dutch/German state border. The dialect of Vaals is most distant (18.2%).

4.3 Closest dialect versus closest standard language

In section 4.1, we compared Afrikaans to the other West Germanic standard languages and found Standard Dutch to be closest when measuring full pronunciation distances. In section 4.2, we went into more detail by comparing Afrikaans to the dialects of Dutch. We found the South-Holland dialect of Zoetermeer closest to Afrikaans. Language technology has been extensively developed for standard languages like Standard Dutch, but usually not for dialects like that of Zoetermeer. This brings us to addressing our third research question: If dialects are found which are closer to Afrikaans than Standard Dutch, is the closest one significantly closer to Afrikaans than Standard Dutch is?

Indeed, we found that the Zoetermeer dialect is significantly closer to Afrikaans than Standard Dutch ($t=3.383$, $n=125$, $p<0.001$). Looking at the level of vowel substitutions only, we did not find Zoetermeer significantly closer to Afrikaans than Standard Dutch ($t=1.378$, $n=125$, $p=0.086$), but at the level of consonant substitutions, Zoetermeer is significantly closer than Standard Dutch ($t=6.763$, $n=125$, $p<0.001$). Therefore, we conclude that Afrikaans language technologists using the recycling approach should ideally work with spoken language resources from Zoetermeer; however, in the absence of such resources, they could use Standard Dutch carefully, since the Zoetermeer dialect is relatively close to Standard Dutch.

5. Conclusions

In this article, Afrikaans was compared to three West Germanic standard languages (Dutch, Frisian and German). Unsurprisingly, Afrikaans was found to be most closely related to Dutch. When Afrikaans was compared to 361 Dutch and Frisian dialects, the South-Hollandic varieties were found to be closest to Afrikaans. According to Kloeke (1950), the southern varieties in the province of South-Holland are the main sources of Afrikaans. However, our closest variety – the dialect of Zoetermeer – is found in the centre of the province. We did not specifically find the southern South-Hollandic varieties to be the closest. It is highly likely that the South-Hollandic dialect area has changed since 1652. The strong relationship between Afrikaans and the South-Hollandic varieties can be explained by their vowels. With regard to the consonants, the Town Frisian varieties are most closely related to Afrikaans, probably because they still maintain features which were lost in the South-Hollandic dialects.

The results of this study indicate that, for the development of automatic speech recognition systems for Afrikaans, Standard Dutch is probably the best language from which to "borrow" acoustic data, rather than, say, Flemish. The dialect of Zoetermeer is significantly closer to Afrikaans than Standard Dutch is. Therefore, acoustic data of the dialect of Zoetermeer and other strongly related South-Hollandic dialects would be even better but will probably not be available since developers of automatic speech systems focus on (accents of) standard languages rather than on dialects.

Acknowledgments

We thank Peter Kleiweg for the program which we used for the visualisation of the maps in this article. The program is part of the R*u*G/L^{04} package which is freely available at http://www.let.rug.nl/~kleiweg/L04.

Part of this research was made possible by a research grant from the South African National Research Foundation (FA207041600015) for research on HLT Resources for Closely-Related Languages.

References

Alewijnse, B., J. Nerbonne, L.J. van der Veen and F. Manni. 2007. A computational analysis of Gabon varieties. In P. Osenova, E. Hinrichs and J. Nerbonne (Eds.) *Proceedings of the RANLP Workshop on Computational Phonology.* Borovetz, Bulgaria: RANLP. pp. 3-12.

Bezooijen, R.V. and C. Gooskens. 2006. Waarom is geschreven Afrikaans makkelijker voor Nederlandstaligen dan andersom? In T. Koole, J. Nortier and B. Tahitu (Eds.) *Artikelen van de Vijfde sociolinguïstische conferentie in Lunteren.* Delft: Eburon. pp. 68-76.

Blancquaert, E. 1939. *Tekstboekje.* Antwerpen: De Sikkel.

Blancquaert, E. and W. Pée. 1925-1982. *Reeks Nederlandse dialectatlassen.* Antwerpen: De Sikkel.

Boersma, P. and D. Weenink. 2002. *PRAAT: Doing phonetics by computer.* Amsterdam: Institute of Phonetic Sciences.

Bolognesi, R. and W. Heeringa. 2002. De invloed van dominante talen op het lexicon en de fonologie van Sardische dialecten. *Gramma/TTT: Tijdschrift Voor Taalwetenschap* 9: 45-84.

Conradie, C.J. 1986. *Taalgeskiedenis.* Pretoria: Academica.

De Kleine, C. 1997. The verb phrase in Afrikaans: Evidence of creolization? In A.K. Spears and D. Winford (Eds.) *The structure and status of pidgins and creoles: Including selected papers from the meeting of the Society for Pidgin and Creole Linguistics*. Amsterdam: John Benjamins. pp. 289-307.

De Villiers, M. 1978. *Nederlands en Afrikaans.* Goodwood: Nasou.

Den Besten, H. 1989. From Khoekhoe foreigner talk via Hottentot Dutch to Afrikaans: The creation of a novel grammar. In M. Pütz and R. Dirven (Eds.) *Wheels within wheels*. Frankfurt: Peter Lang. pp. 207-254.

Den Besten, H. 2009. Desiderata voor de historische taalkunde van het Afrikaans. In H. Den Besten, F. Hinskens and J. Koch (Eds.) *Afrikaans. Een drieluik.* Amsterdam: Stichting Neerlandistiek VU. pp. 234-252.

Ehlers, D. and P.V. Beek. 2004. *Oranje boven: Nederlands voor Zuid-Afrika.* Pretoria: Protea Boekhuis.

Gooskens, C. and R.V. Bezooijen. 2006. Mutual comprehensibility of written Afrikaans and Dutch: Symmetrical or asymmetrical? *Literary and Linguistic Computing* 21: 543-557.

Gooskens, C. and W.J. Heeringa. 2004. Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data. *Language Variation and Change* 16: 189-207.

Hajič, J., J. Hric and V. Kuboň. 2000. Machine translation of very close languages. *Proceedings of the sixth Conference on Applied Natural Language Processing*. 29 April to 4 May 2000, Seattle, Washington. pp. 7-12.

Heeringa, W.J. 2001. De selectie en digitalisatie van dialecten en woorden uit de Reeks Nederlandse Dialectatlassen. *TABU: Bulletin Voor Taalwetenschap* 31: 61-103.

Heeringa, W.J. 2004. *Measuring dialect pronunciation differences using Levenshtein distance.* Groningen: Rijksuniversiteit Groningen.

Heeringa, W.J. and J. Nerbonne. 2000. Change, convergence and divergence among Dutch and Frisian. In P. Boersma, P.H. Breuker, L.G. Jansma and J. Van der Vaart (Eds.) *Philologia Frisica anno 1999. Lêzingen fan it fyftjinde Frysk filologekongres*. Fryske Akademy: Ljouwert. pp. 88-109.

Heeringa, W.J., J. Nerbonne and P. Osenova. 2010. Detecting contact effects in pronunciation. In M. Norde, B. De Jonge and C. Hasselblatt (Eds.) *Language contact. New perspectives*. Amsterdam and Philadelphia: John Benjamins Publishing Company. pp. 131-153.

Hinskens, F., P. Auer and P. Kerswill. 2005. The study of dialect convergence and divergence: Conceptual and methodological considerations. In P. Auer, F. Hinskens and P. Kerswill (Eds.) *Dialect change: Convergence and divergence in European languages*. Cambridge: Cambridge University Press. pp. 1-48.

Kessler, B. 1995. Computational dialectology in Irish Gaelic. *Proceedings of the seventh conference of the European Chapter of the Association for Computational Linguistics.* Dublin: EACL. pp. 60-66.

Kloeke, G.C. 1950. *Herkomst en groei van het Afrikaans.* Leiden: Universitaire Pers.

Krech, H. and U. Stötzer. 1969. *Wörterbuch der deutschen Aussprache.* München: Max Hueber Verlag.

Kruskal, J.B. 1999. An overview of sequence comparison. In D. Sankoff and J.B. Kruskal (Eds.) *Time warps, string edits, and macro molecules: The theory and practice of sequence comparison*. Stanford: CSLI. pp. 1-44.

Nerbonne, J. 2015. Various variation aggregates in the LAMSAS South. In M.D. Picone and C.E. Davies (Eds.) *Language variety in the South: Historical and contemporary perspectives.* Tuscaloosa: University of Alabama Press. pp. 723-753.

Nerbonne, J., W.J. Heeringa, E.V. Den Hout, P. Van der Kooi, S. Otten and W. Van de Vis. 1996. Phonetic distance between Dutch dialects. In G. Durieux, W. Daelemans and S. Gillis (Eds.) *CLIN VI: Proceedings of the sixth CLIN meeting.* Antwerp: Centre for Dutch Language and Speech (UIA). pp. 185-202.

Nerbonne, J. and C. Siedle. 2005. Dialektklassifikation auf der Grundlage aggregierter Ausspracheunterschiede. *Zeitschrift für Dialektologie und Linguistik* 72: 129-147.

Raidt, E.H. 1991. *Afrikaans en sy Europese verlede.* Kaapstad: Nasou.

Rayner, M., D. Carter, I. Bretan, R. Eklund, M. Wirén, S.L. Hansen, S. Kirchmeier-Andersen, C. Philp, F. Sørensen and H.E. Thomsen. 1997. Recycling lingware in a multilingual MT system. In J. Burstein and C. Leacock (Eds.) *From research to commercial applications: Making NLP work in practice*. Somerset, NJ: Association for Computational Linguistics. pp. 65-70.

Rietveld, A.C.M. and V.J.V. Heuven. 1997. *Algemene fonetiek.* Bussum: Coutinho.

Scholtz, J.D.P. 1963. *Taalhistoriese opstelle.* Pretoria: J.L. van Schaik.

Van der Merwe, H.J.J.M. 1951. *An introduction to Afrikaans.* Cape Town: A.A. Balkema.

Van der Merwe, H.J.J.M. 1968. *Afrikaans: Sy aard en ontwikkeling.* Pretoria: J.L. van Schaik.

Van Huyssteen, G.B. and S. Pilon. 2009. Rule-based conversion of closely-related languages: A Dutch-to-Afrikaans convertor. In F. Nicolls (Ed.) *Proceedings of the 2009 Conference of the Pattern Recognition Association of South Africa*. PRASA: Stellenbosch, South Africa. pp. 23-28.

Van Reenen, P. and A. Coetzee. 1996. Afrikaans, the daughter of Dutch. In H.F. Nielsen and L. Schøsler (Eds.) *The origins and development of emigrant languages, Proceedings from the second Rasmus Rask colloquium*, Odense University, November 1994, Rask Supplement vol. 6, NOWELE Supplement vol. 17. pp. 71-101.

Wells, J. and J. House. 1995. *The sounds of the International Phonetic Alphabet.* London: Department of Phonetics and Linguistics, University College London.

Zwicker, E. and H. Fastl. 1990. *Psychoacoustics and models.* Berlin: Springer Verlag.