The origin of Afrikaans pronunciation: a comparison to west Germanic languages and Dutch dialects

Wilbert Heeringa, Febe de Wet

Meertens Institute, Variationist Linguistics
wilbert.heeringa@meertens.knaw.nl
Stellenbosch University, Centre for Language and Speech Technology
fdw@sun.ac.za

Abstract

This paper aims to find the origin of the Afrikaans pronunciation with the use of dialectometry. First, Afrikaans was compared to Standard Dutch, Standard Frisian and Standard German. Pronunciation distances were measured by means of Levenshtein distances. Afrikaans was found to be closest to Standard Dutch. Second, the Afrikaans pronunciation was compared to 361 Dutch dialect varieties in the Netherlands and North-Belgium. Material from the *Reeks Nederlandse Dialectatlassen* was used. Afrikaans was found to be closest to the South Holland variety of Zoetermeer, which largely agrees with Kloeke (1950, *Herkomst en Groei van het Afrikaans*).

1. Introduction

Afrikaans is a daughter language of Dutch and is mainly spoken in South Africa and Namibia. Reenen & Coetzee [1] briefly describe the origin of Afrikaans. Nearly 350 years ago, in 1652, Jan van Riebeeck founded a refreshment station at the Cape of Good Hope on the way to the Indies and introduced a Dutch variety. He and the group around him came from the southern part of the Dutch province of South-Holland. Van Reenen & Coetzee refer to Kloeke [2] who claims that Jan van Riebeeck's group is the most important source of today's Afrikaans language. Kloeke writes extensively about the origin of Afrikaans in his *Herkomst en Groei van het Afrikaans* 'Origin and growth of Afrikaans'. Van Reenen & Coetzee also refer to Scholtz [3, p. 254] who does not agree with Kloeke but wonders whether Afrikaans is derived from a common Hollandish language, the Hollandish norm of the second half of the 17th century. However, Van Reenen & Coetzee doubt whether a common Hollandish language already existed in that period.

The South African constitution recognizes 11 official languages. According to the 2001 census data, Zulu is the most widely spoken mother-tongue in South Africa, followed by Xhosa and Afrikaans, with the latter constituting 13.3% of the population. This percentage is lower than the value reported in the 1996 census, when 14.4% of the population indicated that Afrikaans was their first language [4]. This observation can probably be explained by a decline in population growth as well as the fact that many Afrikaans people emigrated during that period. Although English is most often used as the lingua franca in the country, Afrikaans is more frequently used than English in some provinces of South Africa and Namibia.

As explained above, Afrikaans is seen historically as a daughter of Dutch. This paper shows that Afrikaans is linguistically still a daughter of Dutch. In order to prove this, the Afrikaans pronunciation is compared to the pronunciation of the languages in the west Germanic language group: Standard Dutch, Standard Frisian and Standard German. Pronunciation distances are measured with Levenshtein distance, a string edit distance measure. Kessler [5] was the first to use Levenshtein distance for measuring linguistic distances. He applied Levenshtein distance to transcriptions of Irish Gaelic dialect varieties. Later Levenshtein distances was applied to Dutch dialects by Nerbonne et al. [6] (more detailed results are given by Heeringa [7], to Norwegian by Gooskens & Heeringa [8] and to several other dialect families.

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. This paper aims to answer the following question: which of these standard languages is closest to Afrikaans? Afrikaans is also compared to 361 Dutch varieties, found in the Dutch dialect area. This area comprises the Netherlands and North-Belgium. Material from the *Reeks Nederlandse Dialectatlassen* is used. We determine which dialect variety (or dialect region) is closest to Afrikaans. Again pronunciation differences are measured with Levenshtein distance. We also distinguish between vowel and consonant differences.

The aim of this study is twofold. Firstly, this investigation sheds light on the linguistic relationship between Afrikaans and the west Germanic languages, and between Afrikaans and the Dutch dialects in particular. Secondly, the results of this study will provide useful guidelines for the development of speech technology applications for Afrikaans. Human language technology (HLT) is still a relatively new field in South Africa and 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. Development can be accelerated if existing resources from closely related languages can be used. We are specifically interested in constructing a large vocabulary continuous speech recognition system for Afrikaans. This requires large quantities of annotated audio data. Given that very little Afrikaans data is currently available, we would like to investigate the possibility of using data from closely related languages.

2. Data source

2.1. Dutch dialects

In order to study the relationship between Afrikaans and Dutch dialect varieties, it would be preferable to use data from about 1652, because that time period would coincide 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 [9] in the period 1925–1982. The atlases cover the Dutch dialect area, which comprises the Netherlands, the northern part of Belgium, a smaller northwestern part of France and the German county Bentheim.

In the RND, the same 141 sentences are translated and transcribed in phonetic script for each dialect. Blancquaert 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 digitizing the phonetic texts is time-consuming and 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 [10]). The words are selected more or less randomly and may be considered as a random sample. The transcriptions of the 125 word pronunciations were digitized 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 some other dialects.

The RND contains transcriptions of 1956 Dutch varieties. Since it would be very time-consuming to digitize all transcriptions, a selection of 361 dialects has been made (see Heeringa [10]). When selecting the dialects, the goal was to get a net of evenly scattered dialect locations. A denser sampling resulted 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 extra localities were added because of personal interest. In the area in and around Bentheim extra varieties were added because of a detailed investigation in which the relationship among dialects at both sides of the border was studied. Besides the relationship to Standard Dutch and Standard German was studied (see Heeringa et al. [10]).

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

2.2. Languages

In this paper, Dutch dialects are compared to Afrikaans. The 125 words, selected from the RND sentences, were therefore translated into Afrikaans and pronounced by an old male and a young female, both native speakers of Afrikaans. Old males are known to be conservative speakers while young females are usually innovative speakers [11]. In our measurements below we always take 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. To ensure consistency with the existing RND transcriptions, the Standard Dutch transcription is based on the *Tekstboekje* of Blancquaert [12]. However, words such as *komen*, *rozen* and *open* are transcribed as [koˑmə], [roːzə] and [oˑpə]. In the *Tekstboekje* of Blancquaert these words would end on an [n], as suggested by the spelling. For more details see Heeringa [10].

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

The Standard German word transcriptions are based on *Wörterbuch der deutschen Aussprache* [13]. 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 [ʀ]. Because in German both realizations are allowed, for each pronunciation containing one or more <r>'s two variants are noted, one in which the [r] is pronounced, and another in which the [ʀ] is pronounced. More details are given by Heeringa *et al.* [14]. In the measurements below, both realizations will be taken into account.

3. Measuring pronunciation distances

Pronunciation differences are measured with Levenshtein distance. Pronunciation variation includes variation in sound components and morphology. The items to be compared should have the same meaning and they should be cognates.

3.1. 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 [15]. 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, e.g., 1. Assume 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:

hɑrt	delete h	1
ɑrt	replace ɑ by 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 is adapted so that only a vowel may match with a vowel, a consonant with a consonant, the [j] or [w] with a vowel (and vice versa), the [i] or [u] with a consonant (and vice versa), and a central vowel (in our research only the schwa) with a sonorant (and vice versa). In this way unlikely matches (e.g. a [p] with an [a]) are prevented. The longest alignment has the greatest number of matches. In our example we thus have the following alignment:

h	ɑ	r	t	
	æ	r	t	ə
1	1			1

3.2. Operations weights

The simplest versions of this method are based on a notion of phonetic distance in which phonetic overlap is binary: non-identical phones contribute to phonetic distance, identical ones do not. Thus the pair [i,ɒ] counts as different to the same degree as [i,ɪ]. The version of the Levenshtein algorithm used in this paper is based on the comparison of spectrograms of the sounds. Since a spectrogram is the visual representation of the acoustical signal, the visual differences between the spectrograms are reflections of the acoustical differences. The spectrograms were made on the basis of recordings of the sounds of the International Phonetic Alphabet as pronounced by John Wells and Jill House on the cassette *The Sounds of the International Phonetic Alphabet* from 1995 [16]. The different sounds were isolated from the recordings and monotonized at the mean pitch of each of the two speakers with the program PRAAT [17]. Next, for each sound a spectrogram was made with PRAAT using the so-called Barkfilter, a perceptually oriented model. On the basis of the Barkfilter representation, segment distances were calculated. Inserted or deleted segments are compared to silence, and silence is 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 [7, pp. 79–119]. In perception, small differences in pronunciation may play a relatively strong role in comparison to larger differences. Therefore logarithmic segment distances are used. The effect of using logarithmic distances is that small distances are weighted relatively more heavily than large distances. The weights will vary between 0 and 1. In a validation study, Heeringa [7, pp. 178–195] found that among several alternative distances obtained with the Levenshtein distance measure, using logarithmic Bark filter segment distances gives results which most closely approximates dialect distances as perceived by the speakers themselves.

3.3. Vowels and consonants

Besides calculating Levenshtein distances on the basis of all segments (full pronunciation distance) we also calculated distances on the basis of only vowel and consonant substitutions. If distances are calculated solely on the basis of vowels, initially the full phonetic strings are compared to each other using Levenshtein distance. Once the optimal alignment is found, the distances are based on the alignment slots which represent vowel substitutions. Consonant substitutions are calculated mutatis mutandis.

3.4. Processing RND data

The RND transcribers use slightly different notations. In order to minimize the effect of these differences, we normalized their data. The consistency problems and the way we solved them are extensively discussed by Heeringa [10][7]. For the same reason only a part of the diacritics found in the RND is used.

As in earlier studies, we processed diacritics for length (extra short, half long, long), syllabicity (syllabic), voice (voiced, voiceless) and nasality (nasal) (see Heeringa [7, pp. 109–111]). In this study the diacritic for rounding (rounded, partly rounded, unrounded, partly unrounded) is used. The distance between for example [a] and rounded [i] is 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 the [ɯ] and [ɤ] are not included in the phonetic transcription system of the RND, but transcribed as unrounded [u] and [o] respectively.

	Afrikaans	Dutch	Frisian	German
Afrikaans		3.2	4.1	5.1
Dutch			3.8	4.2
Frisian				4.8
German				

Table 1: Average Levenshtein distances between four standard languages

The distance between a monophthong and a diphthong is 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 is calculated as the mean of the distance between the first elements and the distance between the second elements. Details are given by Heeringa [7, p. 108].

4. Results

4.1. Afrikaans versus Dutch, Frisian and German

The Levenshtein distance enables us to compare Afrikaans to other language varieties. Since we selected 125 words, the distance between a variety and Afrikaans is equal to the average of the distances of 125 word pairs. In Table 1 the average Levenshtein distances between Standard Afrikaans, Standard Dutch, Standard Frisian and Standard German are given. The distances represent the average Levenshtein distances, regardless of the length of the alignments the distances are based on. The table shows that Afrikaans is most closely related to Standard Dutch. This confirms that Afrikaans is a daughter of Dutch, as suggested by Kloeke[2], Van Reenen[1] and others. Furthermore, we found Afrikaans closer to Standard Frisian than to Standard German.

4.2. Afrikaans versus Dutch dialects

With the use of Levenshtein pronunciation distances between Afrikaans and 361 Dutch dialect varieties are calculated. The results are shown in Figure 1. In the map the varieties are represented by polygons, geographic dialect islands are represented by colored 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 closest varieties are found in the province of South-Holland. Some close varieties are also found in the provinces of North-Holland and Utrecht. The dialect variety of Zoetermeer is closest to Afrikaans. Kloeke[2] claimed that the dialect of the first settlers was the main source of Afrikaans. These settlers came from southern part of the Dutch province of South-Holland, the area around Rotterdam and Schiedam. Zoetermeer is slightly north of these two locations. The Limburg variety of Raeren is furthest away.

4.2.1. Vowels

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

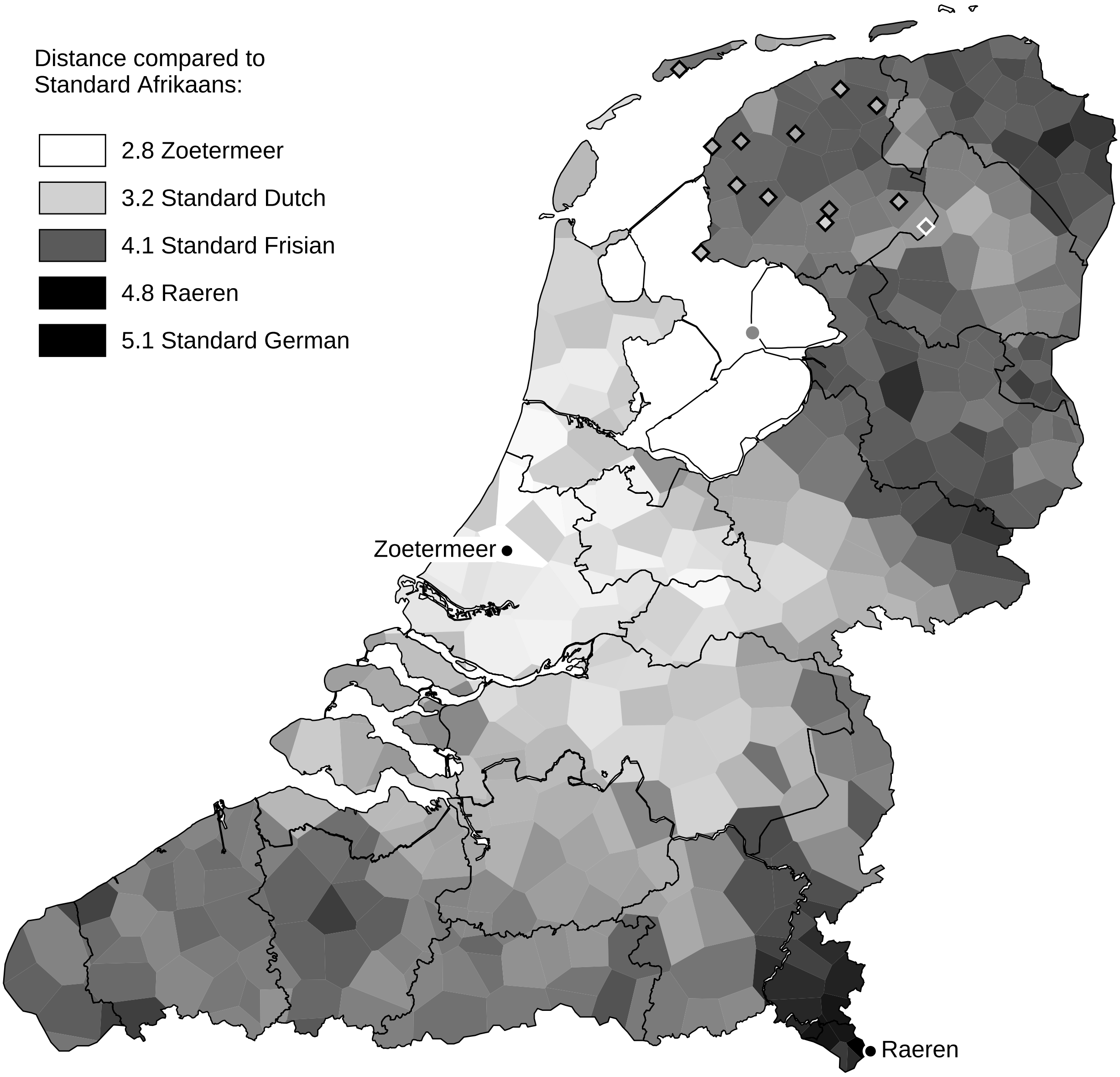

Figure 1: Distances of 361 Dutch dialect varieties compared to Afrikaans. The varieties are represented by polygons, geographic dialect islands are represented by colored dots, and linguistic dialect islands are represented by diamonds. Lighter polygons, dots or diamonds represent dialects which are closest to Afrikaans and darker ones represent the varieties which are most distant. Note that the variety of Zoetermeer is closest to Afrikaans. The IJsselmeer polders (Wieringermeerpolder, Noordoostpolder and Flevopolder) are not under consideration, so they are left white.

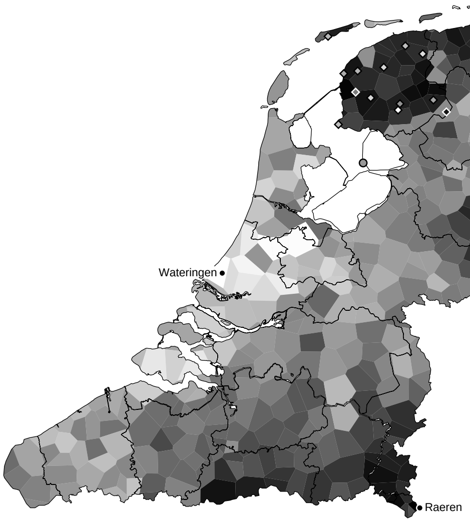

Figure 2: Vowel substitution distances of 361 Dutch dialect varieties compared to Afrikaans. Note that the variety of Wateringen is closest to Afrikaans, and the variety of Raeren is most distant.

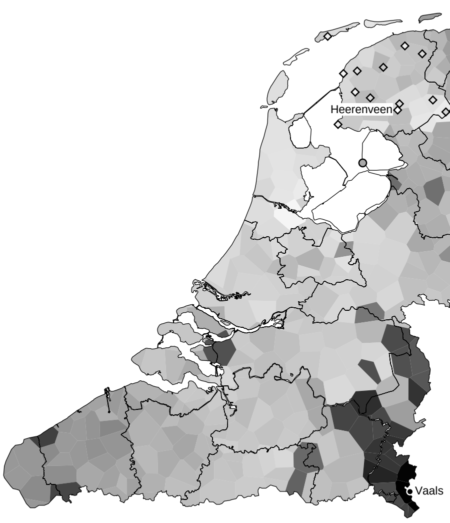

Figure 3: Consonant substitution distances of 361 Dutch dialect varieties compared to Afrikaans. Note that the variety of Heerenveen is closest to Afrikaans, and the variety of Vaals is most distant.

relatively similar to the map in Figure 1. Again the South-Hollandish varieties are close and the southern Limburg varieties are distant. The dialect of Wateringen is closest, and the dialect of Raeren is the most distant. The Frisian varieties and the core Low Saxon varieties found in Groningen and Twente are more distant than in Figure 1. The varieties close to the Dutch/French border in the Belgian province of Brabant are also relatively distant.

Our findings agree with Kloeke [2]. In the summary of his book (p. 262–263) he 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 vocal 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 *are* disappearing, the relationship with the South-Holland varieties is still found when we use the RND data.

4.2.2. Consonants

When consonant substitution distances between the Dutch dialects and Afrikaans are calculated, a completely different picture is obtained, as can be seen in Figure 3. Closest is the town Frisian variety of Heerenveen. Other Town Frisian varieties (Harlingen, Staveren, Bolsward, Midsland and Dokkum), the dialect of Oost-Vlieland and the dialect of Amsterdam are also found among the eight closest varieties. The map shows that the Limburg varieties are again distant.

The strong relationship with the Town Frisian dialects may be explained by the fact that both in Afrikaans and in 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 have [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] respectively, although there may be a current tendency to increasingly unvoice these fricatives.

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

As to the unvoiced fricatives, this phenomenon is partly found in the RND transcription of the South-Hollandish variety 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.

5. Conclusions

In this paper, Afrikaans was compared to the west Germanic standard languages (Dutch, Frisian and German). Afrikaans was found to be most related to Dutch. Van Reenen and Coetzee[1] rightly refer to Afrikaans as a daughter of Dutch. When Afrikaans is compared to 361 Dutch dialects, the South-Hollandish varieties were found to be closest to Afrikaans. According to Kloeke[2] the southern varieties in the province of South Holland are the main source of Afrikaans. However, our closest variety – the dialect of Zoetermeer – is found in the center of the province. We did not specifically find the southern South-Hollandish varieties to be closest. It is likely that the South-Hollandish dialect area has changed since 1652. The strong relationship between Afrikaans and the South-

Hollandish varieties can be explained by their vowels. As regards the consonants, the Town Frisian varieties are most closely related to Afrikaans, probably since they still maintain features which were lost in the South-Hollandish dialects. The southern Limburg varieties are most distant to Afrikaans, both when looking at vowel differences and when considering consonant differences.

The results of this study indicate that, for the development of automatic speech recognition systems for Afrikaans, Standard Dutch is probably the best language to "borrow" acoustic data from. The use of acoustic data of the South-Hollandish 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.

6. Acknowledgments

We thank Peter Kleiweg for the program which we used for the visualization of the maps in this paper. The program is part of the R*u*G/L^{04} package which is freely available at `http:www.let.rug.nl/~kleiweg/L04`. This research is partially supported by an NRF Focus Area grant for research on *HLT Resources for Closely-Related Languages*.

7. References

[1] P. Reenen and A. Coetzee, "Afrikaans, a daughter of Dutch," in *The Origins and Development of Emigrant Languages; Proceedings from the Second Rasmus Colloquium, Odense University, November 1994*, H. F. Nielsen and Lene Schøsler, Eds. 1996, pp. 71–101, Odense University Press.

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

[3] J. du P. Scholtz, *Taalhistoriese Opstelle*, J.L. van Schaaik, Pretoria, 1963.

[4] Statistics South Africa, "Census 2001: Key results," Tech. Rep., Statistics South Africa, Pretoria, 2001, Available as: http://www.statssa.gov.za/PublicationsHTML/Report-03-02-012001/html/Report-03-02-012001.html.

[5] B. Kessler, "Computational dialectology in Irish Gaelic," in *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics*, Dublin, 1995, pp. 60–67, EACL.

[6] J. Nerbonne, W. Heeringa, E. Van den Hout, P. van der Kooi, S. Otten, and W. van de Vis, "Phonetic distance between Dutch dialects," in *CLIN VI, Papers from the sixth CLIN meeting*, G. Durieux, W. Daelemans, and S. Gillis, Eds., Antwerp, 1996, pp. 185–202, University of Antwerp, Center for Dutch Language and Speech (UIA).

[7] W. J. Heeringa, *Measuring Dialect Pronunciation Differences using Levenshtein Distance*, Ph.D. thesis, Rijksuniversiteit Groningen, Groningen, 2004.

[8] Ch. Gooskens and W. Heeringa, "The position of Frisian in the Germanic language area," in *On the Boundaries of Phonology and Phonetics*, D. Gilbers, M. Schreuder, and N. Knevel, Eds., pp. 61–87. Center for Linguistics and Cognition, Groningen, Groningen, 2004.

[9] E. Blancquaert and W. Peé, Eds., *Reeks Nederlands(ch)e Dialectatlassen*, De Sikkel, Antwerpen, 1925–1982.

[10] W. Heeringa, "De selectie en digitalisatie van dialecten en woorden uit de Reeks Nederlandse Dialectatlassen," *TABU: Bulletin voor taalwetenschap*, vol. 31, no. 1/2, pp. 61–103, 2001.

[11] F. Hinskens, P. Auer, and P. Kerswill, "The study of dialect convergence and divergence: conceptual and methodological considerations," in *Dialect change. The convergence and divergence of dialects in contemporary societies*, P. Auer, F. Hinskens, and P. Kerswill, Eds., pp. 1–48. Cambridge University Press, Cambridge, 2005.

[12] E. Blancquaert, *Tekstboekje*, De Sikkel, Antwerpen, 2nd edition, 1939, Nederlandse Fonoplaten van Blancquaert en van der Plaetse, Eerste Reeks.

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

[14] W. Heeringa, J. Nerbonne, H. Niebaum, R. Nieuweboer, and P. Kleiweg, "Dutch-German contact in and around Bentheim," in *Languages in Contact. Studies in Slavic and General Linguistics*, D. Gilbers, J. Nerbonne, and J. Schaeken, Eds., vol. 28, pp. 145–156. Rodopi, Amsterdam and Atlanta GA, 2000.

[15] J. B. Kruskal, "An overview of sequence comparison," in *Time Warps, String edits, and Macromolecules. The Theory and Practice of Sequence Comparison*, D. Sankoff and J. Kruskal, Eds., pp. 1–44. CSLI, Stanford, 2nd edition, 1999, 1st edition appeared in 1983.

[16] IPA, *The Sounds of the International Phonetic Alphabet*, Department of Phonetics and Linguistics, University College London, London, 1995, Available as audio cassette or CD.

[17] P. Boersma and D. Weenink, *Praat: doing phonetics by computer*, Institute of Phonetic Sciences, Amsterdam, 2002, Available at: `http://www.praat.org`.