The relative divergence of Dutch dialect pronunciations from their common source: an exploratory study

Wilbert Heeringa
Department of Humanities Computing
University of Groningen
Groningen, The Netherlands
w.j.heeringa@rug.nl

Brian Joseph
Department of Linguistics
The Ohio State University
Columbus, Ohio, USA
bjoseph@ling.ohio-state.edu

Abstract

In this paper we use the *Reeks Nederlandse Dialectatlassen* as a source for the reconstruction of a 'proto-language' of Dutch dialects. We used 360 dialects from locations in the Netherlands, the northern part of Belgium and French-Flanders. The density of dialect locations is about the same everywhere. For each dialect we reconstructed 85 words. For the reconstruction of vowels we used knowledge of Dutch history, and for the reconstruction of consonants we used well-known tendencies found in most textbooks about historical linguistics. We validated results by comparing the reconstructed forms with pronunciations according to a proto-Germanic dictionary (Köbler, 2003). For 46% of the words we reconstructed the same vowel or the closest possible vowel when the vowel to be reconstructed was not found in the dialect material. For 52% of the words all consonants we reconstructed were the same. For 42% of the words, only one consonant was differently reconstructed. We measured the divergence of Dutch dialects from their 'proto-language'. We measured pronunciation distances to the proto-language we reconstructed ourselves and correlated them with pronunciation distances we measured to proto-Germanic based on the dictionary. Pronunciation distances were measured using Levenshtein distance, a string edit distance measure. We found a relatively strong correlation (r=0.87).

1 Introduction

In Dutch dialectology the *Reeks Nederlandse Dialectatlassen* (RND), compiled by Blancquaert & Pée (1925-1982) is an invaluable data source. The atlases cover the Dutch language area. The Dutch area comprises The Netherlands, the northern part of Belgium (Flanders), a smaller northwestern part of France, and the German county of Bentheim. The RND contains 1956 varieties, which can be found in 16 volumes. For each dialect 139 sentences are translated and transcribed in phonetic script. Blancquaert mentions that the questionnaire used for this atlas was conceived of as a range of sentences with words that illustrate particular sounds. The design was such that, e.g., various changes of older Germanic vowels, diphthongs and consonants are represented in the questionnaire (Blancquaert 1948, p. 13). We exploit here the historical information in this atlas.

The goals of this paper are twofold. First we aim to reconstruct a 'proto-language' on the basis of the RND dialect material and see how close we come to the protoforms found in Gerhard Köbler's *neuhochdeutsch-germanisches Wörterbuch* (Köbler, 2003). We recognize that we actually reconstruct a stage that would never have existed in prehistory. In practice, however, we are usually forced to use incomplete data, since data collections -- such as the RND – are restricted by political boundaries, and often some varieties are lost. In this paper we show the usefulness of a data source like the RND.

Second we want to measure the divergence of Dutch dialects compared to their proto-language. We measure the divergence of the dialect pronunciations. We do not measure the number of changes that happened in the course of time. For

Proceedings of Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology, pages 31–39,
Prague, June 2007.

example if a [u] changed into a [y] and then the [y] changed into a [u], we simply compare the [u] to the proto-language pronunciation. However, we do compare Dutch dialects to both the proto-language we reconstruct ourselves, which we call ***Proto-Language Reconstructed*** (PLR), and to the Proto-language according to the proto-Germanic Dictionary, which we call ***Proto-Germanic according to the Dictionary*** (PGD).

2 Reconstructing the proto-language

From the nearly 2000 varieties in the RND we selected 360 representative dialects from locations in the Dutch language area. The density of locations is about the same everywhere.

In the RND, the same 141 sentences are translated and transcribed in phonetic script for each dialect. Since digitizing the phonetic texts is time-consuming on the one hand, and since our procedure for measuring pronunciation distances is a word-based method on the other hand, we initially selected from the text only 125 words. Each set represents a set of potential cognates, inasmuch as they were taken from translations of the same sentence in each case. In Köbler's dictionary we found translations of 85 words only; therefore our analyses are based on those 85 words.

We use the *comparative method* (CM) as the main tool for reconstructing a proto-form on the basis of the RND material. In the following subsections we discuss the reconstruction of vowels and consonants respectively.

2.1 Vowels

For the reconstruction of vowels we used knowledge about sound developments in the history of Dutch. In Old Dutch the diphthongs /ai/ and /au/ turned into monophthongs /eː/ and /oː/ respectively (Quak & van der Horst 2002, p. 32). Van Bree (1996) mentions the tendencies that lead /eː/ and /oː/ to change into /iː/ and /uː/ respectively. From these data we find the following chains:

ɑi	→	ɛi	→	e	→	i
ɑu	→	ɔu	→	o	→	u

An example is *twee* 'two' which has the vowel [ɑ] in 11% of the dialects, the [ɛ] in 14% of the dialects, the [e] in 43% of the dialects and the [i] in 20% of the dialects.[1] According to the *neu-hochdeutsch-germanisches Wörterbuch* the [ɑ] or [ɑi] is the original sound. Our data show that simply reconstructing the most frequent sound, which is the [e], would not give the original sound, but using the chain the original sound is easily found.

To get evidence that the /ɑi/ has raised to /e/ (and probably later to /i/) in a particular word, we need evidence that the /e/ was part of the chain. Below we discuss another chain where the /i/ has lowered to /ai/, and where the /e/ is missing in the chain. To be sure that the /e/ *was* part of the chain, we consider the frequency of the /e/, i.e. the number of dialects with /e/ in that particular word. The frequency of /e/ should be higher than the frequency of /ɛ/ and/or higher than the frequency of /i/. Similarly for the change from /ɑu/ to /o/ we consider the frequency of /o/.

Another development mentioned by Van Bree is that high monophthongs diphthongize. In the transition from middle Dutch to modern Dutch, the monophthong /iː/ changed into /ɛi/, and the monophthong /yː/ changed into either /œy/ or /ɔi/ (Van der Wal, 1994). According to Van Bree (1996, p. 99), diphthongs have the tendency to lower. This can be observed in Polder Dutch where /ɛi/ and /œy/ are lowered to /ai/ and /au/ (Stroop 1998). We recognize the following chains:

i	→	ɛi	→	ɑi
y	→	œy/ɔi	→	ɑu
u	→	ɔu	→	ɑu

Different from the chains mentioned above, we do not find the /e/ and /o/ respectively in these chains. To get evidence for these chains, the frequency of /e/ should be lower than both the frequency of /ɛ/ and /i/, and the frequency of /o/ should be lower than both /ɔ/ and /u/.

Sweet (1888, p. 20) observes that vowels have the tendency to move from back to front. Back

[1] The sounds mentioned may be either monophthongs or diphthongs.

vowels favour rounding, and front vowels unrounding. From this, we derive five chains:

i	←	y	←	u
ɪ	←	ʏ	←	ʊ
e	←	ø	←	o
ɛ	←	œ	←	ɔ
a	←	←	←	ɑ

However, unrounded front vowels might become rounded under influence from a labial or labiodental consonant. For example *vijf* ‘five’ is sometimes pronounced as [vif] and sometimes as [vyf]. The [i] has been changed into [y] under influence of the labiodental [v] and [f].

Sweet (1888, p. 22) writes that the dropping of unstressed vowels is generally preceded by various weakenings in the direction of a vowel close to schwa. In our data we found that the word *mijn* ‘my’ is sometimes [i] and sometimes [ə]. A non-central unstressed vowel might change into a central vowel which in turn might be dropped. In general we assume that deletion of vowels is more likely than insertion of vowels.

Most words in our data have one syllable. For each word we made an inventory of the vowels used across the 360 varieties. We might recognize a chain in the data on the basis of vowels which appear at least two times in the data. For 37 words we could apply the tendencies mentioned above. In the other cases, we reconstruct the vowel by using the vowel found most frequently among the 360 varieties, working with Occam’s Razor as a guiding principle. When both monophthongs and diphthongs are found among the data, we choose the most frequent monophthong. Sweet (1888, p. 21) writes that isolative diphthongizaton “mainly affects long vowels, evidently because of the difficulty of prolonging the same position without change.”

2.2 Consonants

For the reconstruction of consonants we used ten tendencies which we discuss one by one below.

Initial and medial voiceless obstruents become voiced when (preceded and) followed by a voiced sound. Hock & Joseph (1996) write that weakening (or lenition) “occurs most commonly in a medial voiced environment (just like Verner’s law), but may be found in other contexts as well.” In our data set *zes* ‘six’ is pronounced with a initial [z] in most cases and with an initial [s] in the dialects of Stiens and Dokkum. We reconstructed [s].[2]

Final voiced obstruents of an utterance become voiceless. Sweet (1888, p. 18) writes that the natural isolative tendency is to change voice into unvoiced. He also writes that the “tendency to unvoicing is shown most strongly in the stops.” Hock & Joseph (1996, p. 129) write that final devoicing “is not confined to utterance-final position but applies word-finally as well.”[3] In our data set we found that for example the word-final consonant in *op* ‘on’ is sometimes a [p] and sometimes a [b]. Based on this tendency, we reconstruct the [b].

Plosives become fricatives between vowels, before vowels or sonorants (when initial), or after vowels (when final). Sweet writes that the “opening of stops generally seems to begin between vowels…” (p. 23). Somewhat further he writes that in Dutch the *g* has everywhere become a fricative while in German the initial *g* remained a stop. For example *goed* ‘good’ is pronounced as [guˑᵊt] in Frisian dialects, while other dialects have initial [ɣ] or [x]. Following the tendency, we consider the [g] to be the older sound. Related to this is the pronunciation of words like *schip* ‘ship’ and *school* ‘school’. As initial consonants we found [sk], [sx] and [ʃ]. In cases like this we consider the [sk] as the original form, although the [k] is not found *between* vowels, but only *before* a vowel.

Oral vowels become nasalized before nasals. Sweet (1888) writes that “nothing is more common than the nasalizing influence of a nasal on a preceding vowels” and that there “is a tendency to drop the following nasal consonant as superfluous” when “the nasality of a vowel is clearly developed” and “the nasal consonant is final, or stands before another consonant.” (p. 38) For example *gaan* ‘to go’ is pronounced as [gɑ̹ːn] in the dialect of Dok-

[2] In essence, in this and other such cases, a version of the manuscript-editing principle of choosing the “lectio difficilior” was our guiding principle.

[3] We do feel, however, that word-final devoicing, even though common cross-linguistically, is, as Hock 1976 emphasizes, not phonetically determined but rather reflects the generalization of utterance-final developments into word-final position, owing to the overlap between utterance-finality and word-finality.

kum, and as [gẽ·ə̃] in the dialect of Stiens. The nasalized [ẽ·ə] in the pronunciation of Stiens already indicates the deletion of a following nasal.

Consonants become palatalized before front vowels. According to Campbell (2004) "palatalization often takes place before or after *i* and *j* or before other front vowels, depending on the language, although unconditioned palatalization can also take place." An example might be *vuur* which is pronounced like [fju·ər] in Frisian varieties, while most other varieties have initial [f] or [v] followed by [i] or [y].

Superfluous sounds are dropped. Sweet (1888) introduced this principle as one of the principles of economy (p. 49). He especially mentioned that in [ŋg] the superfluous [g] is often dropped (p. 42). In our data we found that *krom* 'curved' is pronounced [krʊ·m] in most cases, but as [krʊ̃·mp] in the dialect of Houthalen. In the reconstructed form we posit the final [p].

Medial [h] deletes between vowels, and initial [h] before vowels. The word *hart* 'heart' is sometimes pronounced with and sometimes without initial [h]. According to this principle we reconstruct the [h].

[r] changes to [ʀ]. According to Hock and Joseph (1996) the substitution of uvular [ʀ] for trilled (post-)dental [r] is an example of an occasional change apparently resulting from misperception. In the word *rijp* 'ripe' we find initial [r] in most cases and [ʀ] in the dialects of Echt and Baelen. We reconstructed [r].

S*yllable initial [w] changes in [ʋ]*. Under 'Lip to Lip-teeth' Sweet (1888) writes that in "the change of *p* into *f*, *w* into *v*, we may always assume an intermediate [ɸ], [β], the latter being the Middle German *w*" (p. 26), and that the "loss of back modification is shown in the frequent change of (w) into (v) through [β], as in Gm." Since *v* – meant as "voiced lip-to-teeth fricative" – is close to [ʋ] – lip-to-teeth sonorant – we reconstruct [w] if both [w] and [ʋ] are found in the dialect pronunciations. This happens for example in the word *wijn* 'wine'.

The cluster ol+d/t diphthongizes to ou + d/t. For example English *old* and German *alt* have a /l/ before the /d/ and /t/ respectively. In Old Dutch *ol* changed into *ou* (Van Loey 1967, p. 43, Van Bree 1987, p. 135/136). Therefore we reconstruct the /l/ with preceding /o/ or /ɑ/.

3 The proto-language according to the dictionary

The dictionary of Köbler (2003) provides Germanic proto-forms. In our Dutch dialect data set we have transcriptions of 125 words per dialect. We found 85 words in the dictionary. Other words were missing, especially plural nouns, and verb forms other than infinitives are not included in this dictionary.

For most words, many proto-Germanic forms are given. We used the forms in italics only since these are the main forms according to the author. If different lexical forms are given for the same word, we selected only variants of those lexical forms which appear in standard Dutch or in one of the Dutch dialects.

The proto-forms are given in a semi-phonetic script. We converted them to phonetic script in order to make them as comparable as possible to the existing Dutch dialect transcriptions. This necessitated some interpretation. We made the following interpretation for monophthongs:

spelling	phonetic	spelling	phonetic	spelling	phonetic
i	ɪ	***æ***	æ	***u***	u
ī	iː	***ǣ***	æː	***ū***	uː
e	ɛ	***a***	ɑ	***o***	ɔ
ē	eː	***ā***	aː	***ō***	oː

Diphthongs are interpreted as follows:

spelling	phonetic	spelling	phonetic
ai	ɑ·ⁱ	***ei***	ɛ·ⁱ
au	ɑ·ᵘ	***eu***	ɛ·ᵘ

We interpreted the consonants according to the following scheme:

spel-ling	pho-netic	spel-ling	pho-netic	spel-ling	pho-netic
p	p	***f***	f	***m***	m
b	b, v	*þ*	t	***n***	n, ŋ
t	t	***s***	s	***ng***	ŋ
d	d	***z***	z	***w***	w
k	k	***h***	x, h	***r***	r
g	g, ɣ			***l***	l
				j	j

Lehmann (2005-2007) writes that in the early stage of Proto-Germanic "each of the obstruents had the same pronunciation in its various locations…". "Later, /b d g/ had fricative allophones when medial between vowels. Lehmann (1994) writes that in Gothic "/b, d, g/ has stop articulation initially, finally and when doubled, fricative articulation between vowels." We adopted this scheme, but were restricted by the RND consonant set. The fricative articulation of /b/ would be [β] or [v]. We selected the [v] since this sound is included in the RND set. The fricative articulation of /d/ would be [ð], but this consonant is not in the RND set. We therefore used the [d] which we judge perceptually to be closer to the [ð] than to the [z]. The fricative articulation of /g/ is /ɣ/ which was available in the RND set.

We interpreted the *h* as [h] in initial position, and as [x] in medial and final positions. An *n* before *k*, *g* or *h* is interpreted as [ŋ], and as [n] in all other cases. The *þ* should actually be interpreted as [θ], but this sound in not found in the RND set. Just as we use [d] for [ð], analogously we use [t] for [θ]. We interpret double consonants are geminates, and transcribe them as single long consonants. For example *nn* becomes [nː].

Several words end in a '-' in Köbler's dictionary, meaning that the final sounds are unknown or irrelevant to root and stem reconstructions. In our transcriptions, we simply note nothing.

4 Measuring divergence of Dutch dialect pronunciations with respect to their proto-language

Once a protolanguage is reconstructed, we are able to measure the divergence of the pronunciations of descendant varieties with respect to that protolanguage. For this purpose we use Levenshtein distance, which is explained in Section 4.1. In Sections 4.2 the Dutch dialects are compared to PLR and PGD respectively. In Section 4.3 we compare PLR with PGD.

4.1 Levenshtein distance

In 1995 Kessler introduced the Levenshtein distance as a tool for measuring linguistic distances between language varieties. The Levenshtein distance is a string edit distance measure, and Kessler applied this algorithm to the comparison of Irish dialects. Later the same technique was successfully applied to Dutch (Nerbonne et al. 1996; Heeringa 2004: 213–278). Below, we give a brief explanation of the methodology. For a more extensive explanation see Heeringa (2004: 121–135).

4.1.1 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. We determine how one pronunciation might be transformed into the other by inserting, deleting or substituting sounds. 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 Dutch word *hart* 'heart' is pronounced as [hɑrt] in the dialect of Vianen (The Netherlands) and as [ærtə] in the dialect of Nazareth (Belgium). Changing one pronunciation into the other can be done as follows:

hɑrt	delete h	1
ɑrt	subst. ɑ/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 (or opposite), the [i] or [u] with a consonant (or opposite), and a central vowel (in our research only the schwa) with a sonorant (or opposite). In this way unlikely matches (e.g. a [p] with an [a]) are prevented.[4] 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

4.1.2 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 which we use 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.[5] The different sounds were isolated from the recordings and monotonized at the mean pitch of each of the two speakers with the program PRAAT[6] (Boersma & Weenink, 2005). 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. We found that the [ʔ] is closest to silence and the [a] is most distant. This approach is described extensively in Heeringa (2004, pp. 79-119).

In perception, small differences in pronunciation may play a relatively strong role in comparison to larger differences. Therefore we used logarithmic segment distances. The effect of using logarithmic distances is that small distances are weighted relatively more heavily than large distances.

4.1.3 Processing RND data

The RND transcribers use slightly different notations. In order to minimize the effect of these differences, we normalized the data for them. The consistency problems and the way we solved them are extensively discussed in Heeringa (2001) and Heeringa (2004). Here we mention one problem which is highly relevant in the context of this paper. In the RND the *ee* before *r* is transcribed as [eː] by some transcribers and as [ɪˑ] by other transcribers, although they mean the same pronunciation as appears from the introductions of the different atlas volumes. A similar problem is found for *oo* before *r* which is transcribed either as [oː] or [ʊˑ], and the *eu* before *r* which is transcribed as [øː] or [ʏˑ]. Since similar problems may occur in other contexts as well, the best solution to overcome all of these problems appeared to replace all [ɪ]'s by [e]'s, all [ʊ]'s by [o]'s, and all [ʏ]'s by [ø]'s, even though meaningful distinctions get lost.

Especially suprasegmentals and diacritics might be used diffferently by the transcribers. We process the diacritics *voiceless*, *voiced* and *nasal* only. For details see Heeringa (2004, p. 110-111).

The distance between a monophthong and a diphthong is calculated as the mean of the distance between the monophthong and the first element of

[4] Rather than matching a vowel with a consonant, the algorithm will consider one of them as an insertion and another as a deletion.

[5] See http://www.phon.ucl.ac.uk/home/wells/cassette.htm.

[6] The program PRAAT is a free public-domain program developed by Paul Boersma and David Weenink at the Institute of Pronunciation Sciences of the University of Amsterdam and is available at http://www.fon.hum.uva.nl/praat.

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 in Heeringa (2004, p. 108).

4.2 Measuring divergence from the proto-languages

The Levenshtein distance enables us to compare each of the 360 Dutch dialects to PLR and PGD. Since we reconstructed 85 words, the distance between a dialect and a proto-language is equal to the average of the distances of 85 word pairs.

Figures 1 and 2 show the distances to PLR and PGD respectively. Dialects with a small distance are represented by a lighter color and those with a large distance by a darker color. In the map, dialects are represented by polygons, geographic dialect islands are represented by colored dots, and linguistic dialect islands are represented by diamonds. The darker a polygon, dot or diamond, the greater the distance to the proto-language.

The two maps show similar patterns. The dialects in the Northwest (Friesland), the West (Noord-Holland, Zuid-Holland, Utrecht) and in the middle (Noord-Brabant) are relatively close to the proto-languages. More distant are dialects in the Northeast (Groningen, Drenthe, Overijssel), in the Southeast (Limburg), close to the middle part of the Flemish/Walloon border (Brabant) and in the southwest close to the Belgian/French state border (West-Vlaanderen).

According to Weijnen (1966), the Frisian, Limburg and West-Flemish dialects are conservative. Our maps shows that Frisian is relatively close to proto-Germanic, but Limburg and West-Flemish are relatively distant. We therefore created two maps, one which shows distances to PGD based on vowel substitutions in stressed syllables only, and another showing distances to PGD on the basis of consonant substitutions only.[7]

Looking at the map based on vowel substitutions we find the vowels of the Dutch province of Limburg and the eastern part of the province Noord-Brabant relatively close to PGD. Looking at the map based on consonant substitutions we find the consonants of the Limburg varieties distant to PGD. The Limburg dialects have shared in the High German Consonant Shift. Both the Belgium and Dutch Limburg dialects are found east of the Uerdinger Line between Dutch *ik*/*ook*/*-lijk* and High German *ich*/*auch*/*-lich*. The Dutch Limburg dialects are found east of the Panninger Line between Dutch *sl*/*sm*/*sn*/*sp*/*st*/*zw* and High German *schl*/*schm*/*schn*/*schp*/*scht*/*schw* (Weijnen 1966). The Limburg dialects are also characterized by the uvular [ʀ] while most Dutch dialects have the alveolar [r]. All of this shows that Limburg consonants are innovative.

The map based on vowel substitutions shows that Frisian vowels are not particularly close to PGD. Frisian is influenced by the Ingvaeonic sound shift. Among other changes, the [ʏ] changed into [ɪ], which in turn changed into [ɛ] in some cases (Dutch *dun* 'thin' is Frisian *tin*) (Van Bree 1987, p. 69).[8] Besides, Frisian is characterized by its falling diphthongs, which are an innovation as well. When we consulted the map based on consonant substitutions, we found the Frisian consonants close to PGD. For example the initial /g/ is still pronounced as a plosive as in most other Germanic varieties, but in Dutch dialects – and in standard Dutch – as a fricative.

When we consider West-Flemish, we find the vowels closer to PGD than the consonants, but they are still relatively distant to PGD.

4.3 PLR versus PGD

When correlating the 360 dialect distances to PLR with the 360 dialect distances to PGD, we obtained a correlation of r=0.87 (p<0.0001)[9]. This is a significant, but not perfect correlation. Therefore we compared the word transcriptions of PLR with those of PGD.

[7] The maps are not included in this paper.

[8] The Ingvaeonic sound shift affected mainly Frisian and English, and to a lesser degree Dutch. We mention here the phenomenon found in our data most frequently.

[9] For finding the *p*-values we used with thanks: *VassarStats: Website for Statistical Computation* at: http://faculty.vassar.edu/lowry/VassarStats.html.

Figure 1. Distances of 360 Dutch dialects compared to PLR. Dialects 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 more conservative dialects and darker ones more innovative dialects.

Figure 2. Distances of 360 Dutch dialects compared to PGD. Dialects 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 more conservative dialects and darker ones more innovative dialects.

First we focus on the reconstruction of vowels. We find 28 words for which the reconstructed vowel of the stressed syllable was the same as in PGD[10]. In 15 cases, this was the result of applying the tendencies discussed in Section 2.1. In 13 cases this was the result of simply choosing the vowel found most frequently among the 360 word pronunciations. When we do not use tendencies, but simply always choose the most frequent vowel, we obtain a correlation which is significantly lower (r=0.74, p=0).

We found 29 words for which vowel was reconstructed different from the one in PGD, although the PGD vowel was found among at least two dialects. For 28 words the vowel in the PGD form was not found among the 360 dialects, or only one time. For 11 of these words, the closest vowel found in the inventory of that word, was reconstructed. For example the vowel in *ook* 'too' is [ɑu] in PGD, while we reconstructed [ɔu].

Looking at the consonants, we found 44 words which have the same consonants as in PGD.[11] For 36 words only one consonant was different, where most words have at least two consonants. This shows that the reconstruction of consonants works much better than the reconstruction of vowels.

5 Conclusions

In this paper we tried to reconstruct a 'proto-language' on the basis of the RND dialect material and see how close we come to the protoforms found in Köbler's proto-Germanic dictionary. We reconstructed the same vowel as in PGD or the closest possible vowel for 46% of the words. Therefore, the reconstruction of vowels still needs to be improved further.

The reconstructions of consonants worked well. For 52% of the words all consonants reconstructed are the same as in PGD. For 42% of the words, only one consonant was differently reconstructed.

And, as a second goal, we measured the divergence of Dutch dialects compared to their proto-

[10] For some words PGD gives multiple pronunciations. We count the number of words which has the same vowel in at least one of the PGD pronunciations.

[11] When PGD has multiple pronunciations, we count the number of words for which the consonants are the same as in at least one of the PGD pronunciations.

language. We calculated dialect distances to PLR and PGD, and found a correlation of r=0.87 between the PLR distances and PGD distances. The high correlation shows the relative influence of wrongly reconstructed sounds.

When we compared dialects to PLR and PGD, we found especially Frisian close to proto-Germanic. When we distinguished between vowels and consonants, it appeared that southeastern dialects (Dutch Limburg and the eastern part of Noord-Brabant) have vowels close to proto-Germanic. Frisian is relatively close to proto-Germanic because of its consonants.

Acknowledgements

We thank Peter Kleiweg for letting us use the programs which he developed for the representation of the maps. We would like to thank Prof. Gerhard Köbler for the use of his *neuhochdeutsch-germanisches Wörterbuch* and his explanation about this dictionary and Gary Taylor for his explanation about proto-Germanic pronunciation. We also thank the members of the Groningen Dialectometry group for useful comments on a earlier version of this paper. We are grateful to the anonymous reviewers for their valuable suggestions. This research was carried out within the framework of a *talentgrant* project, which is supported by a fellowship (number `S 30-624`) from the Netherlands Organisation of Scientific Research (NWO).

References

Edgar Blancquaert & Willem Pée, eds. 1925-1982. *Reeks Nederlandse Dialectatlassen*. De Sikkel, Antwerpen.

Paul Boersma & David Weenink 2005. *Praat: doing phonetics bycomputer*. Computer program retrieved from http://www.praat.org.

Cor van Bree. 1987. *Historische Grammatica van het Nederlands*. Foris Publications, Dordrecht.

Cor van Bree. 1996. *Historische Taalkunde*. Acco, Leuven.

Wilbert Heeringa. 2001. De selectie en digitalisatie van dialecten en woorden uit de Reeks Nederlandse Dialectatlassen. *TABU: Bulletin voor taalwetenschap*, 31(1/2):61-103.

Wilbert Heeringa. 2004. *Measuring Dialect Pronunciation Differences using Levenshtein Distance*. PhD thesis, Rijksuniversiteit Groningen, Groningen. Available at: http://www.let.rug.nl/~heeringa/dialectology/thesis.

Hans Henrich Hock & Brian D. Joseph. 1996. *Language History, Language Change, and Language Relationship: an Introduction to Historical and Comparative Linguistics*. Mouton de Gruyter, Berlin etc.

Brett Kessler. 1995. Computational dialectology in Irish Gaelic. *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics*, 60-67. EACL, Dublin.

Gerhard Köbler. 2003. *Neuhochdeutsch-germanisches Wörterbuch*. Available at: http://www.koeblergerhard.de/germwbhinw.html.

Winfred P. Lehmann. 1994. Ghotic and the Reconstruction of Proto-Germanic. In: Ekkehard König & Johan van der Auwera, eds. *The Germanic Languages*, 19-37. Routledge, London & New York.

Winfred P. Lehmann. 2005-2007. *A Grammar of Proto-Germanic.* Online books edited by Jonathan Slocum. Available at: http://www.utexas.edu/cola/centers/lrc/books/pgmc00.html.

Adolphe C. H. van Loey. 1967. *Inleiding tot de historische klankleer van het Nederlands*. N.V. W.J. Thieme & Cie, Zutphen.

John Nerbonne & Wilbert Heeringa & Erik van den Hout & Peter van der Kooi & Simone Otten & Willem van de Vis. 1996. Phonetic Distance between Dutch Dialects. In: Gert Durieux & Walter Daelemans & Steven Gillis, eds. *CLIN VI, Papers from the sixth CLIN meeting*, 185-202. University of Antwerp, Center for Dutch Language and Speech, Antwerpen.

Arend Quak & Johannes Martinus van der Horst. 2002. *Inleiding Oudnederlands*. Leuven University Press, Leuven.

Jan Stroop. 1998. *Poldernederlands; Waardoor het ABN verdwijnt*, Bakker, Amsterdam.

Henry Sweet. 1888. *A History of English Sounds from the Earliest Period*. Clarendon Press, Oxford.

Marijke van der Wal together with Cor van Bree. 1994. *Geschiedenis van het Nederlands*. Aula-boeken. Het Spectrum, Utrecht, 2nd edition.

Antonius A. Weijnen. 1966. *Nederlandse dialectkunde.* Studia Theodisca. Van Gorcum, Assen, 2nd edition.