Attraction between words as a function of frequency and representational distance: words in the bilingual brain
Versloot, A.P.; Hoekstra, E.

Published in:
Linguistics

DOI:
10.1515/ling-2016-0028

Citation for published version (APA):
Arjen P. Versloot and Eric Hoekstra

Attraction between words as a function of frequency and representational distance: words in the bilingual brain

Abstract

Bilingual speakers store cognates from related languages close together in their mental lexicon. In the case of minority languages, words from the dominant language often exert influence on their cognates in the minority language. In this article, we present a model describing that influence as a function of frequency and of (dis)similarity (representational distance). More specifically, it is claimed that the strength of the influence of one word upon another is (among others) a function of their frequencies divided by their formal dissimilarity. The model is applied to the distribution of nouns derived from adjectives in Frisian, where the suffix -ens competes with -heid. Of these two suffixes, Frisian -heid is similar to Dutch -heid, whereas Frisian -ens does not have a similar counterpart in Dutch. The model predicts that Frisian derived nouns of which the adjectival bases are similar in form and meaning to Dutch will occur more often with -heid and less often with -ens. It also predicts that this effect will be stronger as the words involved are more frequent. Our findings make it possible to verify the model's quantification of the influence of Dutch words on their cognates.

Keywords: bilingualism, representational convergence, language modelling, word storage, analogy, blocking

1 Introduction to the model

1.1 Psycholinguistics and bilingualism

Psycholinguistic experiments indicate that bilingual speakers store close together words from the two languages which they speak if those words are similar in form or meaning (Dijkstra 2003; Dijkstra 2008; Smits et al. 2006; Smits et al. 2009). Thus a bilingual English-Dutch speaker stores carpenter close to its Dutch equivalent timmerman in the mental lexicon, because of the semantic similarity, and the English word wait (Dutch ‘wacht’) close to the Dutch word weet (English
‘know’) because of the similarity in form, also referred to as formal similarity. Independent research in language acquisition points to the same conclusion: cross-linguistic influence of one word form upon a similar one has been shown to occur in the language acquisition of bilinguals (Hulk and Mueller 2000), in the formation of creole languages (King 2000) and in language contact theory (Van Coetsem 1988; Thomason and Kaufman 1988).

Words which are similar both in form and in meaning are stored even closer together, and this situation regularly obtains in case a dominant language is genetically close to a minority language. It is well-known that speakers of dialects and minority languages are subject to heavy interference from their second language, i.e. the dominant language, which is prominent in the media, in education, and other domains of speech (for example Barbiers et al. 2005; Barbiers et al. 2008 with primary data for Frisian and Dutch dialects; for Frisian-Dutch language contact, see: Sjölin 1976; Gorter and Jonkman 1995; de Haan 1997; Breuker 2001a,b). This regularly causes the dialect or minority language to change in the direction of the dominant language, leading to language erosion or even language death.

West Frisian (henceforth Frisian) is an example of such a minority language. It is spoken in Fryslân, a province located in the north of the Netherlands, by ca. 450,000 speakers, and about two-thirds of them are mother tongue speakers. All speakers of Frisian are at least bilingual. Frisian is genetically very close to the dominant and official language, Dutch, by which it is heavily influenced (see the references above), as will also become clear from the analysis of Frisian corpus data to be discussed below.

In this article, we want to explore the quantitative nature of the impact that the two languages exert upon each other at the word level in the bilingual brain. We are particularly interested in the factors frequency and similarity as a psycholinguistic expression of the underlying neural substratum of language processing. These factors are reflected not only in the outcome of psycholinguistic experiments (cf. the literature referred to at the beginning of this section), but also, as we will see, in the composition of corpora.

1.2 From analogy to similarity to representational distance

The over-arching term for the association of information in the brain is *analogy* (e.g. Hofstadter and Sander 2013), which we consider not fundamentally different between languages than within one language. Analogies between words in two languages can offer a clear view of the on-going process, as the participating components may be more easily detectable by their belonging to one
or the other language. In order to formalise and quantify how analogy driven association takes effect in the bilingual brain, we have to identify factors that contribute to and can be operationalized as variables in a model of analogy in the bilingual brain.

Several types of analogy may be distinguished, such as analogy deriving from the form of a word, analogy deriving from the meaning of a word, and so on. The subject of this study is the factor form. Our data set relies on full synonyms in the two languages Frisian and Dutch. As a consequence, meaning is not a variable in the analysis. We will refer to analogy deriving from the form of two words as (their degree of) similarity. Form similarity between Dutch and Frisian may be used to explain specific on-going changes in Frisian. It has been shown that the level of similarity between Dutch and Frisian words influences their morphological behaviour with respect to compounding on a deeper level than simple morpheme-to-morpheme translations (Slofstra, Hoekstra and Versloot 2009: 39). In the same vein, it has been shown that Frisian words that are (nearly) identical in form with their Dutch counterparts, are less inclined to show Frisian-specific patterns of plural formation (Versloot forthcoming).

Similarity of words may provide us with information about the distance between (parts of) the neural representations corresponding to words. If two words are similar, then the representational distance between them will be small, that is, they will converge on many points of their representations. The similarity between the phonological shape of words from different languages is a reliable measure of the representational distance between them (e.g. van Heuven, Coderre, Guo and Dijkstra 2011: 11). Section 3 explains how representational distance has been operationalized to make it measurable by means of Levenshtein Distances with Pointwise Mutual Information (on Levenshtein Distance, see Heeringa 2004). The representational distance between the neural representations of two words is thus calculated by proxy by means of Levenshtein Distance.

1.3 From frequency to representational strength

Frequency is a reoccurring aspect in language processing and language change in general. A word’s frequency is a reliable measure of the strength of its neural representation (e.g. Bybee 1995: 452). Frequent items are accessed and processed easier and more reliably (e.g. Diessel 2007). In the competition between linguistic variants of any sort, frequency of occurrence adds to competitive strength (e.g. Krott, Baayen and Schreuder 2001). It has, for example, been shown that words from Dutch dialects with a low frequency are more prone to levelling by Standard Dutch than words of high frequency (Wieling, Nerbonne and Baayen 2011). A node for a Dutch
word will have more influence (or exert more attraction) on its Frisian equivalent as it is more frequently accessed. Section 3 explains how frequency has been operationalized. The strength of a representation in a neural network can thus be measured indirectly (by proxy) by determining the frequency of the corresponding word in a linguistic corpus.

1.4 How frequency and similarity relate to each other

Frequency and similarity are factors contributing to the force of the attraction which, on a neural level, a (word from a) dominant language exerts on a (word from a) minority language. The question arises whether a formula for the force of attraction can be found in which the contribution of frequency and similarity is made precise. The measurement of frequency and similarity will be presented in detail in Section 3 and Section 4. It will be argued that the force of attraction (A) which a Dutch word exerts on its Frisian counterpart can be quantified by the following formula:

\[ A \sim \log(frequency) / \text{LevenshteinDistance} \]

This formula entails the following two general predictions:

**General Predictions**

(i) the more similar a Frisian and a Dutch word form are, the more similarly they will behave in the grammar;

(ii) the larger the frequency of two equally similar, Frisian and Dutch, word forms, the more similarly they will behave in the grammar.

---

1 A model that incorporates similarity and frequency is Analogical Modeling (AM) proposed by Skousen (1989). It has recently been applied to explain diachronic changes in the distribution of the two rival suffixes –ity and –ness in English and to provide insight into the nature of their synchronic distribution (Arndt-Lappe 2014). Token frequency is abstracted away from in AM, since the model is characteristically used to measure morphological productivity of suffixes. However, we are primarily interested in the impact of absolute similarity and in the direct effect of token frequency in a single cause-effect chain. Our model can thus be viewed as an extension of AM proposed in order to make it fit for accounting for the influence of Dutch items on their Frisian counterparts in the mental lexicon of bilingual speakers.
In order to test these general predictions, we will investigate a linguistic phenomenon in Frisian which is sensitive to interference from Dutch, as described in Section 2. In many of the earlier quoted pieces of research, psycho-linguistic tests are used to study the interaction between language and the neural substratum of the cognitive system. An important difference with this type of research is that we use corpus data from written texts, that is, data that were never intentionally compiled for the purpose of psycholinguistic research. There is evidence that results obtained from corpus data analysis correlate with results obtained from psycho-linguistic tests (Krott, Baayen and Schreuder 2001). Such a convergence of research results strengthens the conclusions arrived at on the basis of the separate data sources and methodologies.

2 Subject and set-up of the investigation and morphological blocking principles

2.1 Subject of the investigation

Frisian has two suffixes, -ens and -heid, which are used to form nouns from adjectives. They often target the same base words and thus exhibit considerable competition (J. Hoekstra 1990; J. Hoekstra 1998; van der Meer 1986; van der Meer 1987; van der Meer 1988; E. Hockstra and Hut 2003). Some examples of derivational pairs are given below:

<table>
<thead>
<tr>
<th>Frisian</th>
<th>Dutch</th>
</tr>
</thead>
<tbody>
<tr>
<td>dúdlik</td>
<td>'clear'</td>
</tr>
<tr>
<td>dúdlikens</td>
<td>dúdlikheid</td>
</tr>
<tr>
<td>'clarity'</td>
<td></td>
</tr>
<tr>
<td>freedsum</td>
<td>'peaceful'</td>
</tr>
<tr>
<td>freedsumens</td>
<td>freedsumheid</td>
</tr>
<tr>
<td>'peacefulness'</td>
<td></td>
</tr>
<tr>
<td>warber</td>
<td>'industrious'</td>
</tr>
<tr>
<td>warberens</td>
<td>warberheid</td>
</tr>
<tr>
<td>'industriousness'</td>
<td></td>
</tr>
<tr>
<td>stom</td>
<td>'stupid'</td>
</tr>
<tr>
<td>stommens</td>
<td>stomheid</td>
</tr>
<tr>
<td>'stupidity'</td>
<td></td>
</tr>
</tbody>
</table>

Some base words take exclusively the suffix -ens, others take exclusively -heid, and many base words are used with both suffixes, but mostly not with the same frequency.

The Frisian suffix -ens does not have a cognate in Dutch that is recognisable for laymen. The Frisian suffix -heid, which is pronounced either as [hit] or [heit], is very similar to its Dutch 2 Some morphologists maintain that the existence of one form blocks the existence of a similar one. For example, in English, *gloriosity is claimed not to exist because it is blocked by the existence of gloriousness (Aronoff, 1976: 45). There are several problems with that account (see Di Sciullo and Williams, 1987:10ff), but there does appear to exist such a tendency in languages, see also Section 2.2.

3 The suffix is etymologically related to the Dutch suffix -nis. Apart from the difference in form, -nis and -ens also have a different distribution, since -ens attaches to adjectival base words whereas Dutch -nis attaches mainly to verbal stems (on Dutch -nis, see de Haas and Trommelen, 1993:245-246). It should be noted that Frisian has a few rare
counterpart \(-heid\). The latter pronunciation is identical to the Dutch pronunciation. The two suffixes also differ with respect to stress: \(-ens\) is unstressed, whereas \(-heid\) bears a secondary stress. Finally, Frisian / Dutch \(-heid\) may be pluralised and diminutivised, whereas Frisian \(-ens\) may not; the number of pluralised or diminutivised instances of \(-heid\) is very small in our corpus. They are excluded from further computations.

2.2 Blocking principles from the theory of morphology

Theories of morphology mostly assume that both words and suffixes with the same meaning and function are subject to a blocking principie (Rainer 1988). Rainer refers to the blocking of possible words by actual words as token blocking. An example of token-blocking is the blocking of German \#Blassheit \(\) by Blässe ‘paleness’ (cf. blass ‘pale’), and the blocking of \#Mutigkeit ‘litt. courageness’ by Mut ‘courage’ (cf. mutig ‘courageous’). Token-blocked words are well-formed morphologically, Rainer (1988: 162) argues, since they may occur as slips of the tongue, they may occur in child language even after children have learnt the correct word and they may be used for special effect, for example by poets. Token-blocking is sensitive to the frequency of the blocking word. Type-blocking, in contrast, involves a systematic curtailing of the domain of one suffix by another. Such a domain may be phonologically or semantically defined. An example is the relation between two competing suffixes: German \(-beit\) and \(-ität\). A word like *Grotteskität* is type-blocked by Grotteskheit, since the domain of \(-heit\) is defined to apply to adjectives bearing stress on the final syllable, whereas \(-ität\) is not subject to that restriction. Type-blocking is not sensitive to frequency, according to Rainer. Type-blocked words, it is implied, are not formed as slips of the tongue, nor in poetical language nor in child language.

It is not clear whether or how this distinction applies to the rivalry between Frisian \(-ens\) and \(-heid\).

The examples discussed by Rainer involve a description of the mental lexicon of monolingual native speakers. The examples which we will discuss are from Frisian, a language that is for almost no speaker the only language which they speak fluently: virtually all Frisian speakers are fluent in Dutch. Furthermore, the references to psycholinguistic literature make it clear that accessing words in one language will result in secondary activation of similar words in the other language of bilingual speakers. The simultaneous activation of similar words, regardless of the language which they belong to, indicates that there is only one mental lexicon for bilingual speakers, but it is structured differently from the mental lexicon of monolingual speakers: it will

\[\text{words in which } \text{-(e)nis} \text{ may appear, such as in tsjûgenis 'testimony', showing a preference for verbal stems, just as in Dutch.}
\]

To sum up, Frisian \(-ens\) has no equivalent in Dutch that is recognisable to laymen.
contain more items, at smaller distances from each other, and it is not a priori clear what the effects will be of this difference.

On the face of it, the rivalry between the two suffixes discussed here has characteristics both of token blocking and of type blocking. This rivalry is like token blocking in that the formation of the rival item is not radically blocked. In fact, there is often no blocking at all. However, the distribution between the two suffixes is regularly bimodal, so that one suffix is preferred with a particular item or the other suffix. The lack of a robust blocking effect might be due to speakers’ uncertainty about their Frisian, so that blocking appears to be a mere tendency of those who have inadequately mastered the language, in fact, the standard itself is absent. Just as in the study by Rainer, we notice a frequency effect for token blocking. On the other hand, the suffix –ens shows a sign of type-blocking in the domain of Frisian words with word final stress. Monosyllabic words have a clear preference for –ens, that is, 102 out of the 116 monosyllabic items have this preference. Nevertheless, the morphological distribution of and rivalry between the two suffixes may or may not be different from what is reported in the literature for monolingual speakers, for example, for the rivalry between –ity and –ness in English (e.g. the previously mentioned study of Arndt-Lappe 2014). An investigation of the morphological distribution of the Frisian suffixes falls outside the scope of this study, which aims at establishing the relation between frequency and similarity in assessing the influence of Dutch –heid formations upon the token frequencies of similar formations with the suffixes –ens and –heid in Frisian, although we included stress position as a controlling factor in further testing.

2.3 Set-up of the investigation

The part of a word to which a suffix is attached will be referred to as a base. A base may itself contain another suffix, as in reason+able+ness. Thus the notion of base is recursive. Every base can theoretically produce two lemmas, for example, the base dúdlik ‘clear’ returns both dúdílikens and dúdílikheid ‘clarity’. As they are synonymous and built on the same base, we consider the total number of tokens of dúdílikens and dúdílikheid as belonging to one ‘item’, dúdílik+suffix.

The distribution of the suffixes -heid and -ens was investigated by means of corpus data from written Frisian. The morphological database of the Frisian Language Corpus\(^4\) was used to identify a large number of nouns, either ending in -ens or in -heid and taking an adjective as a base form. Both derivations, with -ens and with -heid, were constructed from all identified adjectival base

---

forms. Frisian texts written in the period 1980-2000 were then checked for occurrences of the constructed forms. The raw data thus obtained comprise a list of slightly more than 700 adjectives that appear with either the suffix –ens, or the suffix –heid, or with both. Adjectives were ignored if they occurred with neither suffix. Furthermore, only items were used having a frequency of more than 3 in order to guarantee a minimal robustness of the data. The chance that an observed distribution of let’s say 4x –ens vs. 0x –heid differs from an observed distribution of 0x –ens vs. 4x –heid by mere chance is 0.029 (Fisher’s Exact Test)\(^5\), which is below the generally accepted level of significance of 0.05. Using the minimum item frequency of 4 (sum of all tokens of an adjective X + –ens or –heid) keeps distortion by chance within bounds. After the elimination of the low-frequent items, the list shrank to 336 items with 11,167 tokens. All tests and conclusions rely on that dataset.

3 Measuring frequency and similarity and the model’s further predictions

3.1 Predictions from the model about frequency and similarity

Word frequency will be determined by measuring the corpus frequencies of the Frisian words in a late-20th century text corpus.\(^6\) We take the Dutch frequencies for granted as being (quite) similar to their Frisian counterparts.

A Frisian base can be similar to a Dutch one with respect to either its meaning, or its form, or both. We took the most literary translation of the Frisian base word into Dutch, leaving the meaning ‘distance’ constant and close to zero, and computed the formal similarity of the phonetic form of the base as Levenshtein Distances (LD) with PMI (Pointwise Mutual Information) segment distances, based on Dutch and Frisian dialects.\(^7\) Examples include:

\[
\begin{align*}
\text{F. trystens} & \sim \text{D. triestheid} \text{‘sadness’}, \text{LD} = 0.00000 \\
\text{F. strangens} & \sim \text{D. strengheid} \text{‘strictness’}, \text{LD} = 0.00478 \\
\text{F. meagerens} & \sim \text{D. magerheid} \text{‘meagreness’}, \text{LD} = 0.01532 \\
\text{F. grutskens ‘pride’} & \sim \text{D. trots, LD} = 0.01746 \text{ (D. grootsheid means ‘grandeur’)} \\
\text{F. wurgens} & \sim \text{D. moeheid ‘tiredness’}, \text{LD} = 0.02883
\end{align*}
\]

\(^5\) The applet can be found at: [http://www.langsrud.com/fisher.htm](http://www.langsrud.com/fisher.htm). It is based on Agresti (1992).


\(^7\) For the meaning of Levenshtein-PMI, see Wieling, Margaretha and Nerbonne (2012).
In this way, similarity could be measured. The conceptual model outlined in Section 1 can now be applied to the distribution of -ens and -heid in Frisian. As there are no Dutch nominalisations ending in -ens, highly frequent items (in Dutch with -heid) will favour Frisian nominalisations in -heid, if they are sufficiently similar in form. Thus the model makes the following two specific predictions:

(1) Prediction about similarity:

Frisian nominalisations of which the bases are similar or identical in form and meaning to their Dutch equivalents will occur less often with -ens (and more often with -heid) than Frisian nominalisations with bases different from Dutch, all other things being equal.

(2) Prediction about frequency:

Frisian nominalisations which are more frequent will occur less often with -ens (and more often with -heid) than Frisian nominalisations which are less frequent, given a certain level of similarity between the Frisian and the Dutch form of the base.

In addition, we claimed in the introduction that the attraction which a Dutch word exerts on its Frisian equivalent involved a specific formula which is paraphrased informally as below.

(3) Prediction about force of attraction

The force of attraction is proportional to the measure of frequency divided by the measure of representational distance.

Before the results are presented in Section 4, we will explain in the next section, Section 3.2., how the number of forms was computed.

3.2 Computing amounts of forms

Every attestation in the corpus is either an instance of -ens or of -heid. To give an example: the corpus contains 2 instances of helderens and 8 of helderheid ‘clearity’. This implies that the proportion of -ens forms for this item is 20%. In this way, we can compute a '%-ens' for every item (and, trivially, '%-heid' = 1 - '%-ens').
Table 1: Four examples from the dataset of 336 items.

<table>
<thead>
<tr>
<th>Frisian</th>
<th>Fr - phonetic</th>
<th>Dutch</th>
<th>D - phonetic</th>
<th>LD-PMI</th>
<th>-ens</th>
<th>-heid</th>
<th>%-ens</th>
</tr>
</thead>
<tbody>
<tr>
<td>goederjousk</td>
<td>gu.ədrj.wsk</td>
<td>goedgeefs</td>
<td>‘generous’</td>
<td>0.01848</td>
<td>10</td>
<td>0</td>
<td>100.0%</td>
</tr>
<tr>
<td>ienlik</td>
<td>i.əlak</td>
<td>eenzaam</td>
<td>‘lonely’</td>
<td>0.03061</td>
<td>9</td>
<td>2</td>
<td>81.8%</td>
</tr>
<tr>
<td>helder</td>
<td>hɛldɾ</td>
<td>helder</td>
<td>‘clear’</td>
<td>0.00407</td>
<td>2</td>
<td>8</td>
<td>20.0%</td>
</tr>
<tr>
<td>bekend</td>
<td>bakɛnt</td>
<td>bekend</td>
<td>‘famous’</td>
<td>0.00000</td>
<td>0</td>
<td>41</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

Figure 1: Frequency distribution for %-ens per item; note that ‘>=0.0’ means ‘>=0.0 and < 0.2’, etc. Read: of all items in the dataset, 19% have a %-ens between 0% and 20%; when only high frequency items (N > 10) are included, 24% of the items fall in this class; etc.

Figure 1 above indicates the proportion of items of the total of items for five cohorts of percentage tokens per item in -ens. The graph shows that almost 50% (grey bar) of the items in our dataset has a proportion of 80% or more of the tokens in -ens. Of the total of 336 items, 221 show the suffix -ens in 50% or more of their tokens, which is a clear majority. To avoid extreme values produced by items of low frequency - for which purpose we already excluded items with less than four tokens -, the overall figures are compared to the figures for items with item frequencies over 10. Because a high item frequency correlates with a low %-ens (cf. discussion in 4), the bar “>=0.8” is slightly lower for high frequency items and, vice versa, the bar “>=0.0” somewhat higher. Still, the two distributions are very similar.

If the individual tokens randomly selected –ens and –heid, the computed percentages would show a normal or at least centralised distribution. However, tokens ending in –ens and –heid tend to be clustered per item. Almost every item, irrespective of the categorical preference on a higher level (cf. the factors mentioned in Section 3 and Section 4) has an individual preference for either of
the two endings. This is illustrated in Figure 1 by the bimodal pattern, with the peaks on the extreme values $\geq 0\%$ and $\geq 80\%$. Note further that 47\% of all items with an item frequency $> 10$ has exactly either 100\% or 0\% tokens with $-ens$.

In the evaluation of the correlations between the dependent variable of the proportion of tokens with $-ens$ per item and the independent variables of frequency and similarity, we want to determine their relevance by statistical testing. We applied a correlation and regression analysis to the similarity (LD) as the independent and the proportion of tokens with $-ens$ per item as the dependent variable. However, the residues are not normally distributed, which disqualifies this type of analysis. This is probably due to the bimodal distribution of the dependent variable (the proportions of $-ens$ and $-heid$). We therefore converted the dependent variable into a binary one by counting every item with a proportion of $-ens < 50\%$ as being a “heid-item” and with a proportion of $-ens \geq 50\%$ as being an “ens-item”. Items which have been categorised in this manner will be referred to as bimodally categorised items.

Consecutively, we tested the correlation between the factors of similarity and frequency on the one hand, and the bimodally categorised items on the other, in a logistic regression model. Finally, we checked whether the measure of ‘attraction’ between Frisian and Dutch items in the brain, expressed by the (log of the) frequencies divided by the formal dissimilarity produced a statistically significant prediction for the choice between $-heid$ and $-ens$.

It should be mentioned that there is a range of prosodic, semantic and stylistic factors, the most important being syllable structure, that additionally guide the choice between $-heid$ and $-ens$. The factor of syllable structure, in particular the place of the stress, was included as an additional independent variable in the logistic regression model. Because the suffix $-ens$ is unstressed, but $-heid$ bears secondary stress, words that have the stress on the final syllable of the base (including monosyllabic words) are expected to show $-ens$ more often. Stress is a binary variable

---

8 There are 9 items with exactly 50\% of their tokens in $-heid$ and 50\% in $-ens$. One of the reviewers suggested us to use the actual tokens with $-ens$ or $-heid$ as exemplars of the dependent variable in the logistic regression analysis. We feel that this would strongly enlarge the effect of token frequency, as token frequency is one of the independent variables. Using all attestations as exemplars in the regression analysis implies that items with a high token frequency are more often evaluated in the analysis, increasing the impact of token frequency in the final equation. But see also footnote 14.

9 In the corpus Early Modern Frisian (1550-1800), the suffix $-ens$ is only attested in a small set of 29 nouns with an adjectival root, which is monosyllabic in 24 instances. Remarkably, for 9 out of the 10 words that show both $-ens$ and $-heid$ in this period, the date of attestation of the $-heid$ form is earlier than of the $-ens$ form. This suggests that the
in this study: final yes/no. Our primary interest, however, is the question whether the features mentioned in our hypothesis have a significant impact on the choice between –heid and –ens.

4 Results

4.1 Prediction 1 and 2: similarity of Frisian bases with Dutch and impact of the item’s frequency

Our prediction 1 states that similarity of the Frisian base with the Dutch cognate will facilitate the use of –heid (and disfavour the use of –ens). Our prediction 2 states that high frequency of the Frisian (and Dutch) derivation will disfavour the use of -ens (and facilitate the use of –heid). The LD may be assumed to be normally distributed (Anderson-Darling normality test, p = 0.201).

To obtain a normally distributed independent variable, we work with the logarithm of the frequency figures (Anderson-Darling normality test, p = 0.144) (see DeHaene 2003 for a merithorical reason). It is worth noting that the two variables of frequency and similarity are largely independent. Their correlation is only -0.120 ($r^2 = 0.014$, $p = 0.028$), which is basically negligible.

A logistic regression model with LD, log(frequency) and stress position as independent variables and the bimodally categorised items as dependent variable, with 0 = ‘-heid’ and 1 = ‘-ens’, shows the following descriptives:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Avg</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.3217</td>
<td>0.2301</td>
</tr>
<tr>
<td>2</td>
<td>0.3767</td>
<td>0.1602</td>
</tr>
<tr>
<td>3</td>
<td>0.3452</td>
<td>0.4754</td>
</tr>
</tbody>
</table>

Overall Model Fit...
Chi Square= 76.0627; df=3; p= 0.0000

Coefficients and Standard Errors...
Variable Coeff. StdErr  p
1  2.9191  0.7002  0.0000
2  -2.4512 0.8164  0.0027
3   1.8699 0.3327  0.0000

The current distribution of –ens and –heid is the result of a 19th century innovation in the emerging standard language which favours the suffix –ens because it is more distinct from Dutch.

The applet can be found at: [http://www.xuru.org/st/DS.asp#CopyPaste](http://www.xuru.org/st/DS.asp#CopyPaste). The site does not mention any literature on which the applet is based.

The applet can be found at: [http://statpages.org/logistic.html](http://statpages.org/logistic.html). It is based on Hosmer and Lemeshow (1989).
Intercept 0.2488

Odds Ratios and 95% Confidence Intervals...

<table>
<thead>
<tr>
<th>Variable</th>
<th>O.R</th>
<th>Low -- High</th>
<th>0.0862 = 11.6009</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>18.5248</td>
<td>4.6962 -- 73.0742</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.0862</td>
<td>0.0174 -- 0.4269</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>6.4874</td>
<td>3.3795 -- 12.4535</td>
<td></td>
</tr>
</tbody>
</table>

About the Odds Ratios, the applet website mentions: “The odds ratio for a predictor tells the relative amount by which the odds of the outcome increase (O.R. greater than 1.0) or decrease (O.R. less than 1.0) when the value of the predictor value is increased by 1.0 units.” The LD was scaled to the range from 0 for identical forms in Frisian and Dutch, such as blau ‘blue’ to 0.99 for the entirely different word forms Frisian lilk, Dutch boos ‘angry’. The log(freq) range was rescaled to the range between 0.2 and 1.

Because both the value ranges and the averages of the three variables are similar, the coefficients and Odds Ratios in the model are fairly compatible.

The model shows that apart from the stress placement, which indeed makes a significant contribution to the preference for either –heid or –ens, the LD and the log(freq) have a mirrored effect, as predicted by the hypotheses 1 and 2, with a similar effect size: coefficient values: LD = +2.9, log(freq) = – 2.5.

4.2 Prediction 3: attraction between words

As both considered variables, frequency and similarity, have a significant impact on the lexical distribution of the suffixes –ens and –heid in Frisian, in line with the hypotheses formulated, we tested the impact of the combination of both variables on the distribution of forms with –ens and –heid. For the purpose of this test, we constructed a new variable Attraction =

---

12 Partially overlapping forms such as Frisian boeden, Dutch behoedzaam ‘careful’ have values around 0.6.

13 The word mooglikheid ‘possibility’ has the value 1 and represents – at least for this corpus – the most frequent derivation of an adjective on –heid or –ens.

14 Following a suggestion made by one of the reviewers, we tested the 160 items with a 100% preference in the corpus for either –heid or –ens separately. All three variables have a statistically significant contribution to the model and the coefficients are: LD: 4.7085, log(freq): -3.6652; stress: 2.8875. For the 176 items with variation, the model as a whole is statistically significant (p = 0.0149). The coefficient values are: LD: 1.6415 (p = 0.069), log(freq): -1.9035 (p = 0.063); stress: 1.1218 (p = 0.022). When using the individual exemplars of the latter group of 176 items, all p-values are smaller than 0.0001 and the coefficient values are: LD: 1.6212, log(freq): -5.1042; stress: 0.6071, with a clear over-weighting of the frequency (see footnote 8).
The variable gives a linear prediction for the proportion of words in –ens and –heid as shown in Figure 2. This relation is also confirmed in a logistic regression model with Attraction as the independent variable and the bimodally categorised items as dependent variable, with 0 = ‘-heid’ and 1 = ‘-ens’.

108 cases have Y=0; 205 cases have Y=1.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Avg</th>
<th>SD</th>
<th>Attraction (scaled to the range 0-1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.1207</td>
<td>0.1381</td>
<td></td>
</tr>
</tbody>
</table>

Overall Model Fit... Chi Square = 28.8369; df=1; p = 0.0000

Coefficients, Standard Errors, Odds Ratios, and 95% Confidence Limits...

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coeff.</th>
<th>StdErr</th>
<th>p</th>
<th>O.R.</th>
<th>Low -- High</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-5.1421</td>
<td>1.1046</td>
<td>0.0000</td>
<td>0.0058</td>
<td>0.0007 -- 0.0509</td>
</tr>
<tr>
<td>Intercept</td>
<td>1.2745</td>
<td>0.1809</td>
<td>0.0000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

![Figure 2: The relation between Attraction (X-axis) and the proportion of bimodally categorised items with a preference for the suffix –heid (Y-axis). The graph shows that higher Attraction leads to more items with the ending –heid. p = 0.002 (2-tailed).](image)

We used a class width of 20 units of the variable Attraction. Only the last two classes are wider in order to keep a substantial amount of items per class. The figures above the points show the number of items per class. Items with LD = 0 in the denominator had to be eliminated, leaving 313 items.

The freq/LD-space where the Attraction-values are taken from, is not linearly distributed, as illustrated in Figure 3. High levels of Attraction (Z-axis) are only attained for a combination of high values in the numerator (= frequency) and low values in the denominator (= LD).

---

15 The multiplication factor is only to scale the variable to values with not too many digits or decimals.

16 The applet can be found at: [http://vassarstats.net/corr_big.html](http://vassarstats.net/corr_big.html). The site contains extensive background documentation at [http://vassarstats.net/textbook/index.html](http://vassarstats.net/textbook/index.html).
Figure 3: The freq/LD-space as expressed in the variable *Attraction*.

As outlined in Section 4.1, both the LD-variable and the log(freq)-variable show a normal distribution, which implies that most items have low *Attraction*-values, while high *Attraction* values are relatively rare. This is reflected in Figure 1 with fewer items with 100% *-heid* than with 100% *-ens*. This skewness is also reflected in the number of items per class in Figure 2.

Figure 2 shows that an increase of *Attraction* implies a higher chance for an item to have a preference for *-heid*. The linear correlation shows that there is not one *Attraction*-threshold value that divides the set of items into two categories.

5 Concluding remarks

The choice between *-ens* and *-heid* in Frisian nominalisations turns out to be sensitive to the degree of their representational convergence with Dutch word forms and the strength of their neural representation. Words’ representational convergence or nearness in the brain was calculated by proxy by Levenshtein Distance. The neural strength was calculated by proxy by determining the frequency in a corpus. The force of *Attraction* (A) between Dutch words and their Frisian counterparts is defined by the formula:

$$ A \sim 2 \log(\text{frequency}) / \text{LevenshteinDistance} $$

On this *Attraction*-scale, the items that undergo a strong attraction from Dutch, tend to have a preference for the suffix *-heid*, that also exists in Dutch, while words exposed to little attraction from Dutch prefer the suffix *-ens*, that is unique to Frisian.
Our research hypothesized on the basis of results reported in the psycholinguistic literature that the linguistic system of a minority language such as Frisian, where speakers have been at least passively bilingual for several centuries, is profoundly shaped by the characteristics of the bilingual mind of its speakers, also as reflected in its standardised and written form. A written corpus, so to speak, bears the fingerprint of the cognitive system of the language users who composed the texts. This allowed us to get a grasp on the question why certain bases are more used with –heid, others with –ens. In addition, it allowed us to represent the influence of Dutch items on their Frisian counterparts in the mental lexicon by means of a formula making explicit the strength of this influence in terms of frequency and similarity.

Finally, if Frisian is not spoken and written more than has been done up till now, the formula we have proposed implies that Frisian is slowly yet inexorably changing in the direction of Dutch. It would be interesting to conduct diachronic research in order to establish the speed with which the process of convergence takes place. Though a task for future research, this could potentially result in a prediction as to whether and, if so, when, for a given process, convergence will be completed.

**Acknowledgements**

We would like to extend our gratitude to Martijn Wieling and John Nerbonne for making the LD-computations for us, and to Harald Baayen and an anonymous reviewer of Linguistics for their useful comments.

**References**


Barbiers, Sjef, Johan van der Auwera, Hans Bennis, Eefje Boef, Gunther De Vogelaer & Margreet


