Linguistic determinants of the intelligibility of Swedish words among Danes

Sebastian Kürschner, Charlotte Gooskens, and Renée van Bezooijen
University of Groningen

Abstract

In the present investigation we aim to determine to which degree linguistic factors contribute to the intelligibility of Swedish words among Danes. We therefore correlated the results of an experiment on word intelligibility with eleven linguistic factors and carried out regression analyses. In the experiment, the intelligibility of 384 frequent Swedish words was tested among Danish listeners via the internet. The choice of eleven linguistic factors was motivated by their contribution to intelligibility in earlier studies. The highest correlation is found between word intelligibility and phonetic distances. Also orthography, different syllable numbers, neighbourhood density, and word frequency contribute significantly to intelligibility. However, the correlation is very low ($R = .51$), and the percentage of explained variance is not higher than 26% ($R^2 = .26$). The low correlation is attributed to the higher number of idiosyncrasies of single words compared with the aggregate intelligibility and linguistic distance used in earlier studies. Based on observations in the actual data from the intelligibility experiment, we suggest further steps to be taken to improve the predictability of word intelligibility.

1. Introduction

Danish and Swedish are closely related languages within the North Germanic language branch. The two languages are mutually intelligible to such a high degree that in Danish-Swedish communication speakers mostly use their own mother tongues, a mode of communication termed „semi-communication“ by Haugen (1966). In previous research it was shown that intelligibility scores correlate highly with global phonetic distances between the languages involved (cf. e.g. Beijering et al. 2008, Gooskens 2007). Hence, linguistic factors
play a major role in determining mutual intelligibility. Additionally, it is often assumed that attitudes and prior exposure to the variety in question play a role (e.g. Delsing & Lundin Åkesson 2005). However, correlations are low and the direct relationship is difficult to prove (Van Bezooijen & Gooskens 2007).

Earlier research has mostly involved testing text understanding. Intelligibility scores were based on the text as a whole. This means that the influence of different linguistic dimensions such as textual and sentence context, morphology, and phonology could not be distinguished. In our study, we wanted to determine the role of linguistic factors in more detail. Therefore, we chose to focus on single words instead of sentences or texts. We conducted an internet experiment assessing the intelligibility of isolated Swedish words among Danish subjects, excluding the influence of sentence context. The underlying assumption here is that word recognition is the key to speech understanding. If the listener correctly recognizes a minimal proportion of words, he or she will be able to piece the speaker’s message together. In particular, we tested the impact on word intelligibility of linguistic factors such as segmental and prosodic phonetic distance, Swedish sounds lacking in the Danish sound system, word frequency, and orthography. In this way we hoped to obtain more detailed information on the precise role of various linguistic factors in the intelligibility of Swedish words among Danes.

The way in which we tested intelligibility on the internet may be relevant for research in other experimental disciplines within the humanities such as psycholinguistics, neurolinguistics, and psychology. Furthermore, the algorithms we used to measure linguistic distances might be of interest to any discipline in need of tools for automatic comparison of numbers or strings, for example in history and literary studies. The computationally based methods for intelligibility and distance measurement are also highly relevant for inter-disciplinary studies combining political and linguistic sciences concerned with the multi-lingual Europe.

2. Experiment

To test word intelligibility, an internet-based experiment was conducted.1 In this experiment, Danish subjects were confronted with 384 isolated Swedish nouns. These nouns were
randomly selected from a list of 2575 highly frequent words. In a pre-test, we assured that all these nouns were known to subjects from the test group, i.e. pupils aged 16-19. The 384 words were read aloud by a male Swedish native speaker and recorded in a professional sound studio. Each subject heard one quarter, i.e. 96 of the 384 Swedish words and was requested to write the Danish translation into a text field within ten seconds. Prices were promised to the participants, and especially to the best-scoring participants, to stimulate the subjects to make an effort to do well. The choice of the words and the order of presentation were randomized in order to reduce tiredness effects. Since the word blocks were automatically assigned to the subjects in random order, some word blocks were presented to more subjects than others. The lowest number of subjects who heard a word block was five, the highest number 19, with an average of nine subjects per word block.

36 monolingual Danish pupils aged 16-19 participated in the experiment. Since we are interested in intelligibility at a first confrontation, we needed subjects who had had little contact with the test language. We therefore excluded subjects living in regions close to the Swedish border. As an extra precaution, we also had the subjects translate a number of Swedish non-cognates. Such words should be unintelligible to subjects with no prior experience with the language. Indeed, hardly any of the non-cognates were translated correctly. An exception is formed by the word *flicka* ‘girl’ (Danish *pige*), which was translated correctly by 73% of the subjects. This word is probably known to most Danes as a stereotypical Swedish word, among others because it was used in the popular Danish pop song *sköna flicka* (‘beautiful girl’) by Kim Larsen. Furthermore, one non-cognate was translated correctly by 50% of the subjects and 3 three other non-cognates were translated correctly once by three different subjects. On the basis of these result we decided not to exclude any of the 36 subjects.

The results were automatically categorized as right or wrong through a pattern match with expected answers. Those answers which were categorized as wrong were subsequently checked manually by a Danish mother tongue speaker. Responses which deviated from the expected responses due to a mere spelling error were counted as correct identifications. Spelling errors were objectively defined as instances where only one letter had been spelt
wrongly without resulting in another existing word. So, for example the mistake in ærende (correct ærinde) ‘errand’ is considered a spelling mistake and therefore counted as correct (only one wrong letter without resulting in another existing word), while aske (correct æske ‘box’) was not counted as correct because the spelling mistake results in an existing word meaning ‘ash’. Some Swedish words have more than one possible translation. For example the Swedish word brist ‘lack’ can be translated into Danish brist or mangel, both meaning ‘lack’. Both translations were counted as correct. In the case of homonyms, both possible translations were accepted as correct. For example, Swedish här can be translated correctly into Danish hær ‘army’ or her ‘here’.

After this procedure, we had obtained a score of zero (word not identified) or one (word identified) per word for each subject. We then calculated the percentage of correct translations per word. This percentage was the intelligibility score per word. The obtained scores were subsequently correlated with linguistic factors to determine the degree to which linguistic factors determine intelligibility.

We only investigated the intelligibility of cognates since non-cognate forms should, almost by definition, be unrecognizable. Cognates are historically related word pairs that still bear the same meaning in both languages. We use a broad definition of cognates, including not only shared inherited words from Proto-Nordic, but also shared loans such as Swedish/Danish perspektiv ‘perspective’, which is loaned from the same Latin source in both languages. We also excluded words that have a cognate root but a derivational morpheme that is different from the corresponding Danish cognate. So, for example, the word pair Swedish undersökelse Danish undersøgning ‘examination’ was excluded from the analyses. Of the 384 Swedish nouns, 347 proved to be cognate with Danish nouns.

3. Factors considered for explanation

In this section we will explain eleven factors that we considered to be possible determinants of the variance in the intelligibility scores. Most of the factors are known to play a role in word intelligibility from psycho-phonetic literature. Other factors are assumed to play a role in the special case of Swedish-Danish communication by Scandinavian scholars. We aimed to
include as many factors as possible. However, we were limited by the fact that they had to be quantifiable, since we wanted to test their contribution to intelligibility statistically.

3.1 Levenshtein distance

As mentioned in the introduction, global phonetic distances between languages have proved to be good predictors of intelligibility of whole texts. Also at the word level small phonetic distances can be assumed to correlate with high intelligibility scores, while large distances can be expected to correlate with low intelligibility scores. We measured the phonetic distances by means of the Levenshtein algorithm.

Levenshtein distance is a measure of string edit distance based on the smallest number of operations necessary to map a string of sounds from one variety on the corresponding string in another variety (cf. Heeringa 2004). Insertions, deletions, and substitutions are possible operations. The example in Figure 1 shows the calculation of string edit distance between Danish and Swedish *guld* ‘gold’, pronounced as [gul] in Danish and as [guld] in Swedish. First, the two strings are aligned, with identical sounds being matched with each other (cf. [g] and [l]). Subsequently, the number of operations necessary to change one string to another is calculated. In our example two sounds are identical and therefore they do not add any costs. In contrast, operations are necessary for the vowel which has to be substituted, and for the final sound in Swedish which has to be inserted in order to change the Swedish pronunciation into the Danish. Since operations have to be conducted at two slots, the Levenshtein distance is 2. To relate the distance to word-length, we divide it by the highest number of sounds in the longest string involved, i.e. 4 in the example. The normalised distance is $2/4 = 0.5$, i.e. 50 % for our example.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Swedish</strong></td>
<td>g</td>
<td>u</td>
<td>l</td>
<td>d</td>
</tr>
<tr>
<td><strong>Danish</strong></td>
<td>g</td>
<td>u</td>
<td>l</td>
<td></td>
</tr>
<tr>
<td>subst.</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>insert.</td>
<td></td>
<td></td>
<td></td>
<td>= 2</td>
</tr>
</tbody>
</table>

Figure 1: Calculation of Levenshtein distance.
In order to obtain distances which are based on linguistically motivated alignments that respect the syllable structure of a word or the structure within a syllable, the algorithm was adapted so that in the alignment a vowel may only correspond to a vowel and a consonant to a consonant. The semi-vowels [j] and [w] may also correspond to a vowel or the other way around. The central vowel schwa [ə] may correspond to any sonorant. In this way, unlikely matches – like [o] and [t] or [s] and [e] – are prevented.

Additionally, the operations were weighted according to the following conditions. The difference was only increased by 0.5 if (1) the difference was not based on qualitative, but only on quantitative grounds (cf. [u] vs. [u:]), (2) the difference was only based on [+/-tense], cf. [u] vs. [u], or (3) the corresponding sounds were [R] and [r], the first of which is used in Standard Danish, and the second of which is used in Standard Swedish. In all other cases, the distance was increased by 1.0.

Levenshtein distance was calculated automatically for all 347 pairs of cognates. The mean distance across all words was 52.1%. Eight word pairs had a distance of zero, for example Swedish ämne – Danish emne ‘subject’ that are both pronounced as [ɛmne]. Six word pairs had the maximum distance of 100%, for example Swedish ljud [juːd] Danish lyd [lyːˈd] ‘sound’.

### 3.2 Foreign sounds

When a listener being confronted with a language (variety) for the first time hears unusual or unknown sounds, he may be distracted and this may influence intelligibility negatively. To explore the effect of this factor we listed for each Swedish word the number of sounds which do not exist in the Danish sound system. The following sounds are described in the literature as foreign (from Karker, Lindgren & Løland 1997; Nordentoft 1981):

- The postalveolar-velar fricative [ʃ], cf. Swedish aktion [akʃuːn].
We only considered single sounds, i.e. in our list of foreign sounds we did not include any combinations of sounds which exist in Danish but are phonotactically uncommon, cf. [l̥] or [ŋn]. Neither did we include sounds which are possible in the Danish system, but in contrast to Swedish do not establish a phonemic opposition, such as long plosives, some voiced consonants, and the vowel [u]. 46 of the words contained a retroflex consonant or a postalveolar-velar fricative. Three words even contained two foreign sounds: koordination ‘coordination’ [kuːɛ̃naːjʊn], ordning ‘order’ [oːŋŋ], and stjärna ‘star’ [ʃæːn].

3.3 Word length

Previous research has shown that word length plays a role in word recognition (Wiener & Miller 1946, Sharpf & Van Heuven 1988). According to these studies longer words are better recognized than shorter words. This, in turn, is explained in terms of the relationship between word length and the number of ‘neighbours’, i.e. competing word forms that are very similar to the stimulus word (see Section 3.7). Longer words have fewer neighbours than shorter words (Vitovitch & Rodriguez 2005). Furthermore, redundancy increases with word length, which is assumed to also enhance intelligibility (see Section 3.7). Swedish words were annotated for word length in terms of the number of phonetic segments. The mean word length across all words was 5.57 segments. The four longest words consisted of 12 segments, for example uppmärksamhet [ʊpːmaɛːrsam̩ːhɛt] ‘attention’, while the shortest word had only one segment, ö [œː] ‘island’.

3.4 Word accent differences

Van Heuven (1985) found that correct recognition of words was severely reduced and delayed if medial or final stress was shifted to the initial syllable in Dutch words. Extrapolating this result he hypothesized that unexpected stress positions play an important (negative) role not only in understanding the mother tongue but also speech in a closely related variety (Van Heuven this volume). For each Swedish word, we annotated whenever the place of the word accent was different from that in the corresponding Danish cognate, assuming that such a difference makes the word more difficult to identify. Danish kontekst [kɔntɛksːt] vs. Swedish
kontext [kɔntɛkst] ‘context’ may serve as an example of word accent differences, which were found in ten of the word pairs.

3.5 Differences in number of syllables

Cognates between Danish and Swedish can differ in the number of syllables, cf. Danish mångde [mæŋdə] vs. Swedish mängd [mɛŋd] ‘quantity’. Since a missing or extra syllable could cause confusion in word identification, we annotated instances with different syllable numbers. Ten of the Swedish words contain one syllable extra compared to the corresponding Danish word, while 22 words have one syllable less. Two Swedish words were even two syllables shorter than the Danish cognate, namely choklad [ʃɔklɑːd] (Danish chololade [tʃoɡolɑːdə]) ‘chocolate’ and tjänst [tʃɛnst] (Danish tjeneste [tʃeːnɛst]) ‘service’.

3.6 Lexical tones

According to Van Heuven (this volume), in ideal circumstances the contribution of word prosody to the process of word recognition is a modest one. Because word prosody is a slowly varying property of the speech code, it will normally not be needed in the recognition of words. However, when communication suffers from noise, prosody fulfils the role of a safety catch. Listening to speech in closely related language is similar to listening to speech in noise. Therefore differences in presence and realization of lexical tones are predicted to be detrimental to word recognition. In Danish, no lexical tones are used, while Swedish has two word tones, an acute accent (or Accent I) and a grave accent (Accent II). Minimal pairs occur, e.g. `anden (Acc. I), definite singular form of and ‘duck’ vs. `anden (Acc. II), def. sg. form of ande ‘spirit’. We annotated the Swedish test words with the word accent. 253 words had accent I and 94 had accent II.

3.7 Stød

Danish has a special prosodic feature at the word level which does not occur in Swedish. The so-called ‘stød’ is a kind of creaky voice. It occurs in long vowels and in voiced (sonorant) consonants. Presence versus absence of stød creates an abundance of minimal contrasts, for
example [hɛn’u] ‘hands’ versus [hɛnø] ‘happen’. We assumed that the absence of this phenomenon in corresponding Swedish words may cause confusion on the part of the Danish listeners. However, since stød is also missing in several Danish dialects to the south of the stød isogloss without any reported influence on intelligibility, the influence on intelligibility may be limited. Nevertheless, this has never been tested experimentally. We annotated the words which include a ‘stød’ in the database. 164 words have a stød, and 161 words have no stød.

3.8 Neighbourhood density

Neighbours are linguistically defined as word forms that are very similar to the stimulus word and may therefore serve as competing responses. For an extensive description of the neighbourhood activation model, see Luce & Pisoni (1998). Since a high neighbourhood density enlarges the number of possible candidates for translation, we assume that the higher the density is, the lower the number of correct identifications will be. Short words in general have a denser neighbourhood. From this we would predict that the possible advantage of short words being more frequent than long words (see Section 3.11), is neutralised by the neighbourhood density problem.

Here we define neighbourhood density as the number of Danish words which deviate from the Swedish stimulus in only one sound, disregarding the correct counterpart. For example, the Swedish word säng ‘bed’ with the correct Danish translation seng has five Danish neighbours: syng ‘sing’, senge ‘beds’, hæng ‘hang’, and stæng ‘close’, while the Swedish word adress ‘address’ has no neighbours. For each Swedish word we counted the number of neighbours in Danish. The mean number of neighbours is 0.93. 244 words have no neighbours and the largest number of neighbours was 16 for the Swedish word ö ‘island’.

3.9 Etymology

Gooskens, Bezootjen & Kürschner (accepted) showed that loan words that have been introduced into both Swedish and Danish are easier to understand by Danish subjects listening to Swedish than native cognate words. Presumably this is due to the fact that the loan words
were affected by fewer sound changes differentiating Swedish from the other Nordic languages than native words. The Swedish words in our database were categorized according to their etymology. We distinguished between native words and loan words. All words originating in Proto-Germanic which, as far as we know, have been present in Swedish at all times are defined as native words. There were 196 native words. All words which were newly introduced as loans from other languages are defined as loan words, i.e. even words of Proto-Germanic origin which have been lost in Swedish and were re-introduced through language contact. For this reason, also the quite high number of Low-German words, which have been introduced due to the strong language contact with the Hanseatic league in the Middle Ages, is part of the loan word group. 151 of the words were loan words.

3.10 Orthography

There is evidence that knowledge of orthography influences spoken word recognition (e.g. Ziegler & Ferrand 1998, Chereau, Gaskell & Dumay 2007). The evidence comes from experiments with words that differ in degree of sound-to-spelling consistence and from recent neuroimaging research (Blau et al. 2008). Doetjes & Gooskens (accepted) correlated the percentages of correct translations of 96 words with simple Levenshtein distances between the Swedish and Danish pronunciations and got a correlation of .54. Next, they measured the Levenshtein distances again but this time corrected the distances in such a way that they took into account that Danes may be able to use the Danish orthography when decoding certain Swedish spoken words. The corrected distances showed a higher correlation with the intelligibility scores ($r = .63$), which provides evidence that Danes have support from their own orthography when hearing Swedish words.

Danish is generally described as the Scandinavian language which has gone through the most drastic sound changes (Brink & Lund 1975, Grønnum 1998). The number of sound-letter correspondences has therefore decreased heavily. In contrast, the sound changes in Swedish have not differentiated spoken and written language to such a high extent. In some cases, spoken and written language have even converged because people tend to pronounce the words as they are spelled (Wessén 1965:152, Birch-Jensen 2007). It is therefore plausible that
Danes may use their orthographical knowledge in the identification of Swedish words. To measure the help Danes might get through their orthography, for each word we counted the number of Swedish sounds which (1) did not match with a corresponding sound in Danish, but (2) were equivalent with the corresponding letter in Danish. For example, consider the different pronunciations of the words for ‘hand’: Danish \( \text{hånd} \) [høn?] vs. Swedish \( \text{hand} \) [hænd]. The final consonant is not pronounced in Danish but it can be assumed that Danish subjects identifying the Swedish word make use of their knowledge of Danish orthography which includes the consonant. For this reason the insertion of the \( d \) was given one point in this example. The mean number of sounds that could be identified by means of the Danish orthography was 1.27, with a minimum of 0 (118 of the words) and a maximum of 6 in one case.

### 3.11 Danish word frequency

We assume that the token frequency of words may influence correct identification, since frequent words are more likely to come to the subjects’ minds immediately than infrequent words. The activation of a word that was recognized before remains high for a long time, and never fully returns to its previous resting level. Highly frequent words therefore have a permanent advantage in the recognition process (Luce & Pisoni 1998).

Since we make assumptions about the performance of the Danish subjects, the frequency in their mother tongue must be decisive. We therefore annotated all words for token frequency in Danish. The numbers were based on the frequency list of a large written language corpus, the Korpus 2000.\(^2\) The most frequent word was \( \text{dag} \) ‘day’, which occurred 222159 times in the corpus. There were seven stimulus words which did not occur in the corpus and thus had a frequency of 0. The least positive frequency was found for \( \text{overføring} \) ‘transmission’, which occurred 11 times.

### 4. Results

The intelligibility test resulted in an overall percentage of 57.7 % correct identifications of the Swedish cognates among the Danish subjects. For each word, the total number of correct
identifications was divided by the number of subjects tested with this word. In this manner, we obtained an intelligibility score ranging from zero (the test word was not understood by any of the subjects) to one (the test word was understood by all subjects tested) for each of the test words. This score was correlated with the linguistic factors described in Section 3. The results are shown in Table 1:

<table>
<thead>
<tr>
<th>Factor</th>
<th>Correlation ($r$)</th>
<th>Significance ($p$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Levenshtein distance</td>
<td>-0.33</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Foreign sounds</td>
<td>-0.13</td>
<td>&lt; .05</td>
</tr>
<tr>
<td>Word-length of Swedish words</td>
<td>0.25</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Word accent difference</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td>Difference in syllable number</td>
<td>-0.25</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Tone accents</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td>Stød</td>
<td>-0.02</td>
<td></td>
</tr>
<tr>
<td>Neighbourhood density</td>
<td>-0.17</td>
<td>&lt; .01</td>
</tr>
<tr>
<td>Etymology</td>
<td>-0.1</td>
<td></td>
</tr>
<tr>
<td>Orthography</td>
<td>0.21</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Word frequency</td>
<td>0.04</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Correlation of the intelligibility scores with linguistic factors

In general, correlations and the amount of variance explained by each predictor are low. Relatively, Table 1 reveals the highest correlation between the intelligibility scores and Levenshtein distance, which confirms results from previous research that phonetic distance is an important predictor of intelligibility (see Section 1). Nevertheless, the correlation is rather low ($r = -0.33$), so other factors must affect intelligibility as well. Significant correlations can be observed for word length of Swedish words ($r = 0.25$), difference in syllable number ($r = -0.25$), neighbourhood density ($r = -0.17$), and orthography ($r = 0.21$).

Five factors do not correlate significantly with intelligibility. Three of these factors are prosodic (word accent difference, tone accent and stød). This confirms the observation by Van Heuven (this volume) that the contribution of word prosody to the process of word recognition is a modest one. The low correlation of word frequency with intelligibility may be explained by the fact that all words were selected from a database with frequent words so that all words were in fact highly activated. Contrary to our expectations (see Section 3.9), also etymology does not correlate significantly to intelligibility.
We conducted multiple regression analyses to identify exactly how much of the variance can be explained by the considered factors, and which factor combination reveals the best prediction for intelligibility (cf. Table 2). The analyses show that the linguistic factors can explain the variance partly, but not to a high extent. In enter mode including all factors, we arrive at a correlation of .51 with $R^2 = .26$, i.e. only 26% of the variation can be explained through the factors considered. The scatter plot of the enter-mode analysis in Figure 2 shows that the considered factors only explain a small part of the variation, since the data points are wide-spread and only loosely packed about the regression line.

The stepwise analysis reveals five models. Levenshtein distance is revealed as the most important factor with a correlation of .34. The second model includes orthography, which

<table>
<thead>
<tr>
<th>method</th>
<th>entered variables</th>
<th>model</th>
<th>result</th>
</tr>
</thead>
<tbody>
<tr>
<td>enter</td>
<td>all linguistic factors</td>
<td>all linguistic factors</td>
<td>$R = .51, R^2 = .26$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$p &lt; .001$</td>
</tr>
<tr>
<td>stepwise</td>
<td>all linguistic factors</td>
<td>Levenshtein distance</td>
<td>$R = .34, R^2 = .11$</td>
</tr>
<tr>
<td>1st step</td>
<td></td>
<td></td>
<td>$p &lt; .001$</td>
</tr>
<tr>
<td>stepwise</td>
<td>all linguistic factors</td>
<td>Levenshtein distance, orthography</td>
<td>$R = .44, R^2 = .19$</td>
</tr>
<tr>
<td>2nd step</td>
<td></td>
<td></td>
<td>$p &lt; .001$</td>
</tr>
<tr>
<td>stepwise</td>
<td>all linguistic factors</td>
<td>Levenshtein distance, orthography, different syllable number</td>
<td>$R = .47, R^2 = .22$</td>
</tr>
<tr>
<td>3rd step</td>
<td></td>
<td></td>
<td>$p &lt; .001$</td>
</tr>
<tr>
<td>stepwise</td>
<td>all linguistic factors</td>
<td>Levenshtein distance, orthography, different syllable number, neighbourhood density</td>
<td>$R = .49, R^2 = .23$</td>
</tr>
<tr>
<td>4th step</td>
<td></td>
<td></td>
<td>$p &lt; .001$</td>
</tr>
<tr>
<td>stepwise</td>
<td>all linguistic factors</td>
<td>Levenshtein distance, orthography, different syllable number, neighbourhood density, word frequency</td>
<td>$R = .50, R^2 = .24$</td>
</tr>
<tr>
<td>5th step</td>
<td></td>
<td></td>
<td>$p &lt; .001$</td>
</tr>
</tbody>
</table>

Table 2: Results of multiple regression analyses with the intelligibility scores as dependent variable and the linguistic factors as independent variables.
results in a correlation of .44. Adding different syllable number, model 3 reveals a correlation of .47, which rises to .49 in model 4 when neighbourhood density is integrated. Finally, model 5 includes word frequency with a correlation of .50. Even though word frequency as a separate variable correlates low with the intelligibility results, it is thus identified as a relevant factor in the prediction of word intelligibility in a combined model. The remaining factors (foreign sounds, word length, word accent differences, lexical tones, stød and etymology) do not add significantly to the model.

![Scatter plot of the multiple regression analysis in enter mode](image)

Figure 2: Scatter plot of the multiple regression analysis in enter mode

5. Discussion

Compared to earlier studies on linguistic predictors of intelligibility, the degree to which the intelligibility covaries with phonetic distances is low. Earlier studies showed high correlations between intelligibility scores and Levenshtein distance. Gooskens, e.g., obtained a correlation of $r = -0.80$ ($p < .001$) for intelligibility scores with Levenshtein distance between varieties of the Scandinavian languages Danish, Norwegian, and Swedish. Beijering et al. (2008) even found an overall correlation of $r = -0.86$ ($p < .01$) between intelligibility scores and Levenshtein distances of Copenhagen Danish and a range of other Scandinavian varieties. In our study, the correlation with Levenshtein distance is only $r = -0.33$ ($p < .001$). The five
factors combined in model 5 are revealed as most important for the intelligibility of Swedish words by Danes. Still, the correlation \( R = .50 \) and the percentage of explained variance with this combination of factors is only 24 \% \( (R^2 = .24) \). It remains to be discussed why the considered factors cannot explain more of the variance and which other factors are likely to play an additional role in intelligibility. In what follows, we will consider some possible explanations.

The main reason for the low correlation in the current study is that we focus on the intelligibility of single words rather than aggregate intelligibility. Whereas the aggregate intelligibility score – which is obtained as the mean of all single word scores in a whole corpus – may be consistent, the intelligibility of single words may be influenced by rather unpredictable factors such as prosodic differences (cf. voice quality, speech rate, etc.) and idiosyncratic characteristics of the single words.

In order to get an impression of such idiosyncratic characteristics we had a closer look at the mistakes that the listeners made. A number of different categories of mistakes can be distinguished. First, we found that many subjects confused the stimulus with (or searched for help in) another foreign language they had learned. Swedish *art* ‘sort’, e.g. was often translated into Danish *kunst* ‘art’, supposedly through confusion with English. Checking the corpus for words which are potentially confusable with English and German words could reveal an additional factor for intelligibility. Nevertheless, finding potential candidates for this kind of confusion is a hard task since the confusability is not always obvious.

Second, the mistakes give reason to believe that the way in which we operationalised the neighbourhood factor may not be optimal. It looks as if not the number of neighbours is decisive but rather how close the neighbours are to the test words. For example, the Swedish word *fel* [fe:l] ‘mistake’ was translated with Danish *fæl* [fæ’l] ‘foul’ by a majority of the listeners, probably due to the fact that *fæl* is phonetically closer to *fel* than the correct *fejl* [faj’l]. Examples like this suggest that qualitative characteristics of neighbours are more decisive in word identification than the total number of neighbours. Nevertheless, it is challenging to operationalise such a qualitative neighbourhood model, because the question of
how to measure similarity between sounds is difficult, particularly when addressed in two specific languages (cf. the following point).

Third, a number of sounds cause problems for the listeners because they are confused with non-corresponding phonemes in Danish. For example Swedish /a/ is always pronounced as [a] while in Danish it is mostly pronounced as [æ] except in combination with /r/. Therefore the Swedish stat [stɑːt] ‘state’ is translated as Danish start [sɑːst] ‘start’ instead of stat [sɑːt] by many listeners. Swedish /u/ is often confused with /ø/ or with /y/ which results in the translation of Swedish luft ‘air’ with Danish left ‘lift’ instead of the correct luft, cf. also Swedish frukt ‘fruit’ translated as Danish frygt ‘fear’ instead of the correct frugt. On the other hand, Doetjes and Gooskens (accepted) showed that Danes in general have no problems understanding words with a /u/ that is pronounced as [u] in Swedish and as [o] in Danish. This can probably be explained by the fact that the two sounds are so similar that the Danes think they hear a /o/, when a Swede pronounces a /u/. Disner (1983:59) showed that there is large phonetic overlap between Danish /ol/ and Swedish /u/.

Also some consonants are confused. Danish, e.g., has no voicing distinction but an aspiration-based distinction in plosives. Therefore, the Swedish difference between voiced and voiceless plosives corresponds to a difference between aspirated and unaspirated sounds at word onset in Danish. This is probably the reason why Swedish klass ‘class’ is translated as glas ‘glass’ instead of the correct klasse by the Danish listeners. These examples all show that the role of sound differences on the intelligibility differs per sound pair and probably is very language dependent. In order to model intelligibility more successfully, communicatively relevant sound distances therefore need to be incorporated into the Levenshtein algorithm.

The three kinds of mistakes discussed here give some indications of how we may proceed to improve the predictability of word intelligibility. However, it may turn out that there is a limit to the extent to which the model can be improved. Clearly some factors pertain to only a limited number of words and also the combination of factors plays a role. The listener may use different strategies for each word to match it with a word in his own language. Furthermore, such a model may have to be language dependent since each language combination provides different challenges to the listener.
The experiment may be found on the internet at http://www.let.rug.nl/lrs. It is possible to participate in the test with a guest account (login: germanic, password: guest). We thank Johan van der Geest for programming the experimental interface and databases.

The corpus provides texts from different written language genres (journals, magazines, fiction) from 1988–1992 and consists of 28 million words. The frequency lists are freely accessible at http://korpus.dsl.dk. To our knowledge, there is no corpus of comparable size available for spoken Danish.

References


L. Brink and J. Lund (1975), *Dansk rigsmål: lydudviklingen siden 1840 med særligt henblik på sociolekterne i København* (Copenhagen).


G. Doetjes and C. Gooskens (accepted). 'Skriftsprogets rolle i den dansk-svenske talesprogsforståelse', *Språk och Stil*.


E. Wessén (1965), Svensk språkhistoria. Ljudlära och ordböjningslära (Stockholm).
