LINGUISTIC DETERMINANTS OF THE INTELLIGIBILITY OF SWEDISH WORDS AMONG DANES

SEBASTIAN KÜRSCHNER, CHARLOTTE GOOSKENS AND RENÉE VAN BEZOOIJEN

Abstract *In the present investigation we aim to determine to which degree various linguistic factors contribute to the intelligibility of Swedish words among Danes. We correlated the results of an experiment on word intelligibility with eleven linguistic factors and carried out logistic 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 was found in the negative correlation between word intelligibility and phonetic distances. Also word length, different syllable numbers, foreign sounds, neighbourhood density, word frequency, orthography, and the absence of the prosodic phenomenon of 'stød' in Swedish contribute significantly to intelligibility. Although the results thus show that linguistic factors contribute to the intelligibility of single words, the amount of explained variance was not very large (*R^2 *(Cox and Snell)* $= .16$, R^2 *(Nagelkerke)* $= .21$*) when compared with earlier studies which were based on aggregate intelligibility. Partly, the lower scores result from the logistic regression model used. It was necessary to use logistic regression in our study because the intelligibility scores were coded in a binary variable. Additionally, we attribute the lower correlation 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.*

International Journal of Humanities and Arts Computing 2 (1–2) 2008, 83–100
DOI: 10.3366/E1753854809000329
© Edinburgh University Press and the Association for History and Computing 2009

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, Gooskens and Heeringa, 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 are important factors (e.g. Delsing and Lundin Åkesson, 2005). However, correlations between intelligibility scores and the latter two factors are low, and the direct relationship is difficult to prove (Van Bezooijen and 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 and textual context.[1] 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 of linguistic factors such as segmental and prosodic phonetic distance, Swedish sounds lacking in the Danish sound system, word frequency, and orthography on word intelligibility. 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 multilingual Europe.

2. EXPERIMENT

To test word intelligibility, an Internet-based experiment was conducted.[2] 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.[3] 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 from the city of Uppsala 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. Prizes 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 seven, the highest number 19, with an average of 10.5 subjects per word block.

52 secondary school pupils, all mother tongue speakers of Danish aged 16–19 who grew up with no additional mother tongue, 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 10 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 68 per cent of the subjects. This word is probably known to most Danes as a stereotypical Swedish word. For example, it was used in the popular Danish pop song *sköna flicka* ('beautiful girl') by Kim Larsen. On the basis of these results we decided not to exclude any of the 42 remaining 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 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. The obtained scores were subsequently used as the dependent variable in a regression model with several linguistic factors as covariates (see Section 3) to identify the degree to which these 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 borrowed from the same Latin source in both languages. We also excluded words that have a cognate root but a derivational morpheme that is different between the corresponding cognates in Swedish and Danish. So, for example, the word pair Swedish *undersökning* – Danish *undersøgelse* '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 (cf., e.g., Van Heuven, this volume). 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, aggregate phonetic distances between languages are good predictors of intelligibility of whole texts (cf. e.g. Beijering et al., 2008, Gooskens, 2007). 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.

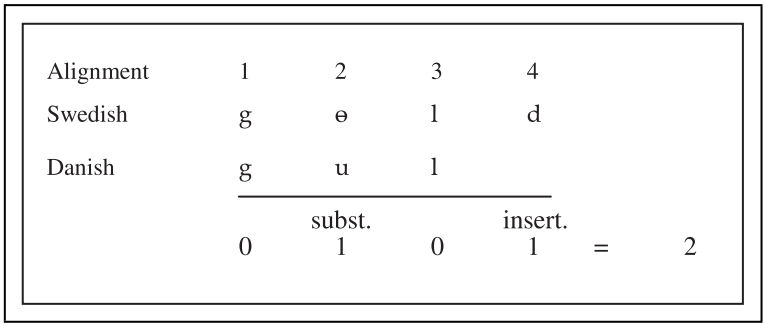

Figure 1. Calculation of Levenshtein distance.

Levenshtein distance is a measure of string edit distance based on the smallest number of operations necessary to map a given string on another string. Applied in linguistics, a string of sounds from one variety can thus be mapped 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 [gɵld] 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 transform one string into 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 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 by the number of alignments, i.e. 4 in the example. The normalised distance is $2/4 = 0.5$, i.e. 50 per cent for our example.

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.

The Swedish test words were transcribed by a phonetician who is a mother tongue speaker of Swedish, and the corresponding Danish words were represented by their pronunciation in Standard Danish. Levenshtein distance was calculated automatically for all 347 pairs of cognates. The distance was only calculated for segments, i.e. we did not include any prosodic features other than

segment length in our calculations. Instead, prosodic features are integrated in Sections 3.4, 3.6, and 3.7. The mean segmental distance across all words was 52.1 per cent. Eight word pairs had a distance of zero, for example Swedish *team* – Danish *team* 'team' that are both pronounced as [tiːm]. Six word pairs had the maximum distance of 100 per cent, for example Swedish *ljud* [jʉːd] Danish *lyd* [lyːˀð̩] 'sound'. For each word, Levenshtein distance was coded as a fraction representing a percentage. The distribution of the distances across the data proved to be normal.

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 (cf. Van Heuven, this volume). 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 as foreign in the literature (from Karker, 1997; Nordentoft, 1981):

- Retroflex consonants produced according to the phonological rule that [r] and a following alveolar consonant merge, cf. Swedish, *art* [ɑːʈ] 'sort', *bord* [buːɖ] 'table', *alternativ* [altəɳatiːv] 'alternative', *orsak* [uːʂɑːk] 'cause', *parlament* [paɭamɛnt]
- The postalveolar-velar fricative [ʃj], cf. Swedish *aktion* [akʃjuːn].

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. [lj] 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 vowels [ʉ] and [ʊ].

For each word, the number of foreign sounds was coded. 46 of the words contained a retroflex consonant or a postalveolar-velar fricative. Three words contained two foreign sounds: *koordination* [kuɔɖɪnaʃjuːn] 'coordination', *ordning* [oːɖɳɪŋ] 'order', and *stjärna* [ʃjæːɳa] 'star'.

3.3 Word length

Previous research has shown that word length plays a role in word recognition (Wiener and Miller, 1946, Scharpff and 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

(Vitevitch and Rodriguez, 2005). Furthermore, redundancy increases with word length, which is assumed to enhance intelligibility as well (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* [ɵp:mærksamheːt] 'attention', while the shortest word had only one segment, *ö* [øː] 'island'.

3.4 Word stress differences

Van Heuven (this volume) found that correct recognition of words was severely reduced and delayed if stress was shifted to the initial syllable in Dutch words with medial or final stress. Extrapolating this result he hypothesized that unexpected stress positions play a (negative) role not only in understanding the mother tongue but also a closely related variety (Van Heuven, this volume). For each Swedish word, we annotated whether the place of the word stress was different from that in the corresponding Danish cognate, assuming that such a difference makes the word more difficult to identify. The coding was categorical, either 1 when word stress was different, or 0 when this was not the case. Danish *kontekst* [ˈkɔntɛ̯gsd] vs. Swedish *kontext* [kɔnˈtɛkst] 'context' may serve as an example of word stress 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 *mængde* [mɛŋ'də] vs. Swedish *mängd* [mɛŋd] 'quantity'. Since a missing or extra syllable could cause confusion in word identification, we annotated instances with different syllable numbers by coding the number of additional or lacking syllables. Ten of the Swedish words contained one syllable extra compared to the corresponding Danish word, while 22 words had one syllable less. Two Swedish words were even two syllables shorter than the Danish cognate, namely *choklad* [ɧɔklɑːd] (Danish *chokolade* [ɕogolæːðə]) 'chocolate' and *tjänst* [ɕɛnst] (Danish *tjeneste* [tjɛ̯ːnəsdə]) '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 a closely related language bears similarities to listening to speech in noise

(cf. Van Heuven, this volume). 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. *ánden* (acc. I), definite singular form of *and* 'duck' vs. *ànden* (acc. II), definite singular form of *ande* 'spirit'. Accent I is most similar to the Danish stress accent, while there is no 'musical accent' comparable to the Swedish accent II in Danish. We therefore hypothesize that words with accent II may distract the Danish subjects when hearing such Swedish words. We coded the word accent for each test word, using a binary categorical variable. 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'ɐ] 'hands' versus [hɛnɐ] '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. We used a binary categorical variable to code for each word if it included a 'stød' or not. 164 words had a 'stød', and 161 words had 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 and 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 four 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 and coded it into the database. The mean number of neighbours was 0.93, which means that on average each correct answer has one competing incorrect answer. 244 words had no neighbours and the largest number of neighbours was 16 for the Swedish word *ö* 'island'.

3.9 Etymology

Work in progress by Gooskens, van Bezooijen and Kürschner 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. Additionally, on average loan words are longer than native words.[4]

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 could tell from etymological dictionaries such as Hellquist (1980) and Wessén (1960), 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. We used a binary categorical variable to code the etymology for each word.

3.10 Orthography

There is evidence that knowledge of orthography influences spoken word recognition (e.g. Ziegler and Ferrand, 1998; Chéreau, Gaskell and Dumay, 2007). The evidence comes from experiments with words that differ in degree of sound-to-spelling consistency and from recent neuroimaging research (Blau et al., 2008). Doetjes and 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 $r = .54$. 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 and Lund, 1975; Grønnum, 1998). As the spelling remained conservative, the number of sound-to-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 has even converged because people tend to pronounce the words as they are spelt (Wessén, 1965: 152; Birch-Jensen, 2007). Because of the Danish conservative orthography, it is 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 *hånd* [hɔnˀ] vs. Swedish *hand* [hand]. 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 number of such helpful letters was coded into the database for each word. The mean number of sounds per word 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 and 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 90.[5] The most frequent word was *dag* 'day', which occurred 222,159 times in the corpus. There were seven stimulus words which did not occur in the corpus and thus had a frequency of 0. The smallest positive frequency was found for *overføring* 'transmission', which occurred 11 times. Since the raw frequency data was heavily positively skewed, we changed the coding of this variable by recalculating it as log frequency. Based on log frequency, the data was normally distributed.

Table 1. Point-biserial correlation of the intelligibility scores with continuous linguistic factors.

Factor	Correlation (r)	Significance (p)
Levenshtein distance	$-.27$	$< .001$
Foreign sounds	$-.11$	$< .001$
Word length of Swedish words	.21	$< .001$
Difference in syllable number	$-.17$	$< .001$
Neighbourhood density	$-.13$	$< .001$
Orthography	.13	$< .001$
Log frequency	.01	not sign.

4. RESULTS

The intelligibility test resulted in an overall percentage of 61 per cent correct identifications of the Swedish cognates among the Danish subjects. To identify the independent contribution of each of the linguistic factors, we correlated the intelligibility scores with each factor separately. Since the intelligibility scores were coded in a binary variable, we had to calculate correlation coefficients considering this coding scheme. To correlate binary variables with continuous variables, point-biserial correlation coefficients using Pearson correlations are commonly used. We used this calculation for correlating the intelligibility scores with the continuous variables coding linguistic factors. The results are listed in Table 1.

The results show that apart from log frequency, all linguistic factors coded in continuous variables correlate significantly with the intelligibility scores. 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 much lower than in previous research which dealt with aggregate distance and intelligibility scores ($r = -.27$ compared to, e.g., $r = .86$ in Beijering, Gooskens and Heeringa, 2008). The correlations with word length ($r = .21$) and difference in syllable numbers ($r = -.17$) are comparatively high as well. Additionally, neighbourhood density ($r = -.13$), orthography ($r = .13$), and foreign sounds ($r = .11$) correlate significantly with the intelligibility scores.

In order to identify the independent contribution of the variables which were coded categorically, we conducted logistic regression analyses with only one covariate each. The results of these analyses are found in Table 2.

Table 2 shows that lexical tones and 'stød' do not explain the variance to a significant extent. By contrast, word stress difference and etymology are found to explain parts of the variance. Nevertheless, the amount of explained variance is low for both factors.

Table 2. Results from logistic regression analyses in enter method with the intelligibility scores as dependent variable and a single categorical linguistic factor as covariate.

	Step	Model χ^2	Significance	-2LL	Cox and Snell R^2	Nagelkerke R^2
	0			4921.00		
Word stress difference	1	6.432	$p < .05$	4914.57	.00	.00
Lexical tones	1	3.442	not sign.	4917.56	.00	.00
Stød	1	1.806	not sign.	4919.19	.00	.00
Etymology	1	23.787	$p < .001$	4897.21	.01	.01

We were not only interested in the contribution of each single factor, but we also wanted to find out which combination of factors served best to explain intelligibility. To identify which factor combination reveals the best prediction for the intelligibility scores, we conducted regression analyses with multiple factors. The intelligibility scores were defined as the dependent variable, and the eleven linguistic factors were chosen as covariates. Since the dependent variable was binary and thus did not meet the requirements of linear regression models, we used a generalized linear model in binary logistic regression. Table 3 summarizes the results of the regression analyses, conducted first with the enter method to identify the effect of all factors in combination and then with the forward method to identify the best stepwise combination of factors.

Table 3 summarizes the results of two linear regression analyses. The improvement of the model for each step and the significance of the improvement are calculated based on the χ^2 score. The -2 Log likelihood (-2LL) indicates how poorly the model fits the data: The more the value of -2LL is reduced in comparison to the beginning state or the previous step, the better the model fits, i.e. the higher is the contribution of the added factor. We report two 'pseudo' R^2 scores (Cox and Snell as well as Nagelkerke R-square) which – comparable with linear regression – serve to indicate the model's effect size. These R^2 scores are calculated cautiously and therefore seem rather low in comparison with R^2 scores from linear regression models. The analyses show that the linguistic factors can explain the variance partly, but not to a very high extent. Including all factors, we arrive at a χ^2 of 624.99 with R^2 (Cox and Snell) = .16, and R^2 (Nagelkerke) = .22.

The stepwise analysis, done in the forward method, reveals eight models. Levenshtein distance is revealed as the most important factor ($\chi^2 = 286.14$, R^2 (Cox and Snell) = .08, R^2 (Nagelkerke) = .10). The second model explains the variance to a higher extent by including word length ($\chi^2 = 461.05$, R^2 (Cox and Snell) = .12, R^2 (Nagelkerke) = .16). Steps 3 to 7 include different syllable number, foreign sounds, neighbourhood density, log frequency, and orthography.

Table 3. Results of binary logistic regression analyses with the intelligibility scores as dependent variable and all linguistic factors as independent variables.

Method	Model	Model χ^2	Significance	-2LL	Cox and Snell R^2	Nagel -kerke R^2
Enter						
Step 0				4810.03		
Step 1	all linguistic factors	624.99	$p < .001$	4185.04	.16	.22
Forward						
Step 0				4810.03		
1st step	Levenshtein distance	286.14	$p < .001$	4523.89	.08	.10
2nd step	previous factor +word length	461.05	$p < .001$	4348.98	.12	.16
3rd step	previous factors +different syllable no.	508.73	$p < .001$	4301.30	.13	.18
4th step	previous factors +foreign sounds	544.35	$p < .001$	4265.67	.14	.19
5th step	previous factors +neighbourhood density	567.85	$p < .001$	4242.18	.15	.20
6th step	previous factors +log frequency	593.03	$p < .001$	4217.00	.15	.21
7th step	previous factors +orthography	610.45	$p < .001$	4199.58	.16	.21
8th step	previous factors +'stød'	618.31	$p < .001$	4191.72	.16	.21

Finally, in step 8 'stød' is added, resulting in $\chi^2 = 618.31$, R^2 (Cox and Snell) $= .16$, and R^2 (Nagelkerke) $= .21$. Although log frequency as a separate variable has a low correlation with the intelligibility results, it is nonetheless identified as a relevant factor in the prediction of word intelligibility in a combined model. The same goes for 'stød', which did not explain the variance to a significant extent when used as the only covariate in a logistic regression model. The remaining factors (word stress difference, lexical tones, etymology) do not add significantly to the model, although both word stress difference and etymology were identified to explain parts of the variance significantly when used as the only covariate in logistic regression. Presumably, this might partly be attributed to the binary categorical coding scheme of both lexical tones and etymology.

Section 3 showed that the variables were not totally independent from each other but interact in certain dimensions. For example, word length presumably correlates negatively with frequency and the number of neighbours. This interaction might weaken the regression models. We therefore conducted a multicollinearity analysis. We calculated the variance inflation factors (VIF) of each of the predicting variables to see if the variables had a strong linear relationship with any of the other predictors. Since the mean VIF of 1.43 is higher than 1, we need to assume that our regression model is slightly biased by multicollinearity. The highest VIF is 2.46 (for word length). Since none of the variables thus reveals a VIF higher than 10, collinearity is not a serious problem for the model.

Collinearity diagnostics reveal that the strongest collinearity exists between word length, Levenshtein distance, and lexical tones. The reason for the collinearity between word length and Levenshtein distance is probably that Levenshtein distance increases with longer words. We tried to reduce this effect by normalizing by the length of the alignment, but the results of the diagnostics reveal that collinearity remains. Word length and lexical tones interact because accent II is impossible in monosyllabic words and thus only found in long words.

The existence of collinearity means that we cannot always precisely decide which of the interdependent factors makes the main contribution to explain the variance. A possible solution to this problem would be to reduce the number of factors, integrating covarying variables into one and the same variable. For example, lexical tones could be somehow integrated into the calculation of Levenshtein distance. Nevertheless, such a solution would cause new problems: When combining segmental and suprasegmental differences into a single measure, how would we know how to weigh the contribution of segmental and prosodic differences, respectively? Since we were mostly interested in tracing the contribution of each of the single linguistic factors and their combination in explaining the intelligibility of isolated words, and since the multicollinearity analysis showed that collinearity does not cause serious problems in our models, we conclude that the current analyses are thus well-suited to reveal models to answer our research questions.[6]

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 (2007), e.g., obtained a correlation of $r = -.80$, $r^2 = .64$ ($p < .001$) for intelligibility scores with Levenshtein distance between varieties of the Scandinavian languages Danish, Norwegian, and Swedish. Beijering, Gooskens and Heeringa (2008) even found an overall correlation of $r = -.86$,

$r^2 = .74$ ($p < .01$) between intelligibility scores and Levenshtein distances of Copenhagen Danish and a range of other Scandinavian varieties. In our study, the correlation with Levenshtein distance reveals lower scores, namely $r = -.27$, $r^2 = .07$ in point biserial correlations, and R^2 (Cox and Snell) $= .08$, R^2 (Nagelkerke) $= .10$ in logistic regression models. The eight factors combined in model 8 are revealed as most important for the intelligibility of Swedish words by Danes. In comparison with earlier studies, this model also reveals a rather low score of R^2 (Cox and Snell) $= .16$, R^2 (Nagelkerke) $= .21$. It remains to be discussed why the factors considered 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.

Partly, the low scores must be ascribed to the rather cautious calculation of R^2 scores in logistic regression modelling. In addition, the reason for the low correlation in the current study is probably that we focus on the intelligibility of single words rather than aggregate intelligibility. Aggregating is a mode of calculation which is known to inflate correlation coefficients because it reduces noise. 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', presumably 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 the number of neighbours is not as decisive as the proximity of the neighbours 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 /ɑ/ mostly corresponds with [ɛ] in Danish. Only in combination with /r/, it is pronounced as /ɑ/ in Danish, too. Therefore Swedish *stat* [stɑːt] 'state' is translated as Danish *start* [sdɑːˀd] 'start' instead of *stat* [sdɛːˀd] by many listeners. Swedish /ɵ/ is often confused with /ø/ or with /y/ which results in the translation of Swedish *luft* 'air' with Danish *løft* '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 in understanding words with an /u/ that is pronounced as [u] in Swedish and as [o] in Danish (cf., e.g., Swedish *fot* [fuːt] – Danish *fod* [foð̰]. This can probably be explained by the fact that the two sounds are so similar that the Danes think they hear an /o/, when a Swede pronounces an /u/. Disner (1983: 59) showed that there is large phonetic overlap between Danish /o/ 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 suggest that the effect of rather fine phonetic differences on the intelligibility is significant and probably 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.

REFERENCES

K. Beijering, C. Gooskens and W. Heeringa (2008), 'Modeling intelligibility and perceived linguistic distances by means of the Levenshtein algorithm', in M. Koppen and B. Botma, eds, *Linguistics in the Netherlands* (Amsterdam), 13–24.

R. van Bezooijen and C. Gooskens (2007), 'Interlingual text comprehension: linguistic and extralinguistic determinants', in J. D. ten Thije and L. Zeevaert, eds, *Receptive Multilingualism*

and intercultural communication: Linguistic analyses, language policies and didactic concepts (Amsterdam) 249–264.
J. Birch-Jensen (2007), *Från rista till chatta. Svenska språkets historia* (Gothenburg).
V. Blau, N. van Atteveldt, E. Formisano, R. Goebel, and L. Blomert (2008), 'Task-irrelevant visual letters interact with the processing of speech sounds in heteromodal and unimodal cortex', *European Journal of Neuroscience*, 28 (3): 500–509.
L. Brink and J. Lund (1975), *Dansk rigsmål: lydudviklingen siden 1840 med særligt henblik på sociolekterne i København* (Copenhagen).
C. Chéreau, M. G. Gaskell and N. Dumay (2007), 'Reading spoken words: Orthographic effects in auditory priming', *Cognition*, 102, 341–360.
L.-O. Delsing and K. Lundin Åkesson (2005), *Håller språket ihop Norden? En forskningsrapport om ungdomars förståelse av danska, svenska och norska* (Copenhagen).
S. F. Disner (1983), *Vowel quality: The relation between Universal and language specific factors* (Los Angeles).
G. Doetjes and C. Gooskens (accepted), 'Skriftsprogets rolle i den dansk-svenske talesprogsforståelse', *Språk och Stil*.
C. Gooskens (2007), 'The contribution of linguistic factors to the intelligibility of closely related languages', *Journal of Multilingual and multicultural development*, 28 (6), 445–467.
N. Grønnum (1998), *Fonetik og fonologi. Almen og dansk* (Copenhagen).
E. Haugen (1966), 'Semicommunication: The language gap in Scandinavia', *Sociological inquiry*, 36, 2, 280–297.
W. Heeringa (2004), *Measuring dialect pronunciation differences using Levenshtein distances* (Groningen).
E. Hellquist (1980 [1922]), *Svensk etymologisk ordbok*. Tredje upplagan. (Malmö). Also available at: http://runeberg.org/svetym/.
V. J. van Heuven (1985), 'Perception of stress pattern and word recognition: recognition of Dutch words with incorrect stress position', *Journal of the Acoustical Society of America*, 78, S21.
V. J. van Heuven (this volume), 'Making sense of strange sounds: (mutual) intelligibility of related language varieties. A review', *International journal of humanities and arts computing*.
A. Karker (1997), 'Det danske sprog', in A. Karker, B. Lindgren and S. Løland, eds, *Nordens språk* (Oslo).
P. A. Luce and D. B. Pisoni (1998), 'Recognizing spoken words: The Neighborhood Activation Model', *Ear and hearing*, 19, 1–36.
A. M. Nordentoft (1981), *Nordiske nabosprog: dansk-norsk-svensk sproglære for lærerstuderende* (Copenhagen).
P. J. Scharpff and V. J. van Heuven (1988), 'Effects of pause insertion on the intelligibility of low quality speech', in W. A. Ainsworth, J. N. Holmes, eds, *Proceedings of the 7th FASE/Speech-88 Symposium* (Edinburgh), 261–268.
M. S. Vitevitch and E. Rodriguez (2005), 'Neighborhood density effects in spoken word recognition in Spanish', *Journal of Multilingual Communication Disorders*, 3, 64–73.
E. Wessén (1960), *Våra ord. Deras uttal och ursprung. Kortfattad etymologisk ordbok* (Stockholm).
E. Wessén (1965), *Svensk språkhistoria. Ljudlära och ordböjningslära* (Stockholm).
F. M. Wiener and G. A. Miller (1946), 'Some characteristics of human speech', in *Transmission and reception of sounds under combat conditions. Summary Technical Report of Division 17, National Defense Research Committee* (Washington, DC), 58–68.
J. C. Ziegler and L. Ferrand (1998), 'Orthography shapes the perception of speech: The consistency effect in auditory word recognition', *Psychonomic Bulletin and Review*, 5, 683–689.

END NOTES

[1] The intelligibility of words without any possible influence from the semantic context can also be tested by presenting words in nonsense contexts, which might resemble usual language decoding more than identifying isolated words. However, since Swedish is characterised by a high number of sandhi phenomena, the correct segmentation of the test words would be an additional task when presented in syntactic context. By using words without any syntactic context we thus made sure that the subjects' task was only to identify words, without the additional need to segment them correctly.

[2] 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.

[3] The list was prepared for investigating several Germanic languages. It was based on the most frequent words occurring in large corpora of both formal language (Europarl, cf. http://www.statmt.org/europarl/) and informal language (Corpus of Spoken Dutch, cf. http://lands.let.kun.nl/cgn/home.htm).

[4] The mean word length of native Swedish words in our dataset was 4.8 sounds, while the mean length of loan words was 5.9 sounds. This difference was significant ($df = 345$, $p < .001$).

[5] 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.

[6] Still, there is a chance that the models are influenced by the fact that some of the variables are not well-balanced. For example, with lexical tones 253 words have accent I, and only 94 have accent two. Only 46 of the 347 words reveal foreign sounds, etc. A possible solution to this problem would be a Latin square design, but this is rather complicated and might even be impossible to build with eleven factors. We therefore chose to include all words into the analysis.