Research plan

The role of orthography for the mutual intelligibility of spoken Swedish and Danish

1. Background

Danish and Swedish are so closely related that Danes and Swedes often communicate each using their own languages, which Haugen (1966) dubbed semicommunication. However, different investigations (Maurud 1976, Bø 1978, Delsing & Lundin Åkesson 2005) have shown that Swedes often have more problems understanding spoken Danish while it is easier for Danes to understand Swedish. The investigations also show that it is easier to read the neighboring language than understanding the spoken form and that the asymmetric intelligibility between Swedes and Danes is only found when the languages are spoken and not when they are written.

A number of explanations for the asymmetric Swedish-Danish intelligibility of spoken language have been given. So far, most attention has been paid to extra-linguistic factors such as attitude and experience. However, no clear relationship has been found between attitude scores and results of intelligibility tests. As far as linguistic explanations are concerned, Gooskens (submitted, accepted) found high correlations between intelligibility scores and linguistic distances measures by Levenshtein distances and Moberg, Gooskens and Nerbonne (accepted) found asymmetric conditional entropies corresponding to the asymmetric intelligibility scores of past investigations.

In the present investigation we will focus on the different relationships between the written and the spoken languages in Sweden and Denmark. These differences are also often mentioned as an explanation for the asymmetric intelligibility [ref.] In late old Danish many sound changes took place and also the Danish spoken language has developed very rapidly during the last 50 years (Brink & ?? ????). As a result there is a large distance between spoken and written Danish. This difference is smaller in Swedish, where the spoken language has developed less rapidly. The orthographies in the two countries have changed less drastically and therefore written Swedish and Danish are much more similar than the spoken languages (see Gooskens and Doetjes submitted). [het zou interessant zijn als men Levenshteinafstanden zou kunnen gebruiken om de historische ontwikkeling van de gesproken en geschreven Zweeds en Deens te quantificeren, bijvoorbeeld met oude en nieuwe radioopnames en teksten].

The different relationships between the spoken and written languages in the two countries have resulted in different points of departures as far as the understanding of the spoken neighboring language is concerned. When a Dane hears a Swedish word he often has more support from the Danish spelling of the corresponding word than a Swede hearing the corresponding Danish word has from the Swedish spelling. This can be illustrated by the word pair Da. *hånd* - Sw. *hand* 'hand'. The pronunciation is different in the two languages [phonetische transcripties toevoegen]. The final consonant is not pronounced in Danish, but a Dane can decode the Swedish word because the final consonant is still written in Danish. On the other hand a Swede has less support from the Swedish orthography when he hears the Danish pronunciation without the final consonant.

Examples which point in the opposite direction can also be given. The Danish word *kirke* 'church', pronounced with [k] at the beginning of the word, is presumably easier for a Swede to understand than the corresponding Swedish word *kyrka* pronounced with [ʃ], is for a Dane. For a Dane it is unexpected that a *k* can be pronounced as [ʃ], but for a Swede it is not unusual that a written *k* corresponds with [k], for example in the word *katt* [kAtˈ] 'cat'.

Most linguists agree (ref.) that the advantage is larger for the Danes because it is the Danish language which has developed most rapidly, but this intuition has never been tested.

The first aim of the present investigation is to answer the following question:

1. Do Danes and Swedes use their own orthography in order to decode the spoken neighboring language?

In order to answer this question, a word decision experiment will be carried out. The results of the experiment will be correlated with different measurements (Levenshtein and conditional entropy, see Section 5) of the distances between the two languages. Different groups of Swedish and Danish listeners will be tested: illiterates and literates; experienced and inexperienced. This will make it possible to answer the following two subquestions of research question 1:

1a. Do listeners with L2 experience have greater advantage when decoding words with frequent and regular sound correspondences and/or grapheme-phoneme correspondences between L1 and L2 than inexperienced listeners?
1b. Do literate Danes and Swedes understand the L2 better than illiterates because of the support from the L1 orthography when decoding the L2?

If question 1 is answered positively, we would also like to answer the second research question:

2. Do Danish listeners use L1 orthography to a higher degree when decoding L2 spoken words than Swedish listeners?

In order to answer this question the listeners will also be asked to translate the test words. In this way we will know whether the listeners are also able to use the correspondences for the correct understanding of words in the neighboring language. In addition, distance measurements will be carried out on a large parallel Swedish-Danish corpus in order to get distance measures that are representative of the languages as a whole. If the Danish listeners show shorter reaction times and the overall distances between Danish orthography and Swedish pronunciation we can conclude that differences in orthographies are indeed (part of) the explanation for the asymmetric Danish-Swedish spoken intelligibility.

2. Stimulus words

The listeners will hear three kinds of L2 words: target words on which the analysis will be based, non-words and non-cognates.

Target words
Approximately 100 [?] target words will be tested in each country. These words will be selected on the basis of the following criteria:

1. Only frequently used words (lexemes) will be tested. In this way we make sure that all concepts are well known and that frequency can be assumed to play no role. The frequency of each word will be looked up in a Danish and a Swedish frequency

dictionary. Only words belonging to the 3000 most frequently used words in both languages will be included

2. The same words will be tested in both languages. All words belong to cognate word pairs with the same meaning in the two languages. 'Cognate' is used in the wide sense of the word. This means that both words belonging to the common Scandinavian heritage vocabulary as well as words which have been borrowed into both languages from other languages later in history will be included. [of beter om allleen erfwoorden en evt. Laag-Duitse leenwoorden te gebruiken om de correspondentieregels te beperken?]. Since we are only interested in sound correspondences, the whole words must be cognates. This means that cognates with historically unrelated suffixes should be excluded. This means that for example the word pair Da. *øv-else* vs. Sw. *öv-ning* 'practise' must be excluded. Alternatively, only monomorphematic words will be selected as test words.
3. Ideally we would like to select words which have the same number of other words with very similar orthographic/phonetic patterns in both languages. Such very similar word forms will provide the listeners with alternatives and make the chance of wrong identifications and longer processing time larger. It is however very difficult to make an exhaustive list of words with similar patterns since we are dealing with spoken as well as written language and since the alternatives are to be found in a language other than the stimulus language. In order to be sure that this factor has no influence on the results, a pre-experiment will be carried out. Swedish and Danish listeners will be asked to listen to the L2 words and translate them into their own language. Cognates which are translated with different numbers of alternatives by the listeners from the two countries will be excluded. [Hoe heet dit fenomeen? Is er geen betere manier om een selectie van cognates te maken die even veel alternatieven hebben?]
4. [Volgens mijn aantekeningen mogen woorden geen taalspecifieke fonemen bevatten. Waarom?].
5. The words will be selected in such a way that a maximal range of linguistic distances as measured by Levenshtein and conditional entropy (see Section 5) will be covered, both between the spoken forms and between written L1 and spoken L2.

   For some of the word pairs the distance between Swedish L1 and Danish L2 should be larger than between Danish L1 and Swedish L2 and for some words it should be the other way round. Likewise words with small distances between the spoken forms and large distances between written L1 and spoken L2 and vise versa should be included.
6. [meer?]

Non-words

Since the task of the listeners is to judge whether they hear an existing word or a non-word (see below), a number of non-words will be included [hoeveel? Hoe worden deze woorden gevonden?]

Non-cognates

Finally a number of non-cognates [20? Ook leenwoorden?] with different degrees of frequency will be included in order to measure the amount of previous experience of the listener with the L2 (see Section 3). Misschien ook en test om de kijken of luisteraars al sound correspondences kennen?

3. Listeners groups

Four different groups of listeners will be tested in each country, see Table 1.

Table 1. The four groups of listeners in each country.

	inexperienced	experienced
illiterates	N=20	N=20
literates	N=20	N=20

We will have experienced as well as inexperienced listeners in order to answer question 1a. whether experienced listeners have greater advantage when decoding words with frequent and regular sound correspondences than inexperienced listeners. Ideally we would select a group of inexperienced listeners who have never been confronted with the L2, but most Danes and Swedes hear or read the neighboring language every now and then via the media. We will therefore search for listeners who have had as little previous contact with the L2 as possible. The listeners will be asked questions about the amount of contact with the L2 in different situations. Furthermore they will be asked to translate a number of frequent non-cognates. The percentage of correctly translated non-cognates will be a measure of experience with the L2 since knowledge of non-cognates is a clear sign of experience.

The group of experienced listeners will be selected among listeners who indicate to have had regular contact with the L2 language and translate a large percentage of the non-cognates correctly. Different degrees of experience will be represented and we expect frequency and regularity to play a larger role with increasing experience. However, very experienced listeners and bilinguals may have a different mental organization of the two languages. For them the sound correspondences and orthography may play a smaller role or no role at all when decoding the test words, and therefore they will be excluded. Also listeners who learned the L2 during courses will be excluded since they may not have learned the L2 words via sound correspondences by learning vocabulary lists from books.

In addition to literates we will test groups of illiterates [dyslectici?]. Since these listeners cannot read they have no support from the L1 orthography when decoding spoken L2 words (question 1b.).[1] It may turn out to be difficult to find enough illiterate listeners who can be matched well with the literate listeners as far as age and intelligibility is concerned. For this reason it may be necessary to test a smaller number of subjects or to leave out one of the two groups of illiterates, or the illiterates from one country.

4. Tasks

The listeners will hear the L2 target words in random order mixed with the non-words and the non-cognates. Half of the subjects will hear the words in one random order and the other half will hear the words in the mirrored order in order to avoid too much influence of learning effects or fatigue. For each word, the listeners will perform two tasks:

1. lexical decision task (reaction time)
2. translation task

[1] An alternative would have been to test children who have not yet learned to read. Anja Schüppert prepares experiments with Danish and Swedish children at the moment.

In the lexical decision task the listeners have to decide whether the L2 word is an existing word in their own language by pressing the yes or the no button as quickly as possible. The reaction time is taken as a measure of how difficult it is to decode the word. For words that are linguistically very different from the corresponding word in the L1 (see Section 5) a longer reaction time is expected than for linguistically similar words. [meten we reactietijden van alle woorden, of slechts van correct vertaalde woorden?]

If the listeners press the yes button, they will be asked to translate the words to the corresponding words in their own language. The literates write down their answers and the answers by the illiterates will be recorded or written down by the research leader. The results of this part of the experiment can be used to select the correctly translated words. It gives us the possibility to see to which extent the listeners can use the orthographic/phonetic correspondences to understand L2 words.

5. Measurements

Two kinds of measurements will be carried out: Levenshtein distances and conditional entropies.

Levenshtein distances
The Levenshtein distance between corresponding words is based upon the minimum number of symbols (letters or phonetic symbols) that need to be inserted, deleted or substituted in order to transform the word in one language into the corresponding word in another language. The more operations are needed, the larger the distance. The distance between two words is symmetric: the distance from a word in language a to the word in language b is the same as the distance from language b to language a.

The cost assigned to each operation can be gradual, expressing for example that [b] is phonetically more similar to a [p] than to [l]. We will assign binary costs to the operations since we deal in part with letters and it is not clear in which way differences between letters and phonetic symbols are gradual. Also for the sake of comparability with conditional entropy binary costs seem a better choice.

The Levenshtein measurements between the spoken varieties are unproblematic. They follow well-described methods (Heeringa 2004). For the Levenshtein measurements from the written form to the spoken form, the method which has been developed by Gooskens and Doetjes (submitted) will be used.

Conditional entropy
Conditional entropy measures the entropy or uncertainty in a random variable when another is known. In our case, a listener hears a phoneme in the L2 and attempts to map it to a phoneme in the L1. We use conditional entropy to measure the uncertainty, and therefore difficulty of predicting a unit in the native language given a corresponding unit in the non-native language. In contrast with Levenshtein distance, conditional entropy can be asymmetric, i.e. it does not hold in general that the entropy of language a given language b is equal to the entropy of language b given language a. Since conditional entropy expresses the regularity and frequency of correspondences, the measurements have to be based on a large corpus. Normally, the entropy expresses the overall distance between two languages, but in our case we will measure conditional entropies between word pairs based on the frequencies and regularities as found in a whole database. We will use a database compiled by members of the VIDI-project. This is a parallel corpus of approximately 2500 word pairs transcribed orthographically and

phonetically in seven Germanic languages, including Danish and Swedish. For a more detailed explanation of conditional entropy see (Moberg, Gooskens and Nerbonne accepted).

Both the conditional entropies and the Levenshtein distances between written L1 and spoken L2 may be asymmetric between the two languages because both the written and the spoken form are different in the two languages. In addition conditional entropies can be asymmetric because the regularity and frequency of the correspondences may vary in both directions.

The Levenshtein distance between the spoken word pairs is symmetric and therefore only one measurement has to be made. Conditional entropies between the spoken varieties can be asymmetric because the regularity and frequency of the correspondences may vary in both directions.

In order to answer our research question seven linguistic distances will be measured per word pair:

1. Levenshtein distances from written Danish to spoken Swedish
2. Levenshtein distances from written Swedish to spoken Danish
3. Conditional entropies from written Danish to spoken Swedish
4. Conditional entropies from written Swedish to spoken Danish
5. Levenshtein distances between spoken Danish and spoken Swedish
6. Conditional entropies from spoken Danish to spoken Swedish
7. Conditional entropies from spoken Swedish to spoken Danish

Table 2. The seven linguistic distance measurements

distances	Levenshtein	conditional entropies
between written L1 and spoken L2	1 (Sw→Da) 2 (Da→Sw)	3 (Sw→Da) 4 (Da→SW)
between spoken L1 and spoken L2	*5 (Da↔Sw)	6 (Sw→Da) 7 (Da→Sw)

* = distance from Danish to Swedish is the same as from Swedish to Danish

6. Analysis

For each of the 8 listeners group (2 countries x 2 literacy groups x 2 experience groups) the results of the two dependent variables (reaction time [slechts voor correct vertaalde woorden?]) will be correlated with the relevant distance measurements from Table 2. [regressieanalyse?]. On the basis of these correlations the research questions will be answered.

1. Do Danes and Swedes use their own orthography in order to decode the spoken neighboring language?

Our first aim is to answer the question whether Danes and Swedes can use their L1 orthography to decode spoken L2 words. If this is indeed the case, we expect high correlations for the reaction times of the literates with the measures between written L1 and spoken L2 (distances 1 and 3 in Table 2 for the Swedish listeners; distances 2 and 4 for the Danish listeners). Also the correlation between the spoken representations (distances 5, 6 and 7) may be high since listeners can be expected to use information about their L1 orthography as well as their L1 pronunciation when decoding L2 words.

1a. Do listeners with L2 experience have greater advantage when decoding words with frequent and regular sound and grapheme/phoneme correspondences between L1 and L2 than inexperienced listeners?
In order to answer this question we compare the results of the correlations of measurements 3, 4, 6 and 7 with results of the group of inexperienced listeners to the results of the experienced listeners.

For the experienced listeners we expect the correlation with the test results to be stronger when distances are measured with entropies (distances 3, 4, 6, 7) than when measured with Levenshtein distance (distances 1, 2, 5), since experienced can profit from regularity and frequency of correspondences and entropies are able to model such correspondences. The effect will be stronger when the amount of experience is larger.

1b. Can literate Danes and Swedes use the L1 orthography when decoding the L2 while illiterates cannot?
Illiterate listeners have no help from the L1 orthography when listening to spoken L2. For illiterate listeners we expect higher correlation with measures between spoken L1 and L2 (distances 5, 6 and 7) than between written L1 and spoken L2 (distances 1, 2, 3, 4). If this is indeed the case, it will be a confirmation of the finding that literate listeners use the L1 orthography when decoding spoken L2.

Our hypotheses are summarized in Table 3.

Table 3. Highest correlations expected with reaction times

	Listeners:			
	experienced		non-experienced	
	illiterates	literates	illiterates	literates
measurements:	spoken→spoken	spoken→written	spoken→spoken	written→spoken
	entropy	entropy	Levenshtein	Levenshtein

Research question 2: can Danish listeners to a higher degree use L1 orthography when decoding L2 spoken words than Swedish listeners?
If we are indeed able to show that literate listeners use their L1 orthography when decoding spoken L2 we also want to investigate whether the asymmetric intelligibility as found in earlier investigations can be explained by differences in the relationship between orthography and pronunciation in the two languages.

We expect Danes to be better able to use L1 orthography when decoding spoken L2 than the Swedes and therefore we expect the correlations with distances 2 and 4 (the distances for the Danes) to be larger than distances 1 and 3 (Swedes). Furthermore we expect the differences between literates and illiterates to be larger for the Danes than for the Swedes.

However, the reaction time does not show us whether the listeners were actually able to understand the test word. Therefore we should also look at the reaction time of the correctly translated words. If the correlation is still higher for the Danes then we know that they are indeed better able to use the correspondences to decode the test words.

Still, the test words are not a random selection of words but have been selected on the basis of a number of criteria (see Section 2). Therefore no strong conclusions about overall linguistic distances between the languages can be drawn. The seven linguistic distance measures listed

in Section 5 should also be carried out on a larger corpus representing the languages as a whole. We will use the corpus mentioned in Section 5. If we are able to show that literate listeners use L1 orthography when decoding L2 words and that the distances between written L1 and spoken L2 is larger for Danes than for Swedes, then we have evidence for the claim that (part of) the asymmetric intelligibility of spoken Danish and Swedish can be explained by the different relationships between orthography and pronunciation in the two languages.