4

Conditional Entropy Measures Intelligibility among Related Languages

Jens Moberg[†], *Charlotte Gooskens*[†], *John Nerbonne*[†], *and Nathan Vaillette*[‡]
[†]University of Groningen
[‡]Dickinson College, Pennsylvania

Abstract

The Scandinavian languages are so alike that their speakers often communicate, each using their own language, which Haugen (1966) dubbed SEMICOMMUNICATION. The success of semi-communication depends on the languages involved, and, moreover, can be asymmetric: for example, Swedish is more easily understandable for a Dane, than Danish for a Swede. It has been argued that non-linguistic factors could explain intelligibility, including its asymmetry. Gooskens (2006), however, found a high correlation between linguistic distance and intelligibility. This suggests that we need to seek linguistic factors that influence intelligibility, and that potentially asymmetric factors would be particularly interesting. Gooskens' distance techniques cannot capture asymmetry. The present paper attempts to develop a model of the success of semi-communication based on conditional entropy, in particular using the conditional entropy of the phoneme mapping in corresponding (cognate) words. Semantically corresponding words were taken from frequency lists and aligned, and the conditional entropy of the phoneme mapping in aligned word pairs was calculated. This gives us information about the difficulty of predicting a phoneme in a native language given

Proceedings of the 17th Meeting of Computational Linguistics in the Netherlands
Edited by: Peter Dirix, Ineke Schuurman, Vincent Vandeghinste, and Frank Van Eynde.

a corresponding phoneme in the foreign language. We also examine the conditional entropy of selected word classes, such as native/loan and function/content words.

4.1 Introduction

The three mainland Scandinavian languages (Danish, Norwegian and Swedish) constitute an interesting linguistic community with respect to mutual intelligibility. They are so closely related that they are sometimes considered dialects of a common, non-existent, language (Maurud 1976, Braunmüller 2002). This linguistic situation enables citizens in Scandinavia to use their native tongues when communicating with their neighbors. Haugen (1966) coined the term SEMICOMMUNICATION for this phenomenon, for which Braunmüller (2002) suggests rather RECEPTIVE MULTILINGUALISM.

4.1.1 Background

It has been noted that semicommunication may be difficult, and several studies, the most prominent being Maurud (1976), Bø (1978), and Delsing and Lundin Åkesson (2005), were carried out in order to investigate how well speakers of the three languages understand the neighboring languages. We calculated the mean percentage of correct answers in the intelligibility tests of these three investigations, and display these per language pair in Figure 4.1. The largest problems are found in the mutual intelligibility between Swedes and Danes. Swedes especially have difficulties understanding Danish (a mean of 27% correct answers as opposed to 37% correct when Danes attempt to understand Swedish). Norwegians understand the neighboring languages best, while Danes and Swedes both have more difficulties understanding Norwegian.

Intelligibility is asymmetric in all of the language pairs in Fig. 4.1, and intelligibility scores are often explained by appeals to attitude and amount of contact. A positive attitude should encourage subjects to try to understand the language in question, whereas a negative attitude will discourage subjects from making an effort. Contact with the language in its written or spoken form is also likely to improve the performance on the test. The good performance by the Norwegians may be explained by the fact that the language variety of the listeners (eastern Norwegians) is linguistically close to both Danish and Swedish, Furthermore, it has been proposed that Norwegians are particularly good at understanding closely related language varieites because the Norwegian dialects are used so extensively. In contrast to many European countries dialects are used by people of all ages and social backgrounds in Norway, not only in the private domain but also in official contexts (Omdal 1995). For this reason Norwegians are used to decoding different language varieties. The influence of this factor on semicommunication has, however, never been tested experimentally. The three Scandinavian studies mentioned above included questions about attitude towards and contact with the test language. The authors assume a relationship between the non-linguistic factors (attitude and experience) and the intelligibility scores, but correlations are low

and the direct relationship is difficult to prove. A third factor, linguistic structure, has been largely neglected so far, mostly due to the absence of a suitable method to measure differences in linguistic structure. In recent years, new methods have been developed for measuring linguistic differences in the area of dialectometry. This makes it possible to measure communicatively relevant linguistic differences among the spoken Scandinavian languages. Linguistic differences can be measured at various linguistic levels, but we shall be concerned exclusively with the phonetic level in this paper.

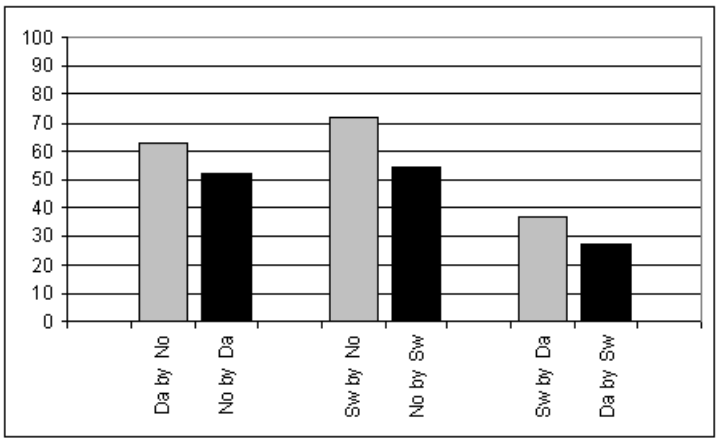

Figure 4.1: Mean percentage correct answers of three spoken intelligibility tests (Maurud 1976, Bø 1978, Delsing & Åkesson 2005). 'Da by No' stands for 'Percentage correct in Danish test by Norwegians', etc.

Heeringa (2004, Chap. 7–8) describes a method for measuring the phonetic distance between dialects and closely related languages by means of the Levenshtein algorithm. This algorithm calculates the minimum cost of transforming one sequence of phonemes to another. Gooskens (2006) used these distance measurements, and found a high correlation between intelligibility and phonetic similarity measured by means of Levenshtein distances ($r = 0.82$, $p \leq 0.01$). However, since the Levenshtein algorithm calculates distances, which are axiomatically symmetric, it cannot provide an account of asymmetric relations in linguistic intelligibility.

4.1.2 Present Paper

The present paper explores the linguistic differences among the Scandinavian languages by means of another measure, conditional entropy, which we apply at the phonemic level. Conditional entropy measures the complexity of a mapping, and is sensitive to the frequency and regularity of sound correspondences between two

languages. Since these two factors could be important to the ease with which a word in one language is understood by speakers of a related language, we hypothesize that conditional entropy corresponds with intelligibility scores. We are motivated to explore conditional entropy because it can model asymmetric remoteness. The conditional entropy between language A and language B is not necessarily the same as between language B and language A. If the asymmetric intelligibility scores reported above (see Figure 4.1) reflect the difficulty of mapping one sound system to another, we may expect conditional entropies to operationalize this difficulty, so that high entropies correspond with low intelligibility between a given pair of Scandinavian languages, and low entropies with high intelligibility. The primary purpose of this paper is to test this hypothesis.

We based our measurements on a database with frequent words in the three languages. This database was divided into different categories which made it possible to test three hypotheses. First, we expected native words to produce a higher conditional entropy between pairs of languages than loan words, since they have evolved in the respective languages for a long time. Loan words entering a language are expected to differ less because they have been borrowed in a similar form and have not had the time to diverge as much.

Second, we expected lower entropies for Latin/Greek/French loan words than for German loan words because the time of borrowing differs. Most German loan words came into Scandinavian in the twelfth and thirteenth century, during the Hanseatic period. French loan words became popular in the sixteenth century (Edlund and Hene 1992). Words imported into the Scandinavian languages were often adapted in some way during the process. Assume that in a borrowed word, sound A becomes sound B in Swedish, sound C in Danish and sound D in Norway. The way that the Scandinavian languages transform this sound to fit their own language is to a certain degree a regular process, meaning that the pairwise relations (between B and C, etc.) are rule governed. However, since the German loan words have been part of the Scandinavian languages for a longer time, they have had more time to change, which means that the regularities may have attenuated. The fact that French, Latin and Greek are less closely related to the Scandinavian languages than German might also mean that the words have been less well integrated into the Scandinavian languages than German words. For this reason the Latin/Greek/French words may to a greater extent have kept their original pronunciation. This might cause lower conditional entropies for Latin/Greek/French loan words than for German loan words.

Third, we make a distinction between function words and content words. Many function words are very frequent, and since they are less essential to the semantic content of sentences, they often occur in unstressed positions. For this reason their form may have been more strongly reduced than that of content words. In addition, very frequent words are also said to be phonologically conservative, i.e. they resist regular changes. Both observations lead us to expect conditional entropies to be higher for function words than for general vocabulary, since in both case they may represent exceptions to rules.

To summarize, our specific research questions are as follows.

1. Do high conditional entropies correspond to low intelligibility scores as found in the literature and *vice versa* (see Figure 4.1)?
2. Can asymmetric mutual intelligibility be modeled by conditional entropies?
3. Is there a difference in conditional entropies between native words and loan words?
4. Is there a difference in conditional entropies between Latin/Greek/French loan words and German loan words?
5. Is there a difference in conditional entropies between content words and function words?

4.2 Conditional entropy

Conditional entropy (CE) measures the entropy, or uncertainty in a random variable when another is known. In the case we have in mind, an interlocutor hears a phoneme in a non-native language and attempts to map it to a phoneme in his own. The conditioning variable is the phoneme heard in the non-native language, and the conditioned variable is the phoneme to be identified.

Conditional entropy is calculated with the following formula:

$$H(X|Y) = -\sum_{x \in X, y \in Y} p(x,y) \log_2 p(x|y) \tag{4.1}$$

As the formula clarifies, CE is always calculated on the basis of the conditional probability of one variable given another.

$H(X|Y)$ is the uncertainty in X given knowledge of Y, i.e. how much entropy remains in X if the value of the variable Y is known. We use CE to measure the uncertainty, and therefore difficulty of predicting a unit in the native language given a corresponding unit in the non-native language.

We note that CE is asymmetric, i.e. it does not hold in general that $H(X|Y) = H(Y|X)$. This means that it will not run into the same conceptual difficulties as the distances used by Gooskens (2006).

4.2.1 Plausibility

As a simplest illustration of how conditional entropy can be used for linguistic units, consider the following. Written Danish words have only one vowel in their grammatical endings, the letter *e*, while Swedish uses *e*, *a* and *o*. This means that a Swedish speaker that encounters the Danish letter *e* has three options when trying to find the equivalent Swedish phoneme. Idealizing now to the situation where this were the only use of the sounds in question, we can see that a Danish speaker, upon encountering Swedish *e*, *a* or *o*, can know that the proper correspondence is *e*. The entropy is therefore higher for Swedish given Danish in this example, and the relationship is asymmetric.

Table 4.1: Corpus of Two Phonetically Transcribed Word Pairs

Danish	Swedish
j ɑ i	j ɑː g
l ɑ ŋ ʔ	l ɔ ŋ #

4.2.2 Example: CE for 2 Danish-Swedish Word Pairs

If the imaginary example of the perfect three-way split in the mapping serves to motivate the idea of using the complexity of the mapping as a model for intelligibility, it suffers from being too simple and from not taking frequency into account. It is too simple in that it is seldom, if ever, the case that one sound is mapping into three (or $n \neq 0$) others, each of which participates in no other mapping. And the measure of complexity intuitively ought to involve frequency—we can also understand more easily if we have a reasonable "guess" about the correspondence, and that guess may be well informed by frequency.

We shore up this intuition using a slightly larger example, with sound segments from two aligned word pairs to calculate the aggregate conditional entropy.

Table 4.1 shows a made-up corpus containing two word pairs with a total of 13 occurrences of sound segments. The sound segments are aligned, mimicking the way a non-native interlocutor might attempt to map a foreign word to one in his own language: /j/ with /j/, /ɑ/ with /ɑː/, /i/ with /g/ and so forth. In the last word pair, Danish glottal stop is aligned with a filler symbol. The frequencies are used to estimate the probabilities needed to calculate conditional entropy (4.1), including $P(d)$, the chance of segment d occurring in Danish; $P(s)$, the chance of s in Swedish; $P(d|s)$, the chance of d in alignment, given s; $P(s|d)$, the converse; and $P(d, s)$, the chance of d and s occurring jointly (in alignment). $P(d, s)$ is used to weight the importance of the conditional probabilities $P(d|s)$ and $P(s|d)$ in the CE formula (4.1).

We illustrate how the conditional entropies would be calculated on the basis of a corpus using the data of Table 4.1 by keeping track of the alignments, including the partial alignments. We thus first align all of the data, obtaining the alignments shown in Table 4.2, which we now discuss.

In the second cell alignment in Table 4.2, Swedish /ɑː/ is matched with Danish /ɑ/. Swedish /ɑː/ occurs only once, so that $P(\text{ɑ}_D|\text{ɑː}_S)$ is therefore 1. Since $-\log 1 = 0$, this contributes nothing to entropy. In the other direction, $P(\text{ɑː}_S|\text{ɑ}_D) = 0.5$: Danish /ɑ/ corresponds to Swedish /ɑː/ in the second word pair and to Swedish /ɔ/ in the second word pair (cell 5). This type of correspondence is the cause of asymmetry in the phoneme mapping complexity: the uncertainty is higher for Swedish speakers because they have more sound segments to choose from than Danish speakers.

All the Swedish segments map uniquely to Danish counterparts so that $\forall s \in$

Table 4.2: Seven Illustrative Segment Alignments and Corresponding Frequencies. From aligned data (as shown), we extract the relative frequencies of the correspondences. The 1:1 frequencies indicate perfect correspondences, therefore conditional probabilities of 1, which correspond to zero contributions to entropy ($-\log_2 1 = 0$). Note that all of the relative frequencies marked with 'S' are perfect (1:1), so that $H(\text{Danish}|\text{Swedish}) = 0$, reflecting the perfect predictability of the Swedish → Danish mapping. 'D(1:2)' in the top row center (cell 2) indicates e.g. that, Danish /ɑ/ is realized in the way indicated in the cell (ɑː) once out of a total of two occurrences (the other is in the bottom row, second position, cell 5). The boldfaced asymmetric alignment frequencies (in cells 2 and 5) contribute to the entropy difference, $H(D|S) = 0.0 < H(S|D) = 0.28$ (in this example set).

Language	1	2	3	
D→	j	ɑ	i	
S→	j	ɑː	g	
	S(1:1), D(1:1)	S(1:1), **D(1:2)**	S(1:1), D(1:1)	
	4	5	6	
D→	l	ɑ	ŋ	ʔ
S→	l	ɔ	ŋ	#
	S(1:1), D(1:1)	S(1:1), **D(1:2)**	S(1:1), D(1:1)	S(1:1), D(1:1)

$Sp(d|s) = 1, -\log_2 p(d|s) = 0$, and the total entropy is zero, corresponding to the perfectly certain mapping. Similarly, five Danish segments map uniquely to Swedish segments, likewise contributing zero to entropy. But one Danish segment /ɑ/ is mapping 50% of the time to Swedish /ɑː/ and 50% of the time to Swedish /ɔ/. We therefore estimate that $p(\text{ɑː}|\text{ɑ}) = p(\text{ɔ}|\text{ɑ}) = 0.5$, and therefore that $-\log_2 p(\text{ɑː}|\text{ɑ}) = -\log_2 p(\text{ɔ}|\text{ɑ}) = 1$, and we use $p(\text{ɑː,ɑ}) = p(\text{ɔ,ɑ}) = 1/7 \approx 0.14$ to weight these contributions to entropy, obtaining $H(\text{Swedish}|\text{Danish}) = 2 \times 0.14 = 0.28$.

Based on the mini-corpus in Table 4.2 $H(S|D) > H(D|S)$ because of the larger number of less certain mappings (in this case the second and fifth elements of the alignments, just discussed). We hypothesize that this is true in general, and that it contributes to the lesser intelligibility of Danish for Swedes.

We turned out to need about 800 words to obtain stable estimations of phoneme mapping entropies, but smaller samples consistently gave good estimations of the relative difference between $H(\text{Lg}_1|\text{Lg}_2)$ and $H(\text{Lg}_2|\text{Lg}_1)$. See Fig. 4.2.

4.3 Material

In order to conduct the entropy measurements, a database containing word lists from the three Scandinavian languages was constructed. The database also contained the same lists in Dutch, Low German, High German and Frisian since we plan to extent our research to these languages as well. The database contains the

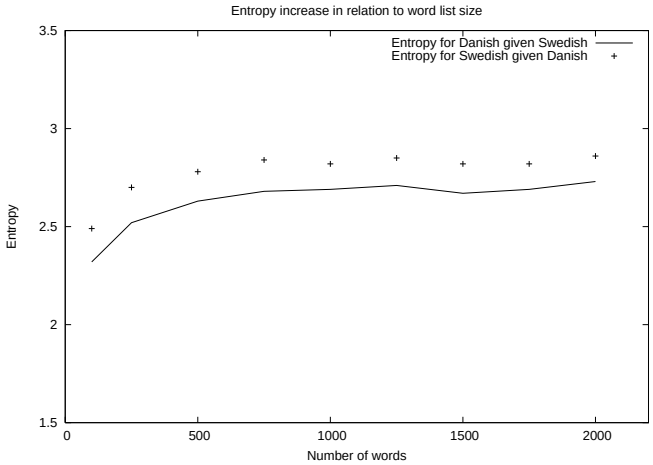

Figure 4.2: Entropy in relation to word list size

most frequent words from two corpora, Corpus Gesproken Nederlands, CGN, and Europarl[1].

CGN is a Dutch corpus of contemporary Dutch as spoken by adults in Flanders and the Netherlands that was collected between 1998 and 2004. This part of the CGN contains a total of 2,626,172 tokens. From this corpus we extracted the 1,500 most frequent words from the category that contained informal speech (the Face-To-Face dialogues).

Europarl is a speech corpus that consists of extracts from meetings held within the European Parliament. They are characterized by monologues by different speakers, including the chairman of the meeting. Europarl is translated into eleven European languages. Our motivation for choosing to extract the Dutch and Swedish version was twofold: firstly, these two languages represent the two Germanic branches of the language tree, West Germanic and North Germanic, that the database was intended to reflect. Secondly, the two languages were part-of-speech tagged, in contrast to for example Danish. We selected the 1,500 most frequent Dutch words from the Europarl, from a total of 889,836 tokens, and the 1500 most frequent Swedish words, from a total of 1,032,144 words. Next the Dutch and the Swedish lists were matched to find the 1,500 most frequent words that are common in the two lists.

The CGN list and the Europarl list were joined, and doublets were removed. The database was later supplemented with function words collected from grammar books. The goal was to make the collection of function words as comprehensive

[1] `http://lands.let.kun.nl/cgn/home.htm` and `http://www.statmt.org/europarl`, both accessed Dec 14, 2006.

as possible. Proper nouns and interjections were removed.

We based the word lists on formal as well as informal speech in order to check for differences regarding the number of loan words for these two categories. Recall that we expect words of common Northern Germanic origin to have been in the individual languages longest, followed by the words borrowed from German, followed by late borrowings from Latin, Greek and French. Recall, too, that the oldest words have the most time to undergo language-specific changes, meaning that they will be less parallel, and contribute therefore more to conditional entropies. In order to be able to investigate this hypothesis we had to ensure that a sufficient number of native words and loan words from different languages were present in our word lists. Europarl contains many words from the domains of politics and economy. These words are often borrowed from French, Latin or Greek (Gooskens et al. submitted, 2007). CGN's informal speech is collected from everyday speech situations where we expect more native words.

The material was translated so that we got word lists of the same words in the seven Germanic languages. All words were transcribed according to standardized speech as in pronunciation dictionaries, but we made no effort to verify that the standard pronunciation was in fact used in the utterances. We also looked up all the words in etymological dictionaries in order to establish from which language the loan words have been borrowed. The final version of the database contains the following information per word and language:

- The corpus from which the word was collected
- Word class
- Function word/content word
- The origin of the word (loan word or native word)
- If the word is a loan word, the language from which the word has been borrowed
- The lexical representation of the word
- The phonetic representation of the word
- Cognate/non-cognate[2] (if a word from one language is a cognate with a word from another language, these words are effectively coindexed at this feature in the database)

On the basis of this information we divided the database into a number of categories, facilitating the calculation of conditional entropies for given subvocabularies of language pairs. These categories, and their sizes in each of the three languages, are shown in Figure 4.3. The grammatical division into function

[2]The term COGNATE is usually reserved for (native) words from different languages that have descended from a common ancestor. We use an extended sense of this term to include as well words in different languages that have a common source via borrowing.

words/content words was only done for the native words since almost all function words are found in that category.

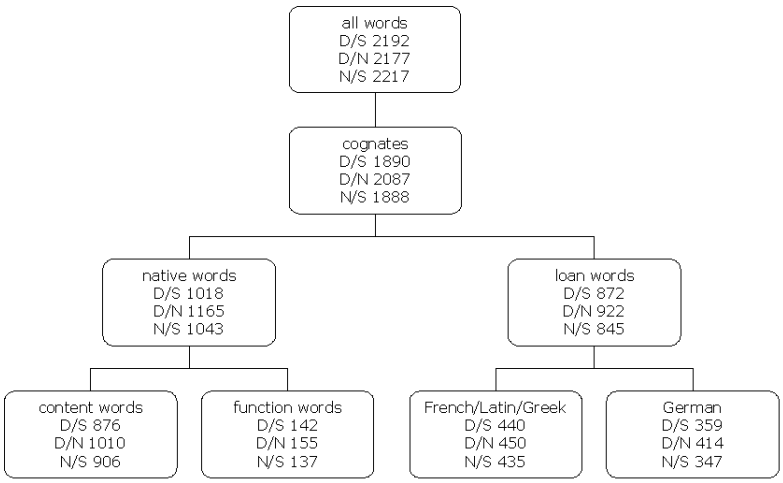

Figure 4.3: Number of Word Pairs for each of the 8 Categories.

4.4 Results

In this section we present the conditional entropies for each of the language pairs measured in both directions and compare them to the mean results of intelligibility tests presented in Figure 4.1. We look at entropies based on the entire word lists as well as subgroups containing different sub-vocabularies (see Section 4.3). To repeat, a low conditional entropy value $H(\text{Native}|\text{Foreign})$ means that mapping from the foreign language to a given native language is relatively simple: correspondences are regular and frequent. Therefore a low conditional entropy is hypothesized to correspond to high intelligibility. On the other hand, a high entropy value means a high level of uncertainty for the listener and a low level of intelligibility.

4.4.1 Danish/Swedish

Since the results of intelligibility tests show that Danes understand Swedes better than vice versa (see Figure 4.1), we expect $H(D|S) < H(S|D)$, i.e. it is less complex to map from Swedish to Danish than it is to map from Danish to Swedish.

Figure 4.4 shows the entropy per category and the divergence from symmetry. The X axis shows the entropy for Danish given Swedish, i.e. the difficulty for a

Swede to predict the Swedish equivalent of a given Danish sound segment. The Y axis shows the entropy for Swedish given Danish. The diagonal represents the completely symmetric situation, where the entropy is the same in both directions. Symbols above the line are categories which have a higher entropy for Swedish given Danish (more difficult for a Swede) while symbols under the line have a higher entropy for Danish given Swedish (more difficult for a Dane).

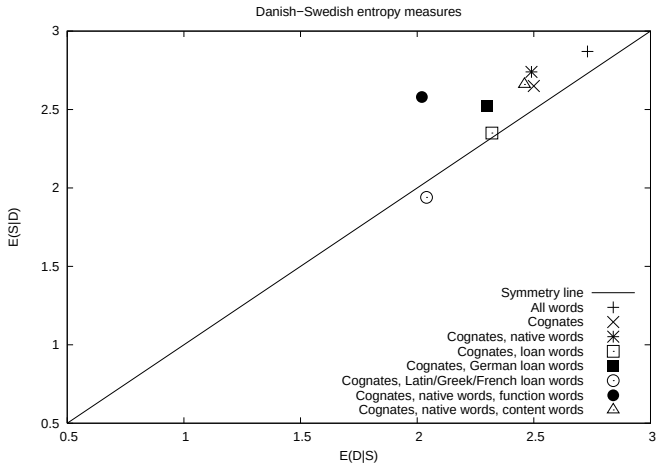

Figure 4.4: Conditional Entropy between Swedish and Danish, noting Asymmetry.

The first conclusion one can draw from these results is that there really is a difference in entropy between Swedish and Danish depending on the direction. For all categories of words, except for Latin/Greek/French cognates, the entropies are higher for Swedes listening to Danes than the other way round. This is what we would expect from the intelligibility tests (Figure 4.1).

As expected, the category consisting of all words has the highest entropy. This can be explained by the fact that this category is the biggest (see Fig. 4.2), and that it contains cognates as well as non-cognates that have no regular sound correspondences. But the group containing only cognates also has high entropy. This could be expected because it contains words of different origin, native as well as loan words, with different sound correspondences. However, when comparing native cognates to cognate loan words, we see that the native words have a higher entropy. These words have had more time to diverge than the loan words and this results in less regular sound correspondences. We see that category containing Latin/Greek/French loan words indeed have lower entropies than the German loan words. This confirms our expectation (see Section 4.1) that the correspondences for these words are more regular due to their later time of borrowing and low level of integration into the languages. Assuming that the words had more or less the same appearance in Danish and Swedish at the time of the borrowing, they have

had little time to diverge along with the respective pronunciation schemas in these countries, which in turn means more regular correspondence and lower entropy.

Contrary to our expectations, the function words have lower entropies than content words. It is possible that this can be explained by the fact that this group consists of so few words in comparison with the other groups (see Fig. 4.2).

4.4.2 Norwegian/Swedish

Figure 4.5 shows the entropy and asymmetry for the Norwegian/Swedish language pair. From the results in Figure 4.1 we expect lower overall entropies than for Swedish/Danish and we also expect the entropies to be higher for Swedish listeners than for Norwegian listeners. Both expectations are fulfilled. The Swedish/Danish entropies for the entire sample ranged between 1.94 (non-Germanic borrowings) to 2.87 (overall) bits while the corresponding Swedish/Norwegian entropies are lower, between 0.87 (non-Germanic borrowings) and 2.28 (overall). In each category of word tested, we found higher entropies and therefore more complex mappings in the Swedish/Danish case than in Swedish/Norwegian. Turning to the second expectation, it also turns out that the Swedish to Norwegian mapping is simpler than the reverse, not only overall, but also in all subcategories of words we examined (with the single exception of the category of non-Germanic loan words, where the Norwegian to Swedish mapping was slightly simpler (0.05 bits)). Among Scandinavian cognates, German borrowings, function words and content words we find lower entropies for the Swedish to Norwegian mapping.

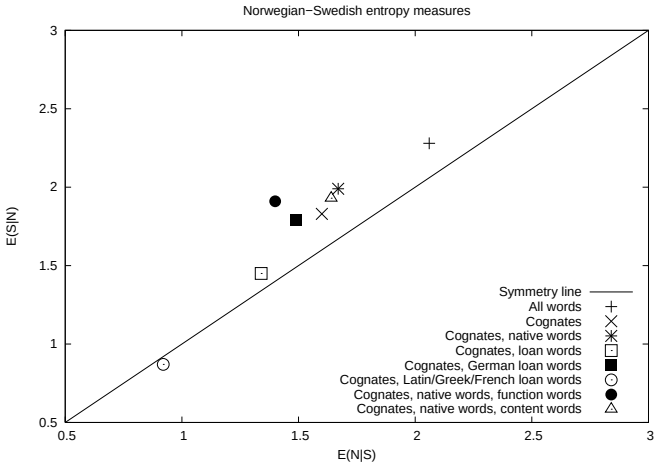

Figure 4.5: Conditional Entropy between Norwegian and Swedish

4.4.3 Danish/Norwegian

Figure 4.6 shows the entropy and asymmetry for the Danish/Norwegian language pair. The overall entropies are higher than for Swedish/Danish but lower than for Swedish/Norwegian. This corresponds with the results of the intelligibility experiments in Figure 4.1. The range between the different categories is not very large (values between 1.90 and 2.46). This can probably be explained by the fact that Danish and Norwegian have a long common history and the east Norwegian variety which represents standard Norwegian in the phonetic transcriptions has had particularly strong influence from Danish until a hundred years ago. This means that the languages were still one language when the loan words were introduced into the languages. This goes for Latin/Greek/French as well as for German loan words and therefore the entropies of these two categories are almost the same. Also the entropies of the cognate native words and the loan words are rather close. Almost all categories are close to the symmetry line. This seems to suggest that the asymmetric mutual intelligibility found in Figure 4.1 can only to a small extent be explained by differences in entropy. We will return to this in Section 4.5.

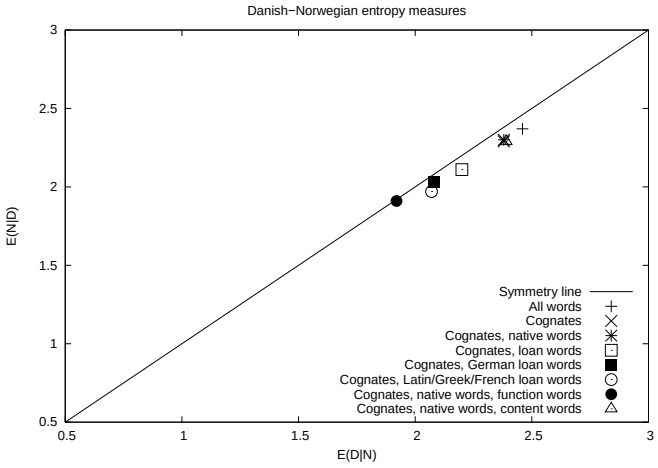

Figure 4.6: Conditional Entropy between Danish and Norwegian

4.5 Conclusions and discussion

The purpose of the present investigation was to explore conditional entropy as a linguistic measure for modeling the mutual intelligibility between closely related languages. Such a measure should also be able to model asymmetric intelligibility between for example Swedish and Danish. In Figure 4.7 we present a scatterplot which shows the relation between scores on intelligibility tests as found in the lit-

erature about semicommunication in Scandinavia (see Figure 4.1) and conditional entropies based on all words, cognates and non-cognates (circles) and on cognates only (triangular shapes). This figure clearly suggests that conditional entropies correspond well with the results of intelligibility tests. The relationship is clearest when all words are included. When the listeners were tested in the intelligibility tests they were also confronted with all words. The measurements based on cognates express pure phonetic measures of difference. Here the relationship with intelligibility scores is less clear, especially due to the fact that the two Norwegian-Danish intelligibility measures are higher than could be expected from the phonetic distances (Fig. 4.7). This might be explained by the small number of non-cognates between Danish and Norwegian (recall Figure 4.3).

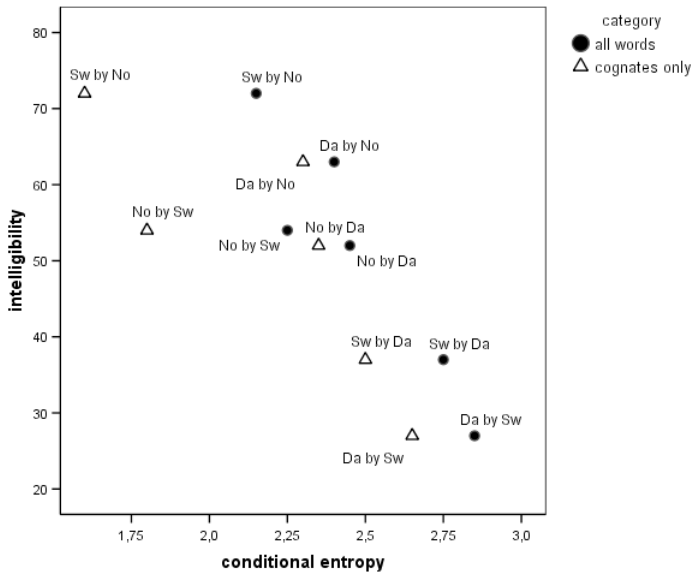

Figure 4.7: Entropy in Relation to Intelligibility

An important motivation for using conditional entropy as a measure of remoteness was that this measure is able to model asymmetric intelligibility. Asymmetric intelligibility is found between all Scandinavian language pairs (see Figure 4.1). This asymmetry was clearly reflected in the conditional entropies of Swedish-Danish (see Figure 4.4) and Swedish-Norwegian (Figure 4.5), but only to a small degree for Danish-Norwegian (Fig. 4.6). As mentioned in Section 4.1, the fact that Norwegians are better at understanding the neighboring languages than Danes and Swedes is mostly explained by the special Norwegian language situation that trains the Norwegians to understand different language varieties. The asymmetric intelligibility is larger for Swedish-Norwegian than for Danish-Norwegian. So maybe

for Swedish-Norwegian, the asymmetry is caused by a combination of language experience and linguistic factors while for Danish-Norwegian linguistic factors play only a minor role.

For all language pairs we found lower entropies for loan words than for native words. We explained this by the fact that loan words have had less time to diverge than native words which has resulted in more regular sound correspondences in the loan words. This explanation is supported by the fact that German loan words have higher entropies than the more recently borrowed Latin/Greek/French loan words. The fact that loan words have lower entropies than native words seem to suggest that a large number of loan words may benefit the intelligibility between the Scandinavian languages. This is an interesting prospect. The worry of linguistic deterioration as a consequence of too many loan words might be toned down if it turns out that loan words favor mutual intelligibility. The idea of having a common Scandinavian policy for acceptance of loan words could also find support in this result.

In future research we will refine the entropy model in several ways. More sophisticated measures will be developed that are able to express the fact that for example consonants are more important for decoding cognates than vowels and that not all phonotactic positions are of equal importance for understanding. The onset is clearly the most important position at least within the Germanic language family. We will also experiment with measurements based on bigrams or trigrams. Mutual intelligibility in Scandinavia is well documented so that the Scandinavian languages formed a good point of departure for our measurement. Our corpus contains more Germanic languages and we will apply our measurement to these languages as well. On this note, we have recently begun a collaboration with colleagues in Nijmegen and Leuven on comprehensibility among various Dutch varieties in the Netherlands and Flanders. Furthermore, we will collect material from other languages pairs which are known to have asymmetric mutual intelligibility, for example Spanish-Portuguese. At present we conclude only that there seems to be a relationship between entropy and the intelligibility experiments reported in the literature. To be more certain we need to conduct intelligibility experiments testing the hypotheses under controlled circumstances.

Acknowledgments

We are grateful to the Volkswagen Foundation, whom we thank for their support of "Measuring Linguistic Unity and Diversity", P.I. Erhard Hinrichs, Tübingen, and also the Dutch Organization for Scientific Research, who fund "Linguistic Determinants of Mutual Intelligibility in Scandinavia", P.I. Charlotte Gooskens. We are also grateful to colleagues T. Zastrow of Tübingen; P. Osenova, K. Simov, V. Zhobov of Sofia; J. Prokić of Groningen; and to two anonymous CLIN referees for discussion and suggestions.

References

Bø, I.(1978), Ungdom og naboland. En undersøkelse av skolens og fjernsynets betydning for nabospråksforståelsen, *Technical Report 4*, Stavanger.

Braunmüller, K.(2002), Semicommunication and accommodation: Observations from the linguistic situation in Scandinavia, *International Journal of Applied Linguistics*.

Delsing, L. and Lundin Åkesson, K.(2005), *Håller språket ihop Norden?*, Nordiska ministerrådet, Copenhagen, Denmark.

Edlund, L. and Hene, B.(1992), *Lånord i svenskan - om språkförändringar i tid och rum*, Förlags AB Wiken, Umeå and Stockholm, Sweden.

Gooskens, C.(2006), Linguistic and extra-linguistic predictors of Inter-Scandinavian intelligibility, *in* J. van de Weijer and B. Los (eds), *Linguistics in the Netherlands*, Vol. 24, John Benjamins, Amsterdam, pp. 101–113.

Gooskens, C., van Bezooijen, R. and Kürschner, S.(submitted, 2007), The lexical profiles of Dutch and Swedish. A contrastive study, *in* B. Los and M. van Koppen (eds), *Linguistics in the Netherlands*, Vol. 23, John Benjamins, Amsterdam.

Haugen, E.(1966), Semicommunication: the language gap in Scandinavia, *Sociological Inquiry* **36**, 280–297.

Heeringa, W.(2004), *Measuring Dialect Pronunciation Differences using Levenshtein Distance*, PhD thesis, University of Groningen.

Maurud, Ø.(1976), Reciprocal comprehension of neighbour languages in Scandinavia, *Scandinavian Journal of Educational Research* **20**, 46–52.

Omdal, H.(1995), Attitudes toward spoken and written Norwegian, *International Journal of the Sociology of Language* **115**, 85–106.