PREDICTING THE ASYMMETRIC INTELLIGIBILITY BETWEEN SPOKEN DANISH AND SWEDISH USING CONDITIONAL ENTROPY

Felicity Frinsel, Anne Kingma, Charlotte Gooskens, Femke Swarte
Department of Applied Linguistics, University of Groningen
c.s.gooskens@rug.nl

Abstract
The languages Danish and Swedish are so similar to each other that they are partially mutually intelligible. Speakers of these languages can communicate with each other each using their own language. The intelligibility between the spoken languages is asymmetrical, however: Danes can understand spoken Swedish better than Swedes can understand spoken Danish. This asymmetry is absent in the written language. In this article, we explore whether this asymmetry can be explained by conditional entropy. Conditional entropy is a way of measuring the amount of regularity in the sound or grapheme correspondences between the two languages. This study is based on the assumption that a high amount of irregularity among the correspondences will impair intelligibility, and conversely, a high amount of regularity will aid intelligibility. The entropy measure can be asymmetrical: the entropy of Swedish from a Danish perspective is not necessarily the same as the entropy of Danish from a Swedish perspective. We calculated the entropies between the two languages for both written and spoken language and compared the results to intelligibility data gathered through a word translation task. In agreement with previous research, the word translation task showed a higher intelligibility of spoken Swedish for Danes than vice versa, and only a very small asymmetry for the written language. The results of the entropy calculations followed the same pattern. In the spoken language, there was a higher entropy of Danish for Swedes than vice versa, but in the written language, there was virtually no asymmetry. This leads us to conclude that entropy is a promising predictor for intelligibility and should be further explored in this context.

Keywords
Receptive multilingualism; Danish; Swedish; asymmetry; conditional entropy; mutual intelligibility.

INTRODUCTION
Research on intelligibility amongst the Scandinavian languages has a long history (Schüppert, 2011). The Scandinavian languages are so alike that their speakers often communicate with each using their own language. Haugen (1966), one of the first to study this phenomenon, originally called this *semi-communication*. However, this communicative mode has also been conceptualized as among others 'intelligibility of closely related languages' (e.g. Gooskens, 2006), 'the Swiss model' or 'plurilingual communication' (Lüdi, 2007), 'inter-comprehension'(Conti & Grin, 2008), 'Lingua Receptiva (LaRa)' (Rehbein, ten Thije & Verschik, 2012; ten Thije, 2013) and

'receptive multilingualism' (Braunmüller, 2002; ten Thije & Zeevaert, 2007; Zeevaert, 2007). In this study, we will use the term *receptive multilingualism* to refer to this way of communicating.

The success of receptive multilingualism depends among others on the languages involved. In the past, intelligibility was often explained by means of non-linguistic factors such as attitude and the amount of contact (Haugen, 1966; Maurud, 1976; Bø, 1978). The first factor, attitude, refers to a listener's opinion of or feeling towards a certain language or language variety. If a listener dislikes a language and/or its speakers, he will probably be less willing to put effort into trying to understand it. This might result in less successful communication. If the listener likes the language variety he is listening to, however, he might understand more, just by trying harder. Schüppert, Hilton and Gooskens (2015), for example, find a low but significant positive correlation (r = .19) between attitude and word intelligibility for Danish and Swedish. The second factor, contact, concerns previous experience the listener has with the speaker's language and other, possibly related, languages. Having learned to speak a language naturally improves a person's ability to understand it. But even if there has only been passive contact, for example by hearing radio programs or hearing tourists speak, the listener might start to recognize certain sound correspondences. In addition, knowing other languages can aid understanding, for example by providing vocabulary. If a native speaker of Dutch encounters Danish for the first time, having experience with a related language like German will improve his understanding of this new language (Swarte, Schüppert & Gooskens, accepted). The Danish word *kartoffel* 'potato', for instance, does not have a cognate in Dutch (the Dutch translation is *aardappel*). A Dutch reader who is familiar with German, however, will immediately recognize the similarity to the German word *Kartoffel* with the same meaning as the Danish word and translate it to Dutch correctly.

In recent years, the focus of intelligibility research has moved away from attitude and contact towards purely linguistic factors that could influence intelligibility (e.g. Gooskens, 2007; Beijering, Gooskens & Heeringa, 2008; Kürschner, Gooskens & Van Bezooijen, 2008). These studies used various ways of measuring linguistic distance or similarity between the languages. On finding occurs very consistently across all of this research: mutual intelligibility of spoken Danish and Swedish is asymmetric (see e.g. Schüppert, 2011). Whereas Swedes and Danes can understand each other's languages on the written level equally well, when it comes to spoken language, Danes perform much better with Swedish than Swedes with Danish.

Most linguistic measures used in previous studies, however, measure the distance between two languages in a symmetric way: there is one absolute distance between Swedish and Danish, regardless of which language is the speaker's language and which language is the listener's language. In this study, we explored the linguistic differences among Danish and Swedish by means of a linguistic measure introduced by Moberg et al. (2007): *Conditional Entropy*, which measures the complexity of a mapping, and is sensitive to the frequency and regularity of sound correspondences between two languages. Swedish /ɛ/, for example, very often corresponds to /e/ in Danish: compare Swedish *sätta* /sɛt:a/ with Danish *sætte* /sedə/ 'to put'. As Danish

and Swedish are related langauges, many such correspondences exist between them, some more regular than others. We aim to explore whether this regulatiry, measured using conditional entropy, might help explain the asymmetry between Danish and Swedish, because entropy is inherently asymmetrical. However, as opposed to Moberg et al. (2007), who have calculated the entropy between these languages on the phonetic level only, we will calculate the entropy between these languages on both the phonetic and orthographic level and compare these to results of spoken and written intelligibility tests, based on the same corpus. Our research questions are as follows.

1. Can asymmetric mutual intelligibility between spoken Swedish and spoken Danish be predicted by conditional entropy?

2. Can the lack of asymmetric mutual intelligibility between written Swedish and written Danish be predicted by conditional entropy?

The next section describes previous research on mutual intelligibility in Scandinavia. The third section explains conditional entropy in more detail and the fourth section describes how we computed our results. In the fifth section, we describe these results. In the final section the results will be summarized and discussed.

BACKGROUND

Receptive multilingualism among the mainland Scandinavian languages Danish, Swedish and Norwegian has been studied extensively in the second half of the past century. Haugen (1966)[48] was the first to investigate this matter by asking people from Norway, Sweden and Denmark how well they could understand the neighboring Scandinavian languages. The lowest degrees of intelligibility were reported for the combination of Danish and Swedish, in both directions. Maurud (1976) experimentally tested how well people in the capital cities of each country understood the other languages and confirmed Haugen's finding. In addition, he found an asymmetry between spoken Danish and Swedish: Danes could understand Swedish better than Swedes could understand Danish. This asymmetry was absent in the written intelligibility. It was suggested that his finding was due to differences in the amount of contact the Danish and Swedish participants had had to each other's languages: Denmark's capital, Copenhagen, is much closer to Sweden than Sweden's capital, Stockholm, is to Denmark (Schüppert, 2011). Later studies that controlled for this factor, however, found the same asymmetry between Danish and Swedish in spoken language and no or little asymmetry in written language (Bø, 1978; Börestam Uhlmann, 1991; Delsing & Lundin Åkesson, 2005).

Another factor to which the asymmetry between Danish and Swedish has been ascribed, other than the amount of previous language contact, is the attitude of the

[48] Published in 1966, but reporting research carried out in the early 50s.

listener to the other language. It has been repeatedly concluded that Danes in general are more positive towards Swedish than Swedes towards Danish, which could explain the fact that Danes understand Swedish better than Swedes understand Danish (Delsing & Lundin Åkesson, 2005; Gooskens, 2006; Schüppert, Hilton & Gooskens, 2015). It has generally been assumed that Danes are more positive towards Swedish than Swedes towards Danish because of extra-linguistic factors such as imposed norms and social connotations. Schüppert, Hilton and Gooskens (2015), however, show that German and Chinese listeners, who are not influenced by these social factors, have similar attitudes towards Danish and Swedish as the Danes and Swedes themselves: the non-Scandinavian listeners too judged Swedish more positively than Danish. This suggests the cause of the asymmetric attitudes is not social or cultural, but linguistic.

Properties of the languages themselves are, after contact and attitude described above, the third possible cause for the asymmetric intelligibility between Danish and Swedish (Gooskens, 2007). Schüppert (2011) hypothesized, based on the low correlations of intelligibility with contact and attitude found by Gooskens (2006) with data collected by Delsing and Lundin Åkesson (2005), that these linguistic factors are in fact the main cause of the asymmetry. The asymmetry establishes itself only in spoken intelligibility, not in the written language. This suggests that the asymmetry is caused by something inherent to spoken language. In an experiment with Danish and Swedish preschoolers, where any influence of previous contact and attitude was ruled out, Schüppert did however not find any asymmetry in spoken intelligibility between the two groups (Schüppert & Gooskens, 2012). This suggests that non-linguistic factors do determine intelligibility. Yet follow-up experiments with adults showed very low correlations between intelligibility and attitude (Schüppert & Gooskens, 2011; Schüppert, Hilton & Gooskens, 2015). Eventually, Schüppert (2011) concludes that the asymmetry is due to linguistic factors which are not relevant at a pre-school age, and hypothesizes the most important of these factors is literacy (Schüppert et al., submitted). Spoken Swedish is more similar to written Danish than spoken Danish is to written Swedish (Doetjes & Gooskens, 2009). This means that a Dane listening to spoken Swedish can use his knowledge of Danish spelling to understand the Swedish words, whereas a Swede listening to spoken Danish is not helped much by Swedish spelling. Schüppert et al. (submitted; see also Schüppert (2011), Chapter 6) confirm in an EEG experiment that literate Danes use their knowledge of Danish orthography when listening to Swedish. Preschoolers are not yet literate, therefore the asymmetry does not arise at that age.

Some of the studies investigating the influence of linguistic factors on intelligibility focused on properties specific to Danish and Swedish and the differences between these languages. Gooskens and Kürschner (2010), for example, investigated the influence of Danish *stød* (a kind of creaky voice) and Swedish tonal accents on intelligibility. Gooskens and Van Bezooijen (2013) performed a detailed error analysis on the results of an intelligibility experiment to determine which exact phonemes in Danish and Swedish caused problems with intelligibility. Other studies focused on more general measures of linguistic distance or similarity which can be applied to any combination of languages, such as articulation rates (Hilton,

Schüppert & Gooskens, 2011), lexical distance (Gooskens, 2007; Gooskens, Heeringa & Beijering, 2008) and phonetic distance (Gooskens, 2007; Beijering, Gooskens & Heeringa, 2008; Gooskens, Heeringa & Beijering, 2008). The current study follows the latter approach.

Heeringa (2004, Chapter 7-8) describes a method for measuring the phonetic and orthographic 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 and/or graphemes to another by counting the number of insertions, deletions and substitutions necessary and dividing them by the word length. The example below (1) shows the Levenshtein distance between the orthographic version of the Danish word *virkelig* and the Swedish *verkligen*, 'really'.

(1)	DA	v	i	r	k	e	l	i	g		
	SW	v	e	r	k		l	i	g	e	n
		0	1	0	0	1	0	0	0	1	1

Distance: 4/10 = 0.40

In (1) the optimal alignment takes up 10 positions. In the transition from one word to the other, there is one substitution (*e* for *i*), one deletion (*e*) and two insertions (*e* and *n*). This means the cost of transforming one word into the other, and therefore the distance between the two words, is 4/10 = .40. Gooskens (2006) used these distance measurements, and found a high correlation between intelligibility and phonetic similarity measured by means of Levenshtein distances ($r = .82$, $p < .01$). However, since the Levenshtein algorithm calculates distances, which are axiomatically symmetric, it cannot prove an account of asymmetric relations in linguistic intelligibility.

Moberg et al. (2007) aimed to solve this problem by using another method, *conditional entropy* (CE), which measures the complexity of a mapping, and is sensitive to the frequency and regularity of sound correspondences between two languages. In other words, Moberg et al. attempt to explain the asymmetrical intelligibility by measuring the amount of entropy in a language combination. As such, it is not a measure of distance per se – it does not measure how similar the two parts of a correspondence are, but simply how predictable the correspondence is in a certain language pair. Given a certain sound (or character) in language A, how predictable is the corresponding sound (or character) in language B? The more predictable this sound is, the lower the entropy. Higher predictability aids intelligibility, therefore the hypothesis is that a low entropy measure correspond to a high intelligibility score. This is indeed what Moberg et al. found for Danish, Swedish and Norwegian. One of the strengths of the entropy measurement is the fact that it can be asymmetrical: the conditional entropy between language A and language B is not necessarily the same as between language B and language A. This is an advantage compared to the Levenshtein distance, which in its basic form is completely symmetrical.

CONDITIONAL ENTROPY

Conditional entropy is calculated with the formula shown in (2). *H*(X|Y) is the entropy of X given Y, that is, the amount of uncertainty regarding the value of X when the value of Y is known. In the case of languages, the conditioning variable Y is the phoneme heard or grapheme read in the non-native language, so Y is the stimulus language (the value of which is known, it is the text the participant is reading or hearing). The conditioned variable X is the phoneme or grapheme to be identified, so X is the reader's native language (the value of which is unknown: the reader is trying to guess which values in his native language correspond to what he is reading or hearing in language Y). *p*(x,y) is the chance that a certain combination of x and y occurs and *p*(x|y) is the chance of the occurrence of x in the case of y. The units of x and y can be anything, but in the case of this study, they represent graphemes or phonemes.

(2)

$$H(X|Y) = -\sum p(x,y) \log_2 p(x|y)$$

(Moberg et al. 2007, p.4)

Table 1: Corpus of two phonetically transcribed word pairs

Danish			**Swedish**		
Phonetically transcribed	Orthographic representation	Translation (English)	Phonetically transcribed	Orthographic representation	Translation (English)
eŋən	ingen	'no'	ɪŋːən	ingen	'no'
kɔmə	komme	'come'	kɔmːa	komma	'come'

Example: CE for two phonetically transcribed Danish-Swedish word pairs

Table 1 shows two phonetically transcribed word pairs with a total of 16 occurrences of sound segments.

The sound segments are aligned, as illustrated in Table 2, such as the way a non-native speaker might attempt to map a foreign word to one in his own language: /e/ with / ɪ /, /ŋ/ with /ŋː/, /ə/ with /ə/ and so forth. For the purposes of this example, the length markers are ignored.

Table 2: Alignment of the two word pairs in Table1, D=Danish and S=Swedish

Language	1	2	3	4
D → S →	e ɪ S(1:1), D(1:1)	ŋ ŋ: S(1:1), D(1:1)	ə ə S(1:1), **D(1:2)**	n n S(1:1), D(1:1)
	5	6	7	8
D → S →	k k S(1:1), D(1:1)	ɔ ɔ S(1:1), D(1:1)	m m: S(1:1), D(1:1)	ə a S(1:1), **D(1:2)**

The frequencies are used to estimate the probabilities needed to calculate conditional entropy according to the formula in (2). For most phoneme pairs, every occurrence of a phoneme in one language is mapped to exactly one phoneme in the other language, making the correspondence completely predictable and the entropy for that pair 0. For example, in the third cell alignment in Table 2, Swedish /ə/ is matched with Danish /ə/. Swedish /ə/ occurs only as a match with the Danish counterpart /ə/, so that p(əD|əS) is therefore 1. Since − log 1 = 0, the entropy for that pair is zero and, therefore, the match is completely predictable from a Danish perspective. Note that the phonemes do not need to be exactly the same. In alignment 1 in Table 2, the Danish phoneme /e/ is mapped to the Swedish phoneme /ɪ/. These are two different phonemes, but both phonemes occur only in this combination. The entropy of the mapping is therefore 0, both from a Swedish and from a Danish perspective.

As shown in Table 2, all Swedish segments map uniquely to Danish counterparts, which indicates the perfect predictability of the Swedish → Danish mapping and H(Danish|Swedish) is therefore 0. However, in the other direction, Danish /ə/ corresponds to Swedish /ə/ in the first word pair (cell 3) and to Swedish /a/ in the second word pair (cell 8). This means that p(əS|əD) = 0.5. Filling in the rest of the formula in (2) yields H(Swedish|Danish) = 0.25. Based on this, it can be stated that in this example the uncertainty is higher for Swedish speakers because they have more sound segments to choose from than Danish speakers and therefore this type of correspondence is the cause of asymmetry in the phoneme mapping complexity.

Example: CE on the orthographic level for two Danish-Swedish word pairs

For this example, the entropy will be calculated for the two word pairs shown in Table 3.

Table 3: Two orthographic word pairs in Danish and Swedish.

Danish		**Swedish**	
Orthographic representation	Translation (English)	Orthographic representation	Translation (English)
ved	‘at’	Vid	‘at’
gøre	‘do’	göra	‘do’

Table 4 shows the alignment of these two word pairs character by character. The two words together produce seven character pairs. As in the previous phonetic example, most of these pairs form a 1:1 correspondence from both directions, as specified below each pair. This means that each letter in one language occurs only as corresponding to one particular letter in the other language. In these cases, $p(x|y)$ is 1, the $\log_2$ of which is 0, which means the entropy for these alignments is 0.

Table 4: Alignment of the two word pairs in Table 3, D=Danish and S=Swedish.

Language	1	2	3	
D →	v	e	d	
S →	v	i	d	
	S(1:1), D(1:1)	S(1:1), **D(1:2)**	S (1:1), D(1:1)	
	4	5	6	7
D →	g	ø	r	e
S →	g	ö	r	a
	S(1:1), D(1:1)	S(1:1), D(1:1)	S(1:1), D(1:1)	S(1:1), **D(1:2)**

There are two exceptions, however: alignment 2 in the first word and alignment 7 in the second word. From a Danish perspective, everything is fine. Reading the Swedish words, the Dane sees an <i> corresponding to an <e> at all times (alignment 2), and an <a> corresponding to an <e> at all times (alignment 7). *H*(Danish|Swedish) is therefore 0. The other way around, however, is more complicated. A Swede reading the Danish words finds an <e> twice (in alignment 2 and alignment 7), which in 50% of the cases corresponds, in his own language, to an <i>, and in 50% of the cases to an <a>.Therefore, $p(x|y)$ is 0.5 for each of these alignments. Filling in the rest of the formula in (2) yields *H*(Swedish|Danish) = 0.29. This result is asymmetric since *H*(Swedish|Danish) > *H*(Danish|Swedish). In other words, a Swede reading Danish has to deal with a higher amount of entropy than a Dane reading Swedish, for these two word pairs. In the given examples, Danish for Swedish readers and listeners has a larger entropy, however, of course there are also examples where Swedish for Danish readers and listeners has a larger entropy.

According to Moberg et al. (2007), at least 800 words are needed to reach stable entropy measures, but calculations based on less words already show the relative differences among language pairs accurately. This is illustrated in Figure 1: the entropy (vertical axis) stabilizes when calculated for around 800 words (horizontal axis), but even before that, the distance between both entropies is constant. Similar tests by Wilbert Heeringa (personal communication), for the combination of Dutch and Frisian, show that word lists consisting of 500 or even fewer words for each language are long enough to reach stable entropy measures.

The lists which are used to calculate lexical distances and Levenshtein distances in many other publications are relatively short lists (for example, Heeringa et al., 2013; and Gooskens, Heeringa & Beijering, 2008 used about 100 words and Gooskens,

2007 used between 200 and 300 words). However, these lists are too short to reliably calculate entropy measures (Moberg et al., 2007). Therefore, for this study new word lists were created consisting of 500 word pairs, 410 of which were cognates. 100 of the words were also used for a word intelligibility experiment. The scripts used to calculate the entropies were written by Wilbert Heeringa.

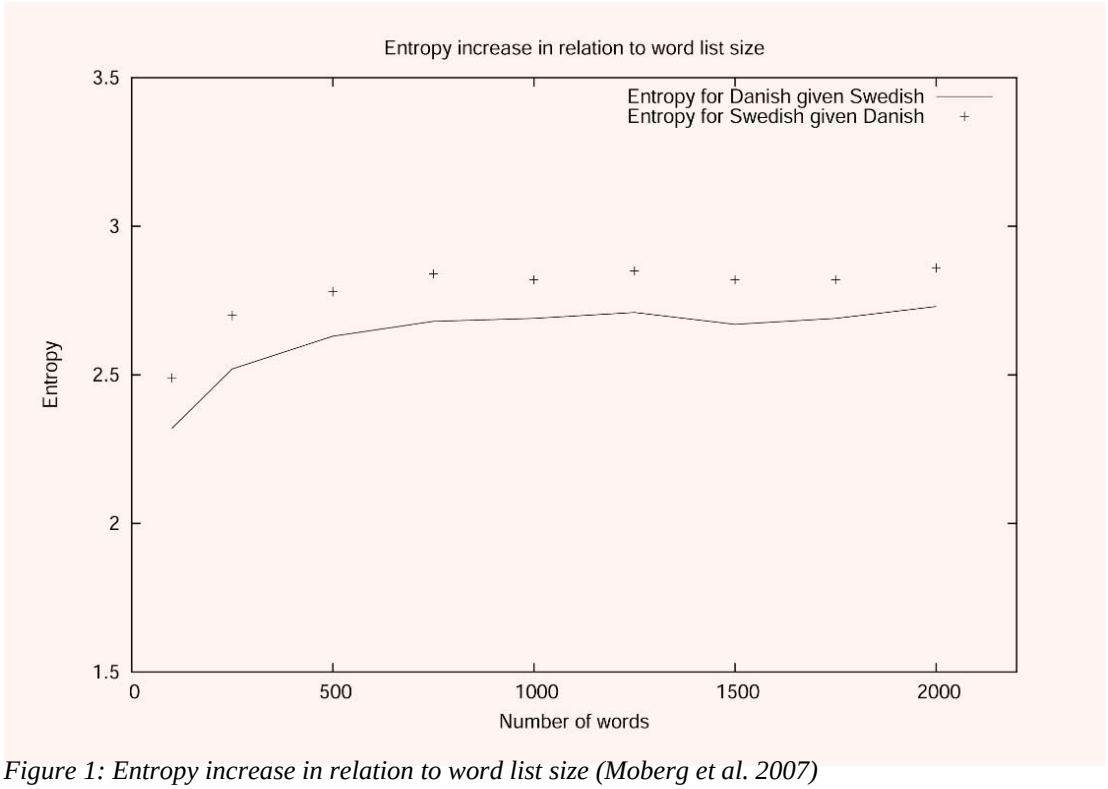

Figure 1: Entropy increase in relation to word list size (Moberg et al. 2007)

MATERIAL

The word list used for this study is based on the British National Corpus[49]. The 500 most frequent words in this corpus were translated into Swedish and Danish to create parallel word lists. Although the word list used in the intelligibility experiment (see 'Measuring intelligibility' below) consists of only nouns. The researcher's choice to use only nouns in the intelligibility experiment is based on the fact that nouns are concrete and therefore easier to understand. In this way, since in the overall project five different languages are used, we try to keep the results as similar as possible. Words from all parts of speech were included for the entropy calculations. The use of words from all parts of speech will make for a better representation of the languages than when only nouns are included, as nouns might behave differently from other word classes when it comes to linguistic similarity. For example, words that were

[49] http://www.natcorp.ox.ac.uk/

borrowed after the two languages separated tend to be influenced less by sound changes than inherited words, simply because they have been part of the lexicon for a shorter time, or because they have foreign properties which make them resistant to the changes (Gooskens, Kürschner & Van Bezooijen, 2012). If both languages borrowed a word from the same source, these words are then more similar to each other than a pair of cognate words which were inherited from a common ancestor language. Gooskens, Kürschner and Van Bezooijen (2012) found that Levenshtein distances between Danish and Swedish were lower for loan words than for inherited words. Loan words are often nouns. Including only nouns in the word list might therefore result in a too low estimate of the entropy between the languages.

But before any list or calculation could be made it had to be determined whether two corresponding words were cognates. The traditional definition of cognate words stresses the shared origin of the words in an older form of the languages, as in this definition from the Oxford English Dictionary (OED): "Coming naturally from the same root, or representing the same original word, with differences due to subsequent separate phonetic development". For this research, however, a broader definition was used. In the situation in which a speaker of one language is trying to understand the words of another language, he does not see the etymological history of a word. The only thing that matters to the reader or listener, is the fact that there is some kind of similarity to the corresponding word in his own language. Therefore, any two words of which the stems are related were considered cognates. This of course includes cognates in the sense of definition of the OED quoted above, but it also includes loan words sharing a common source, such as Danish *skole* and Swedish *skola*, which derived from Latin *schola* and both meaning 'school'. Words which share a base form but have different affixes were considered cognates as well. An example of this type is Danish *ledelse* with Swedish *ledning*, 'management'. Both words are derived from the same lexical item meaning 'to lead' (*lede* in Danish and *leda* in Swedish), but with two different suffixes. Despite this, the words are considered cognates. However, when a word consists of multiple lexical items (as opposed to a lexical item and an affix), and one of them is not related, the complete words were not considered cognates. Take, for example, the compounds *samfund* (Danish) and *samhälle* (Swedish) 'society'. The first parts of these words, *sam-*, are cognates of each other, but the second parts derive from different lexical items, *fund* and *hälle*.. The word pair as a whole was therefore not considered a cognate pair. Word pairs were considered a cognate pair only when the cognates still have the same meaning - as many language learners know, false friends, i.e., word forms that are more similar to the stimulus word than the correct translation, tend to impair intelligibility more than help it.

When there was doubt about whether or not two words shared the same origin, etymological dictionaries were used (Katlev, 2000; Ernby, 2008). Only word pairs with the same meaning in both languages were considered – false friends are no part of this study.

The words used for this study were translated into Danish and Swedish with help from internet sources, dictionaries and native speakers. During this process, some words were removed from the list because they proved to be too hard to be translated

reliably. These cases usually consisted of words from the original English list which simply do not exist, or at least do not exist in the same form, in one or both languages. A word like 'whatever', for example, does not have a clear translation in either of the languages, and any translation approaching the English meaning is a multi-word expression.

Because the basis of the word list was taken from an English-language corpus, the inevitable result is that the final word list is somewhat centered on English. The advantage of this, however, is that both Danish and Swedish are treated the same in this respect. There is no unfair advantage for either language.

For this study 500 words of both languages are translated and after that transcribed in IPA and X-SAMPA according to standardized speech as in the most recent pronunciation dictionaries (Hjorth & Kristensen, 2003-2005; Molbæk Hansen, 1990).

Measuring intelligibility

In order to determine whether entropy can help explain the asymmetry between Danish and Swedish, we compared the results of the entropy calculation to intelligibility data obtained using a subset of the word list. Intelligibility was measured with a word translation task, described in more detail below. This experiment was part of the MICReLa project, currently in progress at the University of Groningen (www.micrela.nl). It aims to map out the intelligibility among related languages in Europe and the factors determining this intelligibility. In this paper, we only present the results for Danish and Swedish. The MICReLa project included several experiments which were carried out online and presented as a game (www.micrela.nl/app). In this article, we will only focus on the word translation task, described below.

Word translation task

For the word translation task, participants were presented with 50 single isolated nouns, written or spoken, and were required to translate them. The language and mode (spoken or written) the participants took the test in were selected randomly for them by the application. They were encouraged to provide an answer even if they had no idea what the word could possibly mean. The words used in this task were the words from the word list used in the MICRela project (Heeringa et al., 2013). This list consists of the 100 most frequent nouns in the British National Corpus (BNC) translated into Danish and Swedish. All of these nouns were included in the 500-word list for the entropy calculations, too. Each participant was randomly assigned 50 words from this list. With this word translation task, the participants cannot use context to derive the meaning of a word: they only have that single word. Therefore, the influence of linguistic factors such as phonetic and orthographic distance as well as entropy between the languages can directly be investigated.

Participants

Participants were gathered by advertising the experiment as an online game. No one was paid for their participation. The participant groups included in the analysis were

selected and matched afterwards by the researchers. Only people who spoke Danish or Swedish natively, had a university education and were between 18 and 33 years old were included in the analysis. All participants had only one native language and no languages other than this language were spoken in their homes when growing up. They grew up in one of the countries included in the project, corresponding to their native language. In this selection, we made sure no Danish participants who had learned Swedish and no Swedish participants who had learned Danish were included.

The number of participants of the written word translation task is 30 of which 15 were Danes (average age: 24;3 / 10 male / 5 female) who took this task in Swedish and 15 were Swedes (average age: 26;6 / 9 male / 6 female) who took this task in Danish. The number of participants of the spoken word translation task is 48 of which 33 were Danes (average age: 23;9 / 22 male / 11 female) who took this task in Swedish and 15 were Swedes (average age: 22;6 / 8 male / 7 female) who took this task in Danish.

RESULTS

The results for the entropy calculations are displayed in Table 5. For orthographic entropy (representing written language), there is hardly any difference between the entropies in both directions (0.85 and 0.89). For the entropy calculated with the phonetic transcriptions of the words, however, a very clear asymmetry is found: the entropy for Swedish given Danish is much higher (1.45) than for Danish given Swedish (1.06). This means that Swedes listening to Danish have to deal with more unpredictability than Danes listening to Swedish. In addition, the phonetic entropy is in both cases higher than the orthographic entropy. This means that the correspondences between the two written languages are overall more regular and predictable than the correspondences between the spoken languages. As written language tends to be more standardized and changes more slowly that spoken language, this is not an unexpected finding.

Table 5: Orthographic and phonetic conditional entropy for Danish and Swedish based on 410 words

	Orthographic entropy	**Phonetic entropy**
H(Swedish|Danish) (Danish for Swedes)	0.85	1.45
H(Danish|Swedish) (Swedish for Danes)	0.89	1.06
Asymmetry	-0.04	0.39

The intelligibility scores from the word translation task, using partly the same words used for the entropy calculations, are displayed in Table 6. As in the previous literature (see Background), there is virtually no asymmetry in written intelligibility (85.5% and 88.0% of correctly translated words), but there is an asymmetry in

spoken intelligibility: the Danish participants understand a larger number of the Swedish words they are presented with (76.3%) than the Swedish participants understand the Danish words they are presented with (64.3%). Therefore, the language combination with the higher entropy (Swedes listening to Danish) had the lower intelligibility score. This result is as we expected.[50]

Table 6: Percentages of correctly translated written and spoken words.

	Written	**Spoken**
Danish for Swedes	85.5 %	64.3 %
Swedish for Danes	88.0 %	76.3 %
Asymmetry	-2.5 %	-12.0 %

CONCLUSION AND DISCUSSION

The spoken word translation task showed that Danes can understand spoken Swedish better than Swedes can understand spoken Danish. On the written word translation task, however, the performance of both groups is virtually the same. This is in line with the results of previous research on mutual intelligibility between Swedish and Danish (Maurud, 1976; Bø, 1978; Börestam, 1987; Delsing & Lundin Åkesson 2005; Gooskens et al., 2010; Schüppert, 2011; Gooskens & Van Bezooijen, 2013).

We have calculated conditional entropy for Swedish and Danish using a word list that contained the words included in the intelligibility task, expanded with more words in order to reach reliable entropy calculations. The entropy, like the intelligibility task, reveals asymmetry between Swedish and Danish on the spoken (phonetic) level, but no asymmetry on the written (orthographic) level. On the spoken level, the entropy in Danish for Swedes is higher than the entropy in Swedish for Danes. This means the higher entropy is found in the language pair where there is lower intelligibility. We have shown that entropy could be a reliable measure when explaining the asymmetry in intelligibility.

Our entropy results are in accordance with the results for conditional entropy published by Moberg et al. (2007), who inspired this article. Moberg et al. found a higher entropy for Swedish given Danish than for Danish given Swedish based on phonetically transcribed word lists for all types of words, and for each of the seven subgroups defined by them except for the subgroup of Latin/Greek/French loan words. Our list contains words of all of Moberg et al.'s groups and, like them, we have found entropy to be higher for Swedish given Danish than for Danish given Swedish – that is, a Swede faces a more complex mapping task when decoding Danish than vice versa. Moberg et al. did not calculate entropy based on orthographic

[50] Note that, for written intelligibility, the combination with the lower entropy has a lower intelligibility score, opposite to what we expected. However, the differences between the two groups are so small in this case that we cannot confidently speak of a 'higher' and 'lower' score.

transcriptions of their words, so we cannot compare their results to ours in that regard.

In order to further establish the value of entropy, this research should be expanded with more languages in addition to Danish and Swedish. In many studies concerning Scandinavian languages cited in this article, Norwegian was included as well, for example. In addition to this, more Germanic languages could be included, and the research could be repeated for other language groups. This will show whether the relation holds beyond the combination of Danish and Swedish.

In addition to expanding the number of languages included, the word list should be expanded. For this study, 410 cognates were used, which is a little bit less than the amount of 500 words needed for reliable measures. When more words are included, the entropy calculations will be more reliable, and the entropies for different language combinations can be more reliably compared to each other.

A more fundamental issue with this study is the way in which intelligibility is measured. The participants who were included in the analysis were specifically selected on not having purposefully learned their neighboring language before. This selection is necessary, as we want to measure the level of intelligibility the participants reach solely by virtue of the knowledge of their own native language; we are not trying to measure how well they can learn a language. Conditional entropy, however, does not measure actual similarity or distance between two languages. Instead, it measures the regularity of the correspondences among the languages. A situation in which a /p/ usually corresponds to a /p/ has the same amount of entropy as when a /p/ usually corresponds to a /f/. For a participant who has never encountered the unknown language before, however, the latter correspondence is not necessarily obvious. Some exposure to the language is necessary for the participant to discern the patterns in the correspondences. Only then, he will get an advantage from a low entropy, that is, high regularity. For this same reason, entropy as an influencing factor of asymmetry can also explain Schüppert's (2011) finding that the Danish-Swedish asymmetry does not exist for preschoolers. After all, preschoolers in general and the participants of Schüppert's study in particular have not yet had much exposure to foreign languages.

As Danish and Swedish are neighboring languages spoken in countries that have good relations to each other, virtually all participants in this study will have had some exposure to the other language in their lives. This means that the regularity of correspondences can very well have an effect on their understanding of that language. However, we have not controlled in any way how much or what kind of exposure the participants have had to that language. The 50 words of the language every participant was presented with during the experiment are not likely to be enough to make a noticeable difference in understanding.

Future research should establish how much and what kind of exposure maximizes the benefit of a low entropy and how likely speakers of both languages are to experience that in their lives. Finally, future research should establish how this knowledge can be used to increase intelligibility among the speakers of the languages without requiring all of them to actively learn the other language.

ACKNOWLEDGEMENTS
We are grateful to Wilbert Heeringa for teaching us how to calculate the entropies, allowing us to use the scripts he has written and supporting us by both answering our questions and solving the technical issues we ran into.

REFERENCES

Beijering, Karin, Charlotte Gooskens and Wilbert Heeringa (2008). Predicting intelligibility and perceived linguistic distances by means of the Levenshtein algorithm. *Linguistics in the Netherlands,* 13-24.

Bø, Inge (1978). *Ungdom og naboland. En undersøkelse av skolens og fjernsynets betydning for nabospråksforståelsen.* [Youth and neighbouring country. An investigation of the influence of school and TV on inter-Scandinavian comprehension.] Stavanger: Rogalandsforskning (rapport 4).

Börestam, Ulla (1987). *Dansk-svensk språkgemenskap på undantag.* [Danish-Swedish language community as a special case.] Uppsala: Uppsala Universitet.

Börestam Uhlmann, Ulla (1991). *Språkmöten och mötesspråk i Norden.* [Language meetings and meeting languages in the Nordic countries.] Nordisk språksekretariats rapporter 16. Oslo: Nordisk språksekretariat.

Braunmüller, Kurt (2002). Semicommunication and accommodation: Observations from the linguistic situation in Scandinavia. *International Journal of Applied Linguistics,* 12(1), 1-23.

Conti, Virginie and François Grin (eds.) (2008). *S'entendre entre langues voisines: vers l'intercompréhension* [Understanding between closely related languages: towards intercomprehension]. Chêne-Bourg: Georg, 79-109.

Delsing, Lars-Olof and Katarina Lundin Åkesson (2005). *Håller språket ihop Norden? En forskningsrapport om ungdomars förståelse av danska, svenska och norska.* [Does the language keep together the Nordic countries? A research report of mutual comprehension between young speakers of Danish, Swedish and Norwegian.] Copenhagen: Nordiska ministerådet.

Doetjes, Gerke and Charlotte Gooskens (2009). Skriftsprokgets rolle i den dansk-svenske talesprogsforståelse. [The role of orthography in spoken language comprehension between Danish and Swedish.] *Språk och stil,* 19, 105-123.

Ernby, Birgitta (2008). *Norstedts etymologiska ordbok* . Norstedts Akademiska Förlag.

Gooskens, Charlotte (2006). Linguistic and extra-linguistic predictors of Inter-Scandinavian intelligibility. In: J. van de Weijer & B. Los (eds.), *Linguistics in the Netherlands,* 23, Amsterdam:John Benjamins, 101-113.

Gooskens, Charlotte (2007). The contribution of linguistic factors to the intelligibility of closely related languages. *Journal of Multilingual and Multicultural Development, 28*(6), 445-467.

Gooskens, Charlotte and Renée van Bezooijen (2013). Explaining Danish-Swedish asymmetric word intelligibility – An error analysis. In: C. Gooskens & R. van Bezooijen (eds.), *Phonetics in Europe: Perception and Production.* Frankfurt a.M.: Peter Lang, 59-82.

Gooskens, Charlotte, Wilbert Heeringa and Karin Beijering (2008). Phonetic and lexical predictors of intelligibility. *International Journal of Humanities and Arts Computing*, 2(1-2), 63-81.

Gooskens, Charlotte, Vincent van Heuven, Renée van Bezooijen and Jos Pacilly (2010). Is spoken Danish less intelligible than Swedish? *Speech Communication, 52,* 1022-1037.

Gooskens, Charlotte and Sebastian Kürschner (2010). Hvilken indflydelse har danske stød og svenske ordaccenter på den dansk-svenske ordforståelse? [What influence do the Danish stød and the Swedish tonemes have on Danish-Swedish word recognition?] In: C. Falk, A. Nord & R. Palm (eds.), *Svenskans beskrivning 30. Förhandlingar vid Trettionde sammankomsten for svenskans beskrivning, Stockholm den 10 och 11 oktober 2008*, 82-92.

Gooskens, Charlotte, Sebastian Kürschner and Renée van Bezooijen (2012). Intelligibility of Swedish for Danes: Loan words compared with inherited words. In: H. van der Liet & M. Norde (eds.), *Language for its own sake.* Amsterdam: Amsterdam Contributions to Scandinavian Studies 8, 435-445.

Haugen, Einar (1966). Semicommunication: The language gap in Scandinavia. *Sociological Inquiry, 36,* 280-297.

Heeringa, Wilbert (2004). Measuring dialect pronunciation differences using Levenshtein distance. PhD thesis. Groningen: Grodil, 46.

Heeringa, Wilbert, Jelena Golubovic, Charlotte Gooskens, Anja Schüppert, Femke Swarte and Stefanie Voigt (2013). Lexical and orthographic distances between Germanic, Romance and Slavic languages and their relationship to geographic distance. In: C. Gooskens & R. van Bezooijen (eds.), *Phonetics in Europe: Perception and Production*. Frankfurt a.M.: Peter Lang, 99-137.

Hilton, Nanna Haug, Anja Schüppert and Charlotte Gooskens (2011). Syllable reduction and articulation rates in Danish, Norwegian and Swedish. *Nordic Journal of Linguistics,* 34(2), 215-237.

Hjorth, Ebba and Kristensen, Kjeld. 2003-2005. *Den danske ordbog.* Copenhagen: Det Danske Sprog- og Litteraturselskab.

Katlev, Jan (2000). *Politikens Etymologisk Ordbog*. Aalborg: Politikens Forlag.

Kürschner, Sebastian, Charlotte Gooskens and Renée van Bezooijen (2008). Linguistic determinants of the intelligibility of Swedish words among Danes. *International Journal of Humanities and Arts Computing, 2*(1-2), 83-100.

Lüdi, G. (2007). The Swiss model of plurilingual communication. In Ö J. D. ten Thije & L. Zeevaert (eds.), *Receptive multilingualism: Linguistic analyses, language policies and didactic concepts*. Amsterdam: Benjamins, 159-178.

Maurud, Øivind (1976). *Nabospråkforståelse i Skandinavia. En undersøkelse om gjensidig forståelse av tale- og skriftspråk i Danmark, Norge og Sverige.* [Neighbouring language comprehension in Scandinavia. An investigation of mutual comprehension of written and spoken language in Denmark, Norway and Sweden.] Stockholm: Nordiska rådet.

Moberg, Jens, Charlotte Gooskens, John Nerbonne and Nathan Vaillette (2007). Conditional entropy measures intelligibility among related languages. In: P. Dirix, I. Schuurman, V. Vandeghinste and F. Van Eynde (eds.), *Computed Linguistics in the Netherlands 2006*: *Selected papers from the 17th CLIN Meeting*. Utrecht: LOT, 51-66.

Molbæk Hansen, Peter. 1990. *Udtaleordbog*. Copenhagen: Gyldendal.

Oxford English Dictionary (OED), online edition. Online at: www.oed.com (retrieved April-September 2014).

Rehbein, Jochen, Jan D. ten Thije and Anna Verschik. (2012). Lingua Receptiva (LaRa). Remarks on the quintessence of receptive multilingualism. *International Journal of Bilingualism*, 16(3), 248-264.

Schüppert, Anja (2011). *Origin of asymmetry. Mutual intelligibility of spoken Danish and Swedish.* PhD thesis. Groningen: Grodil, 94.

Schüppert, Anja and Charlotte Gooskens (2011). Investigating the role of language attitudes for perception abilities using reaction time. *Dialectologia*, 7, 119-140.

Schüppert, Anja and Charlotte Gooskens (2012). The role of extra-linguistic factors for receptive bilingualism: Evidence from Danish and Swedish pre-schoolers. *International Journal of Bilingualism,* 16(3), 332-347.

Schüppert, Anja, Nanna Haug Hilton and Charlotte Gooskens (2015). Swedish is beautiful, Danish is ugly? Investigating the link between language attitudes and intelligibility. *Linguistics*, 53(2), 375-403.

Schüppert, Anja, Johannes Ziegler, Kristina Magnusson, Holger Juul, Kenneth Holmqvist and Charlotte Gooskens (submitted). On-line activation of L1 orthography enhances spoken word recognition of a closely related L2. Evidence from ERP.

Swarte, Femke (in preparation). *Mutual intelligibility in the Germanic Language Area. PhD thesis*. Groningen: Grodil.

Swarte, Femke, Anja Schüppert and Charlotte Gooskens (accepted). Does German help speakers of Dutch to understand written and spoken Danish words? - The role of second language knowledge in decoding an unkown but related language. In: G. De Angelis, U. Jessner and M. Kresic (eds.), *Crosslinguistic Influence and Multilingualism*.

ten Thije, Jan D. and Ludger Zeevaert (eds.) (2007). *Receptive multilingualism: Linguistic analyses, language policies and didactic concepts.* Amsterdam: John Benjamins.

ten Thije, Jan D. (2013). Lingua Receptiva (LaRa). *International Journal of Multilingualism*, 10, 2, 137-139.

Zeevaert, Ludger. (2007). Receptive multilingualism and inter-Scandinavian semicommunication. In: J. D. ten Thije & L. Zeevaert (eds.), *Receptive multilingualism: Linguistic analyses, language policies and didactic concepts*. Amsterdam: Benjamins, 103-135.

Biographical notes

Anne Kingma, BA in linguistics from the University of Nijmegen, currently completing the Research Master's in Linguistics at the University of Groningen. Interested in Computational Linguistics, Language Variation and Change, and Syntax. Contact: anne.s.kingma@gmail.com.

Felicity F. Frinsel, BA in linguistics from the University of Groningen. She applied for the Research Master in *Language and Cognition* at the University of Groningen and has the ambition to continue with a PhD position in the field of theoretical linguistics. Interested in Language Variation, Syntax, Semantics and Psycholinguistics. Contact: felicityfrinsel@hotmail.com.

Femke Swarte, MA in Linguistics and in German from the University of Groningen, is a PhD student in the project *Mutual intelligibility of closely related languages in Europe: linguistic and nonlinguistic determinants* (NWO Vrije Competitie) led by Charlotte Gooskens (University of Groningen) and Vincent van Heuven (Leiden University).

Charlotte Gooskens, PhD, is Associate Professor of Applied Linguistics at the University of Groningen. Her work covers a wide variety of aspects of the mutual intelligibility of closely related languages. She has been working as a principal investigator of large NWO-financed projects (Nederlandse Organisatie voor Wetenschappelijk Onderzoek): *Linguistic determinants of mutual intelligibility in Scandinavia* (NWO Vidi, 2006-2011) and *Mutual intelligibility of closely related languages in Europe: linguistic and nonlinguistic determinants* (NWO Vrije Competitie).