Predicting intelligibility and perceived linguistic distance by means of the Levenshtein algorithm[1]

Karin Beijering[*], Charlotte Gooskens[*] and Wilbert Heeringa[+]

[*]*University of Groningen,* [+]*Meertens Institute*

1. Introduction

In this article, we investigate the predictive value of so-called Levenshtein distances for both intelligiblity scores and perceived linguistic distances. Additionally, we compare two measuring methods, namely normalized and non-normalized Levenshtein distances. The Levenshtein algorithm is a string edit distance measure that quantifies the distance between the pronunciations of corresponding words in different dialects or closely related languages. It calculates the minimal costs required to change a string of segments into another by means of insertions, deletions or substitutions. Kessler (1995) introduced the algorithm for measuring distances between Irish Gaelic dialects. Since then it has been applied successfully to Dutch dialects (Heeringa 2004, pp. 213-278), Sardinian dialects (Bolognesi & Heeringa 2002), and German dialects (Nerbonne & Siedle 2005).

In Gooskens & Heeringa (2004), Norwegian listeners judged linguistic distances between recordings of 14 Norwegian varieties and their own language variety on a scale from 1 (similar to own dialect) to 10 (not similar to own dialect). These subjective distances were correlated with objective Levenshtein distances between these dialects. The correlation was significant and reasonably high (r=.67, p<0.01). A similar score (r=.66, p<0.01) was found by Tang & Van Heuven (2007), who correlated objective linguistic distances with judged similarity measurements between Chinese dialects. Gooskens (2006) correlated Levenshtein distances with objective mutual functional intelligibility scores, i.e. the percentage of correctly translated words from a spoken text, of the closely related languages Danish, Norwegian and Swedish, and found an even stronger correlation (r=.82, p<0.01). Given the higher correlation between Levenshtein

distance and intelligibility scores, the hypothesis arises that Levenshtein distance is a better predictor of intelligibility than of perceived distance. However, since different language varieties were included in the three investigations, it is not possible to conclude that Levenshtein distance is more suitable for predicting intelligibility than perceived distances. In the present investigation we tested both intelligibility and perceived distances of 18 Scandinavian language varieties among Danish listeners and correlated the results with Levenshtein distances in order to answer the following question:

1. How well do Levenshtein distances predict intelligibility and perceived distances between closely related language varieties?

In the Levenshtein algorithm, normalization for word length can be implemented. The effect of normalization is that a pronunciational difference weighs heavier in a short word than in a long word. Previous applications of the Levenshtein algorithm typically employed a word length normalization, which means that the total number of operations (insertions, deletions and substitutions) is divided by the number of alignment slots for a word pair (see example in section 3).

In our investigation, we correlated the intelligibility scores and the perceived distance measurements with normalized as well as non-normalized Levenshtein distances in order to answer the following question:

2. How well do normalized and non-normalized Levenshtein distances predict intelligibility and perceived distances between closely related language varieties?

With regard to *intelligibility*, word length normalization is applied in order to account for the phenomenon that a segmental difference in for example a word of two segments has a stronger impact on intelligibility than a segmental difference in a word of ten segments. However, when predicting *perceived distance* by means of Levenshtein distance, it cannot be assumed that Levenshtein distances should be normalized. The total number of deviant segments in another language variety is likely to be important when judging the distance to one's own language variety (and not whether they occur in long or short words). Accordingly, Heeringa et al. (2006) found non-normalized objective distances to correlate more strongly with perceived distances than normalized distances do.

2. Material

We included recordings and transcriptions of the same text, the fable *The North Wind and the Sun*[2], in 18 different language varieties in our investigation (see Figure 1).

Figure 1. Map of Scandinavia with the location of the 18 Scandinavian language varieties.

From a selection of recordings in more than 50 different Norwegian dialects, we chose eight dialects that form a good representation of the dialectological and geographical diversity of Norway.[3] In addition, we made extra recordings of Faroese (Thorshavn), Standard Swedish (as spoken in Stockholm), four Swedish dialects representing the four major dialect groups (including Finland Swedish), Standard Danish (as spoken in Lyngby, close to Copenhagen) and three Danish dialects spoken on the peninsula of Jutland. The standard varieties of Danish, Norwegian[4] and Swedish all belong to the Mainland Scandinavian branch of the North Germanic language family and are known to be mutually intelligible. Speakers of these varieties can in principle communicate with each other in their own language, though sometimes with some effort. The most recent investigation into the mutual intelligibility in Scandinavia was carried out by Delsing & Lundin Åkesson (2005). Their results show that Danes and Swedes

have the greatest difficulty understanding each other's languages, whereas they have fewer problems understanding Norwegian. Norwegians in general understand both neighboring languages well. So far, the intelligibility of non-standard language varieties in a Scandinavian context has not been investigated. Faroese belongs to the Insular Nordic branch of the North Germanic language family and is almost unintelligible to speakers of Mainland Scandinavian without prior instruction.

The Norwegian version of *The North Wind and the Sun* was first translated into Standard Danish, Swedish and Faeroese. These texts were then presented to the speakers of the non-standard varieties of these languages to be translated and recorded. The 18 text versions comprised between 91 and 111 words, with a mean of 98 words. The 18 recordings were used for listening experiments (see Sections 4 and 5). In addition, phonetic transcriptions were made of all recordings.[5] These transcriptions were used to calculate the Levenshtein distances (see Section 3).

3. Levenshtein distances

Phonetic distances between Standard Danish (Lyngby) and each of the other 17 Scandinavian language varieties were calculated by means of the Levenshtein algorithm. When phonetic transcriptions of two pronunciations are compared with each other, Levenshtein distance is equal to the number of operations needed to transform one transcription into another. There are three types of operations: insertions, deletions and substitutions of phonetic symbols. The power of the Levenshtein distance is that it chooses the operations that transform one pronunciation into another in such a way that the total number of string operations is minimal.

We will illustrate the algorithm with an example. The form *enige* (meaning 'in agreement') is pronounced as [ʔeːni] in Lyngby (Standard Danish) and as [eːnɪgɑ] in Stockhom (Standard Swedish). Ignoring suprasegmentals and diacritics, the pronunciation of Lyngby can be changed into the pronunciation of Stockholm as in (1).

(1)

Lyngby	ʔeni	delete ʔ	1
	eni	substitute i by ɪ	1
	enɪ	insert g	1
	enɪg	insert ɑ	1
Stockholm	enɪgɑ		
			4

In this example all operations (insertion, deletion or substitution of a segment) contribute to the transformation to the same degree, i.e. every operation 'costs' one point. In this way we get a total cost of 4. There is no way to transform the one sequence into the other at a lower cost.

In the example we found that the distance between the two pronunciations equals 4. This is the non-normalized Levenshtein distance. In this paper we study both non-normalized and normalized Levenshtein distance (see question 2 in Section 1). The normalized distance is obtained by dividing the non-normalized distance by the length of the longest alignment, which gives the minimum costs (see Heeringa 2004, pp. 130-132). In our example we get six alignment slots, as shown in (2). The normalized Levenshtein distance is 4/6 = 0.67 or 67%.

(2)

	1	**2**	**3**	**4**	**5**	**6**
Lyngby	ʔ	e	n	i		
Stockholm		e	n	ɪ	g	ɑ
	del			sub	ins	ins

The text *the North Wind and the Sun* consists of 58 different words in most dialects. In that case at most 58 word pairs will be taken into account when a variety is compared to standard Danish. The phonetic distance is calculated as the average of 58 Levenshtein distances that correspond to the 58 word pairs. However, since only cognate pairs (i.e. historically related words) are considered, the number of word pairs was usually lower. Standard Danish (Lyngby) shares the highest percentage of cognates with the variety of Høgsted

(100%), and the lowest percentage with the variety of Thorshavn (86%). On average Standard Danish shares 96% of the cognates with the other 17 varieties[6]. In order to achieve distances which are based on linguistically motivated alignments that respect the syllable structure of a word or the structure within a syllable, the algorithm was adapted so that a vowel may only correspond to a vowel and a consonant to a consonant. The semi-vowels [j] and [w] may also correspond to a vowel or the other way around, their vowel counterparts [i] and [u] may correspond to a consonant or the reverse. The central vowel schwa [ə] may correspond to any sonorant. In this way, unlikely matches like [o] and [t] or [s] and [e] are prevented.

In the example above, all operations have the same cost. In the present study we use graded operation weights. Spectrograms were made on the basis of sound samples from *The Sounds of the International Phonetic Alphabet* (1995). On the basis of the spectrograms, distances were measured between the IPA vowels and pulmonic consonants (see Heeringa 2004, pp.97-107). We used a Barkfilter representation, which we consider as a more perceptually-oriented spectrogram since it has a more or less logarithmic frequency scale, a logarithmic power spectral density and the 24 first critical bands are modeled (see Heeringa 2004, pp. 87-88 for more details). The Barkfilter distances are used as operation weights. In this way the fact that for example [a] and [A] are phonetically closer to each other than [a] and [i] is taken into account.

In validation work Heeringa (2004) found the tendency that Levenshtein distances based on logarithmic gradual segment distances approach perception better than Levenshtein distances based on linear gradual segment distances (see pp. 185-186). Although the Barkfilter represention already is logarithmic in itself since it has a logarithmic power spectral density, the use of logarithmic Barkfilter segment distances still gave some further improvement. This gives the impression that it does not matter so much *to what extent* segments differ, but only t*he fact that* they differ.

4. Intelligibility

The intelligibility of the 18 language varieties was tested in a listening experiment. The listeners were 351 native speakers of Danish between 15 and 19 years of age (average 17.0) from 18 Highschool classes in Copenhagen. Since the listeners lived in Copenhagen, we assumed that they all spoke Standard Danish or at least were familiar with this language variety. Some of the listeners may have been familiar with some of the language varieties presented in the test. However, people living in Copenhagen in general do not have much experience

with the Danish dialects of Jutland or the other Scandinavian dialects. The task of the listeners was to translate the recordings of the fable *The North Wind and the Sun* as precisely as possible into Standard Danish. Due to lack of space, the precise design of the intelligibility experiment cannot be discussed in detail (see Beijering (2007, pp.57-60) for a comprehensive description of this experiment).

4.1. *Results*

The percentage of correctly translated words constituted the intelligibility score of a given language variety. A correctly translated word got one point, partly correctly translated words got half a point. For example, if only the last part of the word *nordenvinden* 'The North Wind' was correctly translated, half a point was given. We excluded Lyngby, representing standard Danish, from the analysis. This recording was only included to check that the test was feasible. Since 99% of the Lyngby words were translated correctly, we conclude that this was indeed the case.

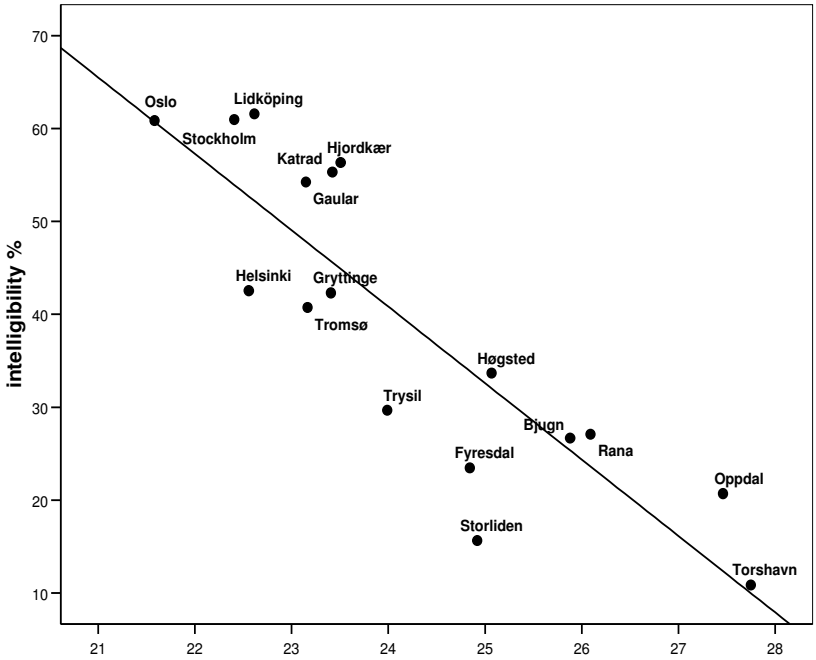

Figure 2. Scatterplot of the correlation between normalized Levenshtein distances and intelligibility scores (r=-.86, p<.01).

The scatterplot in Figure 2 shows the relation between normalized Levenshtein distances and the intelligibility scores for the 17 Scandinavian language varieties. As expected from the fact that Faeroese belongs to another branch of the North Germanic language family, the Faeroese variety from Thorshavn is most deviant from Standard Danish and it was difficult to understand for the Danish listeners. One of the Norwegian dialects (Oppdal) is almost just as deviant and also difficult to understand. The standard languages of Sweden and Norway (Oslo and Stockolm) are most similar to standard Danish and also most easily understood, even more so than the Danish dialects of Hjordkær, Katrad and Høgsted. The strong correlation between the Levenshtein distances and the intelligibility scores (r=-.86, p<.01) shows that intelligibility can be predicted well from the Levenshtein distances (the larger the distances, the more difficult it is to understand the dialects). This result is in line with previous investigations (see Section 1).

In addition to the normalized Levenshtein distances, we also correlated the intelligibility scores with the non-normalized distances. As explained in the introduction, we expected a lower correlation between the non-normalized distances and intelligibility scores, since these distances do not take the length of words into account. We found that the intelligibility scores correlated less strongly with the non-normalized Levenshtein distances (r=-.79, p<.01) than with the normalized Levenshtein distances (r=-.86, p<.01). However, this is only a tendency since the difference is not significant.

5. Perceived distance

The perceived distance between Standard Danish and the 17 language varieties was measured by means of an auditory judgment task. The listeners were matched[7] as well as possible with the listeners in the intelligibility experiment. Fifty-five Highschool pupils from Greve, a place close to Copenhagen, aged 15 to 19 (mean 17.6) participated in the experiment. They were all native speakers of Danish. The pupils listened to versions of *The North Wind and the Sun* in different language varieties and had to judge how (dis)similar each language variety was to Standard Danish on a scale from 0 to 10, 0 meaning 'this language variety is similar to Standard Danish' and 10 meaning 'this language variety is not similar to Standard Danish'. Due to lack of space, the precise design of the perceived distance experiment cannot be discussed in detail (see Beijering (2007, pp.75-77) for a comprehensive description of this experiment).

5.1. *Results*

For each language variety the mean perceived distance was computed. Like in the case of the intelligibility scores, we excluded Standard Danish (Lyngby) from our analysis. The perceived distance to Lyngby was very small (0.06), which again confirms that this language variety is a good representation of Standard Danish. The mean distance judgment over all the dialects is 5.2.

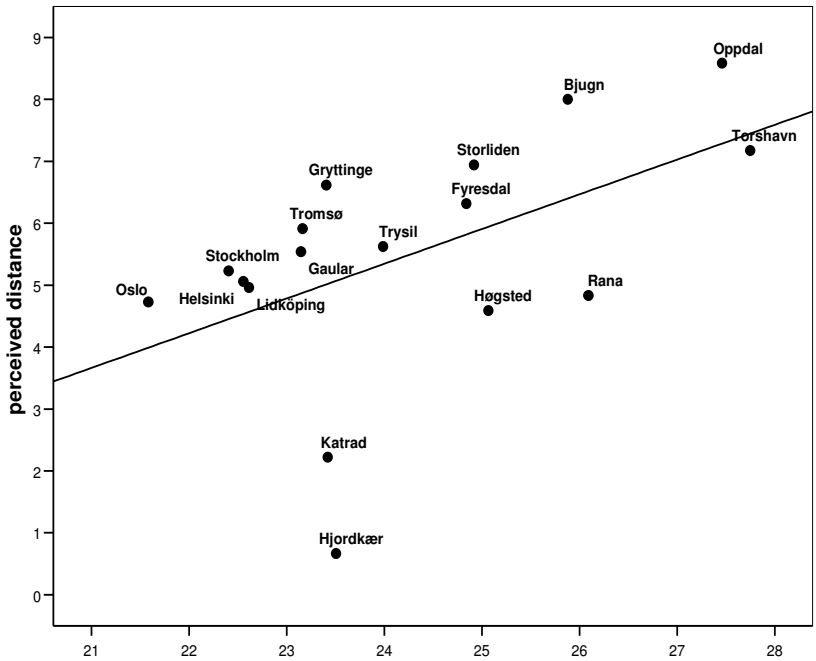

Figure 3. Scatterplot of the correlation between normalized Levenshtein distance and perceived distance (r=.52 , p<.05).

The scatterplot in Figure 3 shows the relation between normalized Levenshtein distances and the perceived distance scores for the different Scandinavian language varieties. The correlation is rather low and only significant at the 5% level (r=.52). In general, the most deviant language varieties according to the Levenshtein distances are also perceived as most deviant. The two Danish dialects of Hjordkær and Katrad are perceived as less deviant than one would expect from the Levenshtein distances. We will come back to this in Section 6.
We expect non-normalized Levenshtein distances to be a better predictor of the perceived distances than normalized distances since perceived distance is likely

to be dependent on the total number of deviant sounds regardless of word length. The listeners base their judgments on recordings of the whole text and it probably does not matter whether deviant sounds are part of long or short words. The correlation with the non-normalized Levenshtein distances was indeed stronger (r=.62, p<.01). However, the hypothesis cannot be confirmed since the difference (r=.52 versus r=.62) is not significant.

6. Conclusions

We measured objective phonetic distances between Standard Danish and 17 other Scandinavian language varieties by means of the Levenshtein algorithm and correlated these distances with intelligibility scores and perceived linguistic distances obtained from speakers of Standard Danish. We found high correlations in both cases. The normalized Levenshtein distances correlated significantly more strongly with the intelligibility scores than with the perceived distances (r=-.86 versus r=.52).[8] The non-normalized Levenshtein distances showed the same tendency, but the difference is not significant (r=-.79 versus r=.62, respectively). The results suggest that normalized Levensthein distance is indeed a better predictor of intelligibility than of perceived linguistic distances. There are three feasible explanations for this.

First, the perceived distances are based on the judgments of the whole recordings, including information at all linguistic levels. This means that we do not know which characteristics of the text have determined the listeners' judgments. It is possible that in some cases a single very deviant word, sound or morphosyntactic feature may have caused a listener to judge the whole text as very deviant. The intelligibility scores, on the other hand, are based on the percentage correctly translated words. This means that the intelligibility of all words in the whole text contributes to the mean score. It could be argued that a better way of measuring the intelligibility would have been to let the listeners answer questions about the text, since this task would have been more comparable to the perceived distance measurements. This method was used in Gooskens (2006) and here a very high correlation with normalized Levenshtein distances was found too (r=-.82). Single deviant sounds and words may play a less important role for intelligibility than for perceived distances.

Furthermore, the Levenshtein distances express phonetic distances leaving out information about differences at other linguistic levels. The impact of differences at other levels (lexical, prosodic, morphological, syntactic) on intelligibility may be different from the impact of these other levels on perceived distance. For example the meaning of an unknown lexeme might become clear

from the context, and therefore it may affect intelligibility less strongly than perceived distance. In Figure 2 we found that the intelligibility scores for Katrad and Hjordkær were quite well predicted by the Levensthein distances, but in Figure 3 we found that these two varieties are perceived as less deviant than predicted from the Levenshtein distances. At the lexical level the two varieties are much closer to Standard Danish than the other varieties. The fact that Katrad and Hjordkær do so much deviate from the regression line in Figure 3 while in Figure 2 they do not, may indicate that lexical variation affects intelligibility less strongly than perceived distance.

Finally, non-linguistic factors such as geographical knowledge or attitude towards the language varieties may have a larger effect on perceived distances than on intelligibility scores. If listeners know where a language variety is spoken, they may use this knowledge when judging the linguistic distance. Listeners may have the idea that geographically close varieties are also linguistically close, and vice versa. On the other hand, geographical knowledge is of no use when trying to understand the text. Similarly, if for some reason listeners have a negative attitude towards the target language variety, they may judge it as more deviant from their own language variety than if they have a positive attitude. This kind of behavior is known from the literature (e.g. van Bezooijen 1994). There are also indications in the literature that there is a correlation between intelligibility and attitude (e.g. Wolff 1959). A positive attitude may encourage subjects to try and understand the language in question, whereas a negative attitude will discourage subjects from making an effort. However, in experimental settings the relation between attitudes and intelligibility has been weak (e.g. Delsing & Lundin Åkesson 2005, Gooskens 2006, van Bezooijen & Gooskens 2007) and the relation is probably less strong than in the case of a judgment task. Therefore, we expect that the effect of attitude has been stronger in our perceived distance experiment than in our intelligibility test.

Our results and the considerations given above show that the Levenshtein algorithm is successful in predicting intelligibility. The correlations are lower in the case of perceived distances, but still significant. When applying the method it is important to be conscious of what one wants to measure. The distances as perceived by laymen are most likely influenced by factors other than phonetic distances. However, this does not mean that the Levenshtein algorithm is not a good method for measuring objective, aggregate phonetic distances.

A second aim of our investigation was to test the hypothesis that normalized Levenshtein distances correlate better with intelligibility scores than non-normalized distances do. The intelligibility results show this tendency but the hypothesis cannot be confirmed since the differences are not significant. The perceived distances correlated more strongly with non-normalized distances than

with normalized distances, but this difference was also insignificant. Heeringa et al. (2006) found the same tendency, but in their study the difference was not significant either. On the basis of these tendencies preference may be given to non-normalized distances when predicting distance judgments by means of Levenshtein distance and to normalized distances when predicting intelligibility.

References

Beijering, K. 2007. *The role of phonetic and lexical distances in the intelligibility and perception of Scandinavian language varieties for speakers of standard Danish.* MA-thesis, University of Groningen.
Available at: www.rug.nl/staff/k.beijering/projects

Bezooijen, R. van. 1994. "Aesthetic evaluation of Dutch language varieties". *Language and communication* 14. 253-263.

Bezooijen, R. van & C. Gooskens. 2007. "Interlingual text comprehension: linguistic and extralinguistic determinants". *Receptive Multilingualism and intercultural communication: Linguistic analyses, language policies and didactic concepts* ed. by J. D. ten Thije & L. Zeevaert, 249-264. Amsterdam: Benjamins.

Bolognesi, R. & W. Heeringa. 2002. De invloed van dominante talen op het lexicon en de fonologie van Sardische dialecten [The influence of dominant languages op the lexicon and phonology of Sardinian dialects]. *Gramma/TTT; tijdschrift voor taalwetenschap* 9. 45-84.

Delsing, L.O. & K. Lundin Åkesson. 2005. *Håller språket ihop Norden?: en forskningsrapport om ungdomars förståelse av danska, svenska och norska* [Does the Language Keep the Nordic Countries Together? A Research Report on How Well Young People Understand Danish, Swedish and Norwegian]. Copenhagen: Nordiska ministerrådet/Nordiska Kulturfonden.

Gooskens, C. 2006. "Linguistic and extra-linguistic predictors of Inter-Scandinavian intelligibility". *Linguistics in the Netherlands* 23 ed. by J. van de Weijer & B. Los, 101-113. Amsterdam: John Benjamins.

Gooskens, C. & W. Heeringa. 2004. "Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data". *Language variation and change* 16. 189-207.

Heeringa, W. 2004. *Measuring dialect pronunciation differences using Levenshtein distance*. Doctoral dissertation, University of Groningen.

Heeringa, W. P., Kleiweg, C. Gooskens & J. Nerbonne. 2006. "Evaluation of String Distance Algorithms for Dialectology". *Linguistic Distances Workshop at the joint conference of International Committee on*

Computational Linguistics and the Association for Computational Linguistics, Sydney, July, 2006 ed. by J. Nerbonne & E. Hinrichs, 51-62.

Kessler, B. 1995. Computational dialectology in Irish Gaelic. *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics*, 60-67. Dublin: EACL.

Nerbonne, J. & C. Siedle. 2005. "Dialektklassifikation auf der Grundlage aggregierter Ausspracheunterschiede". *Zeitschrift für Dialektologie und Linguistik* 72. 129-147.

Tang, C. & V. J. van Heuven. 2007. "Mutual intelligibility and similarity of Chinese dialects. Predicting judgments from objective measures". *Linguistics in the Netherlands* 24 ed. by B. Los & M. van Koppen, 223-234. Amsterdam: John Benjamins.

Wolff, H. 1959. "Intelligibility and inter-ethnic attitudes." *Anthropological Linguistics* 1. 34-41.

---

[1] We thank two anonymous reviewers for useful comments on an earlier version of this paper.

[2] *The North Wind and the Sun* is a well-known text in phonetic research. In *The principles of the International Phonetic Association* (1949) the text is transcribed in 51 different languages.

[3]The recordings were made by Jørn Almberg and Kristian Skarbø. They are made available via http://www.ling.ntnu.no. We thank for their permission to use the material.

[4] In Norway there is no spoken standard language. The Oslo variety represented the standard variety in this investigation because it approximates the written and media language in most aspects.

[5]The phonetic transcriptions of the Norwegian varieties were made by Jørn Almberg. The rest of the transcriptions were made by Andreas Vikran and corrected by Jørn Almberg to ensure consistency.

[6] Since this paper is about Levenshtein distance as a predictor of intelligibility and perceived distance we did not include lexical distance (the percentage of non-cognate words between two language varieties) in the analysis. See Beijering (2007) for the effect of lexical distance on intelligibility and perceived distance of closely related languages.

[7] The intelligibility experiment was carried out in 2006. In 2007 the investigation was extended to perceived distance as well. Therefore, it was not possible to test the same high school pupils.

[8] The correlation between the intelligibility scores and the perceived distances was (r=-.65 , p<.01). This moderately strong correlation indicates that perceived distance and intelligibility scores are two different measurements that cannot be equated with each other. In other words, the question how *(dis)similar* another language variety is to one's own variety, is not the same as how *intelligible* another language variety is.