Kaivapalu, Annekatrin & Maisa Martin. 2017. Perceived similarity between written Estonian and Finnish: Strings of letters or morphological units? *Nordic Journal of Linguistics* 40(2), 000–000.

4 figures
5 tables + 1 in Appendix
1 footnote – auto-tagged
Appendix – includes 1 table

<RECTO>

<CT>
Perceived similarity between written Estonian and Finnish: Strings of letters or morphological units?

<CA>
Annekatrin Kaivapalu & Maisa Martin

<ABSTRACT>
The distance or similarity between two languages can be objective or actual, i.e. discoverable by the tools and methods of linguists, or perceived by users of the languages. In this article two methods, the Levenshtein Distance (LD), which purports to measure the objective distance, and the Index of Perceived Similarity (IPS), which quantifies language users' perceptions, are compared. The data are the quantitative results of a test measuring conscious perceptions of similarity between Estonian and Finnish inflectional morphology by Finnish and Estonian native speakers ('Finns' and 'Estonians') with no knowledge of and exposure to the other ('target') language. The results show that Finns see more similarity between Finnish and Estonian than Estonians do. Also the correlations between LD and the perception results of the Finns are statistically significant while the correlations between the LD and the IPS scores of the Estonians are not. Comments by test participants provide insights into the nature of the perceptions of similarity.

<Keywords> Estonian, Finnish, inflectional morphology, measuring actual and perceived cross-linguistic similarity

<ADDRESSES>

Annekatrin Kaivapalu, Finnish and Finno-Ugric Languages, University of Turku, FI-20014 University of Turku, Finland. annekai@utu.fi

& School of Humanities, Tallinn University, EE-10120 Tallinn, Estonia.kaivapa@tlu.ee

Maisa Martin, Department of Language and Communication Studies, University of Jyväskylä, P.O. Box L 35, FI- 40014 University of Jyväskylä, Finland. maisa.martin@jyu.fi

<HA>1. INTRODUCTION

Exploring the distance between languages has its roots in the areas of dialectology and general linguistics, particularly in language typology (e.g. Herlin & Kotilainen 2004, Kolehmainen, Miestamo & Nordlund 2013). It can be used as a tool for establishing relationships between languages as they are now, and in enquiry into historic developments. The object of such studies is the language system, so objective measures for the actual distance are sought. In the area of second or foreign language acquisition the distance between languages is also of interest, but for different reasons: it is often seen as an explanatory factor in language learning. The linguistic distance between the first language (L1) and the language to be learnt (L2) may be established by using the same measures as in language typology (Schepens, van der Silk & van Hout 2013). However, the distance or similarity which learners perceive, the psycho-typological distance, differs from the objective distance both quantitatively and qualitatively (Ringbom 2007, Jarvis & Pavlenko 2008). Language learners do not seek to grasp the whole linguistic system but want to make sense of what they see or hear. For them the distance between languages is an obstacle and similarity is an aid.

Measuring the degree of similarity between two whole languages is extremely difficult and subject to many sources of error even in the era of automatic parsing programs and large language

corpora (Heeringa et al. 2006). Even very small and limited language areas such as noun inflection, the focus of this article, are hard to compare. Reliable methods for quantifying the distance between languages are scarce (Schepens et al. 2013), although some advances have been made (see e.g. lexicostatistical studies by Gray & Atkinson (2003) and Holman et al. (2008).

The research presented in this article is part of the Receptive Multilingualism (REMU) network (https://www.uef.fi/web/remu2015), which conducts research on two closely related languages, Estonian and Finnish, with the aim of creating a holistic model of the effects of perceived and actual similarity on mutual intelligibility. The current authors concentrate on defining and measuring similarity while other network members work on mutual intelligibility (see e.g. Kaivapalu & Muikku-Werner 2010; Muikku-Werner & Heinonen 2012; Paajanen & Muikku-Werner 2012; Muikku-Werner 2013, 2014a, b; Kaivapalu 2015) and receptive multilingualism in practice (Härmävaara 2013, 2014). In this article we view the distance between morphological forms of Estonian and Finnish through the eyes of (potential) learners, people who have no previous knowledge of the target language (TL). These findings are compared with what can be found using objective measurements.

Alphabetic writing systems make us see written languages as strings of letters which form words and sentences. When we try to decipher words of a language new to us we attempt to read them letter by letter to see if the resulting words might resemble something we recognize from other languages, carrying a meaning that might fit the context. The ability to read even affects the phonological skills which help us remember and repeat words we hear, making it harder for non-literate learners to notice linguistic forms and learn new words (see e.g. Reis & Castro-Caldas 1997, Dellatolas et al. 2003). The image of words as strings of letters has a strong influence on how literate language users perceive language.

Against this background, it is not surprising that the surface distance between two languages is often measured by comparing strings of letters, using what is called the Levenshtein algorithm. It is based on the number of different, missing or additional letters (or sounds, if spoken languages are compared), usually in relation to the total length of the words. There are several versions of the Levenshtein Distance (LD) measure (see e.g. Beijering, Gooskens & Heeringa 2008, Wichmann et al. 2010); the one used in this article (see Section 4) is the one also used in Heeringa et al. (2013, 2014) as it is the most suitable one for comparing morphology. The results are compared with perceptions of similarity by language users (see Kaivapalu & Martin 2014, also Section 4 below). As the perceptions of similarity differ from the Levenshtein Distance, we focus on what factors might cause this discrepancy. We discuss alternative ways of describing the perceived similarity, looking at the phonological and morphological closeness of the languages and on some background factors such as the social, geographical, historical and paradigmatic variety of participants' L1s. The languages discussed here are Finnish and Estonian, two closely related languages with fairly transparent alphabetic orthography. The linguistic level chosen for comparison is noun morphology. Both languages inflect nouns extensively and the two inflectional systems share many features, described in more detail in Section 3 (for examples see also e.g. Remes 2009). Both are basically agglutinative languages but a major difference is the more central role of fusion in Estonian. A large proportion of the lexicon is also of common origin, even if some historical sound changes often obscure immediate perceptions of similarity.

Speakers of Finnish and Estonian are aware of the fact that the two languages resemble each other, i.e. they expect to be able to understand at least something when they encounter L2 speech or writing. Thus when they read a word, they are likely to assume that the same word might exist in their L1. They look for similarity. But what is the conscious perception of similarity based on? The overall aim of this article is to explore this issue: are consciously processed similarity perceptions based

simply on similarities between strings of letters (or sounds in speech, but in this study only written material is included) or on the perceived similarity of linguistic units, such as words, stems, or formatives. The question is obviously related to to an issue of broad concern in language in general, that is, the relationship between surface forms and meaningful units.

The LD assumes symmetry between the languages being compared. In our earlier research (Kaivapalu & Martin 2014:305−308) we have shown that the perceived similarity is not symmetrical: Finns see more similarity in the same set of words than Estonians do. This is also true of the larger set of data of this study. Asymmetry has been found in other language pairs, too (see e.g. Moberg et al. 2006). In this article we search for potential causes of the asymmetry.

One way to approach the differences between actual and perceived distance or similarity between languages is calculating the rank correlation of the results achieved by methods designed to measure the two types of differences (e.g. Jarvis 2016, Letica Krevelji 2016). However, the correlations found in this way are not always qualitatively equal. One reason is the asymmetry mentioned above. Deletions and insertions may also not have an equal role in perceptions. Probing the qualitative differences is the main contribution of this article.

The LD can be calculated for word stems and formatives separately. As the formatives of Estonian are harder to perceive as independent units, due to more extensive fusion, we will also explore these differences to explain the asymmetry. The detailed analysis of the effect of the stem and formative relations will also contribute to the comparison of the two measures themselves. The research questions are:

<NL>

1. While the LD can claim objectivity and measure 'actual' distance or similarity between two sets of linguistic items, to what extent does it explain the perception of the speakers of the two languages?
2. Are insertions and deletions really of equal value, as assumed in the calculations of the LD? How do insertions and deletions relate to the perceived asymmetry between the speakers of Estonian and Finnish? Although an experimental exploration of insertions and deletions is not possible within this article, we will suggest how this could be done.

<HA>**2. BACKGROUND**

The construct and taxonomy of similarity as well as some tentative results of our similarity perception measure (Index of Perceived Similarity, IPS) were discussed in Kaivapalu & Martin (2014). As suggested in the introduction above, similarity can be objective or actual on the one hand and perceived, on the other (Ringbom 2007:7–8, 24–26; Jarvis & Pavlenko 2008:176–182). Learners' reliance on perceived similarity can also be divided into item and system (procedural) (Ringbom 2007:54–58), or process similarity (Martin 2006, Kaivapalu & Martin 2007), depending on the domain of the study. ITEM SIMILARITY refers to the similarity of individual forms, such as sound, letter, morpheme, word or phrase, whereas SYSTEM SIMILARITY is defined as a set of principles for organizing forms paradigmatically and syntagmatically (Ringbom & Jarvis 2009:108–109). Our test is intended to measure item similarity, but the comments of some participants indicate that they do compare the items to other items in the test, thus introducing some aspects of system similarity, e.g. the paradigmatic nature of the inflectional system (see also Kaivapalu 2005:267–272), in their perceptions.

The degree of similarity also varies (Kaivapalu & Martin 2014:291). Items to be compared can be identical or bear a closer or more remote resemblance to each other. The distance between two items can be calculated using, for instance, LD, and the items can thus be ranked at a certain distance

from each other, providing that the properties included in the calculations are valid. Averages of perceived similarity scores given by a sufficient number of participants can also be ranked and the two lists compared, as is done below. Such scores, however, are more dependent on the properties of the task as a whole, the background of the participants, the scoring system etc. The potential for error in each of the methods of measuring similarity is discussed in Section 6 of this article.

Similarity or distance across many language pairs has been studied before. The Levenshtein algorithm has been used in the large European Micrela project (van Heuven, Gooskens & Van Bezooijen 2015), which studies the mutual intelligibility and distance of several Germanic, Romance, and Slavic languages (Heeringa et al. 2013), and particularly by Heeringa et al. (2014) to compare Germanic languages. Another measure of the distance or similarity between languages is the conditional entropy, which measures the amount of regularity in the sound or grapheme correspondences between the two languages (Frinsel et al. 2015). Most studies compare whole words or longer units but the effects of stems and affixes have also been studied before (Heeringa et al. 2014).

The results of such measures of the surface distance are often compared to overall similarity perceptions of languages or dialects, either without determining what exactly is supposed to be similar (e.g. Letica Krevelji 2016) or concentrating on phonological or semantic or functional similarity (e.g. Tang & van Heuven 2015, Jarvis 2016). Morphological similarity has been focused on before mainly for Germanic languages (Heeringa et al. 2009, 2014; Heeringa & Hinksens 2011). The advantage of the similarity test used here (see Section 4 below) is that it does not ask about similarity in general but makes the participants compare given forms, in the knowledge they have been given that they have the same function and represent the same grammatical form. The disadvantage is, of course, that it covers only one area of language, noun inflection. Also, the connection to mutual intelligibility may be weaker than in tests which search for overall similarity.

Explanations for the difference between objective and perceived similarity and for the asymmetry of perceptions have been sought in exposure to dialects (Delsing & Lundin Åkesson 2005, Gooskens 2006, Berthele 2008, Gooskens & Heeringa 2014), with the idea that extensive variation in the L1 would help in seeing similarity in a closely related L2. This is one of the explanations explored also in this article. Another causal relationship can be found by charting the time the participants spend exposed to the L2 by reading or watching TV, etc. (e.g. Delsing & Lundin Åkesson 2005). It is hard to find Finnish participants who have never heard Estonian spoken on TV or seen a few words written on packages etc. The same is true of Estonian people: Finnish is occasionally present in their lives, albeit not very noticeably. The older generation of Northern Estonians was exposed to Finnish TV during the Soviet era but this is not the case with the participants of this study, who are nearly all young adults. We excluded potential participants who had studied the TL or lived in the TL country.

Yet another sometimes quoted explanation for the results is language attitude (Gooskens 2006). With regard to the languages in our study, it has been found that the attitudes of young people in Finland and Estonia to each other and each other's language are generally either fairly neutral or positive (Junttila 2006:151–152).

<HA>3. FINNISH AND ESTONIAN NOUN MORPHOLOGY

Estonian (Est.) and Finnish (Fin.) are basically agglutinative languages, which indicate grammatical roles as well as semantic information by inflectional suffixes. Noun stems are followed by markers for number (0 for singular, *t*/*d* or *i*/*j* for plural) and case endings, e.g. Fin. *talo-i-ssa* 'house-PL-INE', Est. *talu-de-s* 'farm-PL-INE'.[1] The number of cases varies somewhat, depending on the way they are defined, but the most commonly quoted number for both languages is 14. For Finnish, the accusative forms that only exist for some pronouns are not counted (EKG I:§48; Viitso 2003a:32–34; ISK:§81; Karlsson 2015:3). The case inventory is the same for both languages apart from the instructive (only

in Finnish) and the terminative (only in Estonian). The case ending may be followed by a possessive suffix in Finnish and/or by another clitic (in both languages), e.g. Fin. *talo-i-ssa-mme-kin* 'house-PL-INE-POSS.1PL-CLI', Est. *talu-de-s-ki* 'farm-PL-INE-CLI'. In this study noun forms in the singular or plural and in a variety of cases are explored. The combination of the plural marker and a case ending is here called an inflectional formative.

In the agglutinative process the stem and the suffixes interact (Karlsson 1983:277–304). The stem may undergo consonant gradation (ISK:§41; Karlsson 2015:30–40; EKG I:144–170; Viitso2003b:196–198), e.g. Fin. *mu**rr**e* 'dialect.NOM.SG' : *mu**rt**e-i-ssa* 'dialect-PL-INE' or Est. *mu**rr**e* 'dialect.NOM.SG': *mu**rd**e-i-s* 'dialect-PL-INE' or other less regular changes of consonants or vowels, e.g. Fin. *saun**a*** 'sauna. NOM.SG' : *saun**o**-i-ssa* 'sauna-PL-INE', where the stem-final *-a-* is replaced by *-o-* before the plural *-i-* (Karlsson 2015:40–44) or Est. *puna**ne*** 'red.NOM.SG' : *puna**se*** 'red.GEN.SG' where *-ne* in the nominative singular is replaced by *-se* in the genitive singular and all other cases except the nominative singular (EKG I:175–176). Some of these changes are shared by both Finnish and Estonian, some are language specific.

A major difference between the noun morphology of the two languages is the Estonian so-called stem plural (EKG I:208–214; Viitso 2003a:37), an inflected form with no formative, e.g. *lub**a*** 'permission. NOM. SG.': *lub**e*** 'permission. PAR.PL.'. In such cases, mostly in the partitive plural and more rarely in other plural case forms, the stem vowel is replaced by another vowel. Another difference between the Estonian and Finnish case systems is the Estonian genitive singular, which has no case ending, e.g. *tuba* 'room. NOM. SG.': *toa* 'room.GEN.SG'. The case function of genitive is represented only by a special stem allomorph (Viitso 2003a:33). Such forms do not exist in Finnish and can be predicted to be confusing for Finns. They are at one extreme of a continuum between fully agglutinative and fully fusional inflectional forms. The test words represent various stages on this continuum (Figure 1).

<Insert Figure 1 about here>

Figure 1 does not describe historical relationships but current surface forms. While many stems and most of the inflectional material in Estonian and Finnish are of the same historical origin, sound changes have rendered many items different. The same material may also exist in both languages but be distributed in a different way. A good example is provided by the plural markers *-i/j-* and *-t(-te-/-de-)*, which are used in both languages but the *i*-plural is much more common in Finnish and the *t*-plural in Estonian (Remes 2009:94–95).

<HA>**4. DATA AND METHOD**

The two ways of measuring similarity between linguistic forms discussed in this study are the Levenshtein Distance (LD) (Heeringa et al. 2013, 2014) for objective similarity and the Index of Perceived Similarity (IPS) developed by the authors (Kaivapalu & Martin 2014). The data are the quantitative LD and IPS results from a list of test items targeting the variation in actual similarity between Estonian and Finnish inflectional morphology. The test contains 48 pairs of Estonian and Finnish words, chosen on the basis of a historical-comparative and typological analysis of the two languages (Remes 1995, 2009). The test word pairs share the same stem (i.e. the same stem with the same meaning exists in both languages but may have a somewhat different surface form) and represent the same inflectional form (case and number), again with variance in the surface form. The list of test pairs is divided into four similarity categories (mixed in the actual test): (i) similar stem, similar inflectional formative; (ii) similar stem, different formative; (iii) different stem, similar formative; and (iv) different stem, different formative. The hypothesis is that the first category would show the highest degree of similarity, the last category the lowest. When the list is presented to the participants, the first member in each pair is given to the Estonians in Estonian and to the Finns in Finnish, with the familiar language thus offered as the basis for comparison.

Four test groups, students in different universities in Estonia and Finland, were set up: two groups of L1 speakers, one of Finnish (n = 115) and one of Estonian (n = 109), and two groups of L2 speakers of these languages, their L1s Swedish (L2 Finnish, n = 105) and Russian (L2 Estonian, n = 80) respectively. None of the participants had studied the target language (Estonian for Finns, Finnish for Estonians) or lived in the target language country, although most of them were very likely to know that Finnish and Estonian are closely related and some had visited the country of the target language briefly as tourists. Due to limited space, this article deals only with the perceptions of L1 speakers, focusing on the symmetry of perceived similarity; for the perceptions of L2 speakers, see Kaivapalu & Martin 2016.

In the IPS test the participants were asked to rate the 48 pairs of inflected words as similar, somewhat similar, or different. They were also given the opportunity to comment on their choice in writing. The test answers were collected on a computerized form. The IPS score of each test item was then calculated by giving two points for each 'similar' answer, one for 'somewhat similar' and none for 'different' and by adding up all of the points. The results are shown in Table 1 below.

The Levenshtein Distance refers to the 'cost' of moving from one item to the other in the comparison. In this study, the LD is – as usual – implemented as a symmetrical measure, i.e. the distance is considered to be the same regardless of the direction of comparison. The cost is determined by aligning the two items, matching a vowel with a vowel and a consonant with a consonant and then calculating the cost of changing one string for the other (Heeringa et al. 2013). An example from our data, 'rain-INE', is seen in Figure 2.

<Insert Figure 2 about here>

Every difference, insertion, and deletion within the two strings is weighed as 1. The result (here 5) is then divided by the total length of the string (here 10). In the example (Figure 2) the LD is thus 0.5 or 50%. LD refers to surface distance, calculating only what can be seen, without making any

functional assumptions: a different letter means a difference. LDs were calculated for the morphological form as a whole, for the stem, and for the inflectional formative of every test word pair (Heeringa et al. 2014). LD and IPS values were correlated using IBM SPSS version22. In cases of significant correlation between LD and IPS, multiple linear regression analyses (Field 2009) were applied.

<HA>5. RESULTS

In this section, the test results are discussed by comparing the LDs of the test items with the IPS scores of the participant groups in the test as a whole and in the four morphological categories described above. The LDs and IPS scores of individual test items can be found in appendix Table A1. Table 1 presents the total number of word pairs (balanced for the number of participants) rated by the Finns and Estonians as similar, somewhat similar or different, and the IPS for the Finns and Estonians (on calculating IPS, see Section 4 below).

<Insert Table 1 about here>

While the LD is always symmetrical across languages, the IPS need not be. The results show that the Finns see more similarity between Finnish and Estonian inflectional morphology than the Estonians do: the results reveal a statistically significant difference ($p < .001$) between the IPS of the Finns and Estonians. The Finns have also answered similar and somewhat similar more often than different, while the Estonians have perceived most test word pairs as somewhat similar and have chosen different more often than similar. So, the perceptions of the two groups are not symmetrical. To explain the asymmetry of the similarity perceptions of the two participant groups, the LDs of every individual test item were calculated for the morphological form as a whole, for the stem and for the inflectional formative (see Appendix) and then the LD and the IPS values were correlated. A significant correlation ($r = –.751$, $p < .01$) between the LD and IPS of morphological form as a whole was found for Finns (Figure 3). The negative value of r indicates the inverse relationship

between the LD and the IPS: as the LD indicates distance rather than similarity, the lower the LD and the higher the IPS are, the more similar are the word pairs. For the Estonians, no correlation (r = .079, p > .05) was found between the LD and the IPS, as can be seen in Figure 3. Also, the correlations between the LDs of the stems and the morphological formatives and the IPSs of the Finns are statistically significant (r = –.638, p = .000 and –.468, p = .001 respectively) while the correlations between the LD and the IPS scores of the Estonians are not (r = –.120, p = .417 and –.240, p = .100 respectively). This will be explored further in order to find explanations for the asymmetry and variation in the results by comparing the correlations broken down by morphological category and also by examining the comments made by the participants.

<Insert Figure 3 about here>

<Insert Figure 4 about here>

To find out the LD's predictability for the IPS in cases of significant correlation (the results of the Finns), regression analyses were applied (Table 2). According to the multiple linear regression analysis, the best predictor for the similarity perceptions of the Finnish group is the LD of the morphological form (the combinaton of stem and formative) as a whole; this explains 56% of the similarity perceptions. The LD of the stem predicts 40% and the LD of the formative only 21% of the similarity perceptions. The impact of other factors on perceiving similarity is addressed in Section 6 below.

<Insert Table 2 about here>

To investigate the relations between the LD and the IPS in more detail and also for the Estonian group, these relationships were analyzed by similarity categories: the word form pairs with (i) similar stem and similar formative, (ii) similar stem but different formative, (iii) different stem but similar formative, and (iv) different stem and formative, in Finnish and Estonian (Table 3). According to a paired samples t-test, a statistically significant difference between the IPS of the Finns and Estonians was found for three similarity categories: word form pairs with similar stem and similar formative (t =

12.812, $p = .000$), different stem but similar formative ($t = 2.880$, $p = .016$), and different stem and formative ($t = –2.679$, $p = .021$). For the second similarity category, word form pairs with a similar stem but a different formative, no statistically significant difference was found ($t = .145$, $p = .887$). The results show that the Estonians perceive almost as much or more similarity than the Finns do in word form pairs with a higher LD, which indicates a bigger difference in word forms. Especially word forms with a higher LD of the morphological formative have been perceived more similar by the Estonians than by the Finns. These results indicate that the Estonians tend to see more similarity than the Finns do in the more different forms.

<Insert Table 3 about here>

Looking at the individual word form pairs where the Estonians perceived more similarity than the Finns (Table 4), the results indicate two main tendencies. The first one concerns the typologically different partitive plural forms in Estonian and Finnish: the stem plural in Estonian and the agglutinative *i*-plural in Finnish. The second tendency concerns the word pairs where there is a great difference either in the stem or in the plural and case formative in the two languages. The results indicate that the Estonians are better able to see through the surface differences that are due to the discrepancies in the inflectional systems of the two languages. Further proof for this interpretation is provided by the Estonians' comments in Section 6.

<Insert Table 4 about here>

The second research question addresses the relationship of differences, insertions, and deletions to similarity perceptions. In many test word pairs there is something added from the point of view of the Estonians and something deleted from the point of view of the Finns (see Table 5), either at the end of the word (e.g. *raamatu-st* 'book-ELA'– *raamatu-st**a*** 'bible-ELA', *sinis-te* blue-PL'– *sinis-te-n* 'blue-PL-GEN') or in the middle and at the end of the word (e.g. *vangla-s* 'prison-INE'– *vankila-ssa* 'prison-INE'). In most of these word pairs the Finns see systematically more similarity than the Estonians.

<Insert Table 5 about here>

There are some word pairs where the differences between the Finns and the Estonians are very small (Table 5), as Est. *nobeda-i-le* 'speedy-PL-ALL' and Fin. *nope-i-lle* 'speedy-PL-ALL' or Est. *lään-de* 'west-ILL' and Fin. *länte-en* 'west-ILL'. In the word pairs *nobeda-i-le* – *nope-i-lle* 'speedy-PL-ALL' and *sademe-i-s* 'rain-PL-INE' – *sate-i-ssa* 'rain-PL-INE', where from Finns' point of view there is an additional syllable, the Estonians perceived more similarity. Also in the word pairs *järve* 'lake.ILL' – *järve-en* 'lake-ILL' and *pesa* 'nest.PAR' – *pesä-ä* 'nest-PAR' the IPS scores of Estonians are higher. The results provide tentative evidence that deletions and additions are, contrary to what is assumed in the calculations of the LD, not of equal value for similarity perceptions: according to the IPS scores, deletions seem to be more transparent in terms of perceiving similarity than additions. Further evidence for this is provided in the next section. In the word pairs in Table 5, several factors discussed above overlap: the stems are similar but there are considerable differences in the formatives, or the formatives are missing, which is particularly confusing for the Finns.

<HA>**6. DISCUSSION**

In this section, the results presented above are discussed in the light of the comments written by the participants when completing the test. All the comments were collected and classified, and the following categories were found: social, geographical, historical (examples(1a–e) below) and morphological variation, as in (2a, b); awareness of the morphological structure of words (stems and formatives), as in (3a–i); strings of letters, phonology (pronunciation), as in (3e, h); and paradigmatic awareness, as in (4a–c). The comments have been translated into English, with the original language in brackets. For the glosses of the words cited in the comments, please see Table A1 in the Appendix.

The first research question deals with the relationship between the 'actual' distance or similarity between two sets of linguistic items, here operationalized by the use of the LD, and perceived similarity, here the IPS scores. The results confirm what has been shown in previous research, that

unlike the LD, the IPS is not symmetrical: Finns see more similarity than do Estonians. One reason may be the wider variation in Finnish. The five million speakers of Finnish are spread over an area many times larger than Estonia and speak many more dialects in their everyday life than Estonians do. The difference between the standard and spoken language is bigger in Finnish than in Estonian, and many features of spoken Finnish are similar to the Estonians' standard language. Finns are thus exposed to several morphological forms for one function, which may enable them to see more similarity between a given Finnish form and the corresponding Estonian form. Exposure to dialects has also been suggested elsewhere as a factor for mutual intelligibility (e.g. Delsing & Lundin Åkesson 2005, Gooskens 2006, Gooskens & Heeringa 2014). It is also given as a reason for the choice 'similar' in the comments in our test (here translated from Finnish or Estonian, with the original language indicated in parentheses), as in (1a–c):

<NL>

(1) a. Dialect-like Finnish. (Finnish)

b. The Estonian word sounds like spoken Finnish. (Finnish)

c. Sounds like a Finnish slang expression. (Finnish)

Although there is less regional variation in Estonian – because of the smaller number of speakers in a much smaller area and possibly also because of a heavier emphasis on standard language in the educational system – references to dialects can also be found in comments made by the Estonian group, as in (1d–e):

<NL>

(1) d. Resembles some Estonian dialects. (Estonian)

e. Could be in South-Estonian dialect. (Estonian)

Both groups thus search for sources of similarity in language-internal variation, but the Finns do so much more frequently than the Estonians. Dialects are most frequently mentioned, but slang and spoken language are also quite commonly mentioned. References to old forms also occur, indicating knowledge of historical variance. Occasionally also unrelated words or words from another language are discussed, but usually in connection with perceived difference rather than similarity.

In addition to dialects and other variants, internal variation in the standard language is mentioned as in (2a–b).

<NL>

(2) a. In Finnish one can also say *saarten*, this adds to the similarity. (Finnish)
(Fin. *saarien* – Est. *saarte* 'island-PL-GEN')

b. *Laintest* – plural, one can also say *laineist* – the same as in Finnish. (Estonian)
(Fin. *laineista* – Est. *laintest* 'wave-PL-ELA')

Alternative forms seem to be a good route for perceiving similarity. Another interesting finding is that the correlation between the LD and the IPS is significant for the Finnish group but there seems to be no such correlation for the Estonian group. This can be interpreted as suggesting that the Finns pay more attention to strings of letters, the same aspect of similarity that the LD measures, while the Estonians more often compare morphological forms rather than surface similarity. Estonians are also more likely to give explanations which display linguistic knowledge, exemplified in (3a–g):

<NL>

(3) a. The stem of the word is the same. (Estonian)

b. Inflectional endings *-te-* and *-ten* are similar. (Estonian)

c. It seems that the basic form of the word is similar: *kolmas* – the formative of the inessive case is added (Estonian)

d. Only the plural suffixes differ (Estonian)

e. Weak and strong *t* is still the same, and the pronunciation can for example be the same (*kalad – kalat*) (Estonian)

f. Basically the same inflectional ending. (Estonian)

g. The inflectional endings sound different: in Estonian illative, in Finnish inessive. (Est)

Comments employing linguistic knowledge are rare and less explicit among the Finns:

<NL>

(3) h. At the end of words one can find a similar inflectional ending and they sound similar. (Finnish)

i. A similar word, but the inflectional form disappears (*lehtiä – lehti*). (Finnish)

The differences in the comments reflect the Estonians' better ability to analyze the forms beyond the surface strings. It is hard to say whether or not this is due to the less agglutinative nature of Estonian, which requires speakers to interpret function even when there is no surface formative, while in Finnish each morphological function is expressed by an affix (even if the combinations of affixes sometimes fuse into formatives which are not easy to dissect). It is also possible that the explicit teaching of morphology in Estonian schools causes this; certainly it helps the participants to express themselves in linguistic terms.

Finnish agglutinative forms include parts that are familiar to Estonians, even if they may not be used in the context of a particular test word. This may help the Estonians to see the most different word pairs as more similar than Finns do. Estonians are also used to the variation in stem vowels common in Finnish plurals. Other stem vowel changes often seem to throw the Finns off course.

Another problematic issue for the Finns are the stem plurals, which completely lack a formative. This is not surprising, as it deviates from the one morpheme – one meaning principle typical of Finnish. Also *t*-plurals are easier for the Estonians to see as similar to Finnish *i*-plurals. Both exist in both languages but they differ in their distribution (see Section 3). The greater frequency of the *t*-plural

in Estonian and many alternative forms apparently make it easier for the Estonians to see the underlying functional similarity.

In spite of all the advantages for seeing similarity enjoyed by the Estonians and listed above, the Finns clearly see much more similarity than the Estonians do. This links to the second research question, which asks whether deletions and insertions are of equal value in perception. Finnish word forms often contain material which is not present in the corresponding Estonian form, most commonly at the end of the word, within the inflectional formative, but also sometimes within the stem (Table 4). For this reason, Finns looking at Estonian words see something lacking, while Estonians see something unnecessary in the Finnish words. The IPS results of this group of test words show that the fact that something is missing from the viewpoint of L1 is a smaller obstacle to perceiving similarity than the situation where there is something extra. This is also reflected in the comments, which often refer to spoken Finnish or the dialect of the Turku region, both well-known for dropping the final vowel. Dropping the *-n* indicating the genitive form is also very common in spoken Finnish. Accepting missing sounds or letters seems easy for the Finnish participants.

The Estonians, however, seem to find it harder to see similarity when the Finnish word contains material which is not present in the Estonian word. It is possible that additional material provokes the participants unconsciously to search for a function for the extra material while missing material does not trigger a similar response. This is a very tentative suggestion, which needs to be further investigated.

The research questions did not address the issue of item vs. system similarity, since in the test the word pairs were presented one by one and we assumed that they would also be compared only within each pair. Some comments, however, reveal that at least some participants compared word pairs to other word pairs in the list or to other morphological forms of the same word, as in (4a–c):

<NL>

(4) a. Again -*aid*, in other words plural. (Finnish)

b. Estonian nominative singular *lind* Finnish *lintu*(?) ‘bird’. (Estonian)

c. Not a well designed test: it is impossible to understand what has to be compared – this particular word form or also all other forms of the word. (Finnish)

Especially the last comment reveals the participant’s awareness of the paradigmatic character of the inflectional system of her/his L1, Finnish (Paunonen 1983:59). This search for some systematicity is also a natural way to approach an unknown language (see also about applying paradigmatic analogy in the production process, Kaivapalu 2005:267–272). Regardless of whether a language is seen as a set of rules a learner has to acquire or as a multitude of constructions from which regularities emerge, learners are aware of the fact that language items bear some relation to each other.

<HA>**7. CONCLUSIONS**

A simple answer to the question posed by the title of this article, ’Perceived similarity between written Estonian and Finnish: Strings of letters or morphological units?’, is that it depends on the L1 of the participant and on the task. The Finns see more similarity than do the Estonians, and more than half of that similarity is explained by similarity in the strings of letters. No correlation was found between strings of letters and the similarity perceptions of the Estonians. Other sources for perceiving similarity were explored by analyzing the participants’ comments and by comparing the morphology of the two closely related languages.

Comments that explain some of the findings mostly refer to variation in the L1, the language the participant knows well. Another large group of comments are those that display linguistic knowledge, with or without linguistic terminology: stems, endings, rules, sound changes etc. are mentioned as an explanation for similarity or difference. The prevalence of comments varies by group. The L1 Finnish group takes advantage of the variety in Finnish: regional dialects, spoken colloquial forms, slang, and

old or literary forms provide them with more potential for comparison than the Estonian group employs. The Estonians, on the other hand, often refer to morphological knowledge and use more linguistic terminology, probably due to the inclusion of morphological analysis in the school curriculum. The comparison of forms letter by letter is infrequent in the L1 groups, but the fact that it occurs at all reinforces the statistical results: surface forms matter.

The task itself obviously directs participants' attention and influences the results. The task in this study was carefully designed to contain a balance of the different types of relationships between Estonian and Finnish morphological forms. This proved essential, as the similarity perceptions of the Finns and the Estonians differed by the type of difference present in the word pair. The Finns found pairs with (from their point of view) missing letters similar, while Estonians were more able to see through the surface string in word pairs where fusion and agglutination were being compared. We also tried to exclude semantic and functional considerations by emphasizing that both parts of the word pair to be compared had the same meaning and represented the same grammatical form (case and number). A factor we were not able to keep constant was the frequency of the words in each language, but apparently all the words were familiar to the participants, as this did not give rise to any comments.

The ability to perceive similarity between linguistic items depends on linguistic awareness and a conscious processing of language. Most comparisons are item-based, as was expected, but paradigm-level considerations, comparing words within paradigms, employing knowledge of forms outside the test, were also present, as were comparisons within the list of word pairs. Word forms do not exist in a vacuum but invoke other forms and other words. Item similarity overlaps with system similarity.

The test was not systematically designed to explore the issue of the (un)equal value of deletions and insertions. It is actually surprising that users of the LD or other measures of linguistic distance have shown so little interest in this question. In this study, where it is only relevant to some parts of the test, the issue still explains a substantial part of the differences between the groups. Estonians who

encounter insertions find it harder to see similarity than Finns who encounter deletions. This finding invites new research. We need a test that concentrates on deletions and insertions in a way that excludes the alternative explanation in this test, the influence of spoken Finnish and dialects (e.g. the possibility of dropping the final *-n* in spoken Finnish). The hypothesis that additional material in one of the words to be compared prompts participants unconsciously to search for a function for the extra material needs to be tested, although it is not easy to find a way of doing this. It would probably require multiple methods. We also need to explore the roles of conscious and unconscious processing, not only for this issue but for perceiving similarity in general. A reaction time test is already in progress.

If the final goal of the study of similarity across languages is to explore mutual intelligibility or help people learn each other's languages, as it is in the REMU project, it really is the perceptions of language users that matter rather than objective distance. In this article only a small area of language is targeted but the results raise interesting questions and can provide advice and ideas for testing perceived similarity in other areas of related languages.

<HA>**ACKNOWLEDGEMENTS**

Marianna Penttilä and Kärt Kaivapalu helped to sort out the comments in all four languages. Marianna Penttilä, Ingrid Krall and Anastassia Kallikorm helped in collecting the data. Ats Kaivapalu and Scott Jarvis performed the regression analysis. Many thanks to all of them. We are also very grateful to the three anonymus reviewers for their valuable comments.

<HA>**APPENDIX**

<Insert Table A1 exactly here>

<HA>**NOTE**

<Insert endnote here, as style – cpy at the end of the present file>

<HA>**REFERENCES**

Beijering, Karin, Charlotte Gooskens & Wilbert Heeringa. 2008. Predicting intelligibility and perceived linguistic distance by means of the Levenshtein algorithm. *Linguistics in the Netherlands 2008*, 13–24.

Berthele, Raphael. 2008. Dialekt-Standard Situationen als embryonale Mehrsprachigkeit. Erkenntnisse zum interlingualen Potenzial des Provinzlerdaseins. *Sociolinguistica* 22, 87–107.

Dellatolas, G. Nikolaus, Willadino Braga, Ligia do Nascimento Souza, Gilberto Nunes Filho, Elizabeth Queiroz & Gerard Deloche. 2003. Cognitive consequences of early phase of literacy. *Journal of the International Neuropsychological Society* 9(5), 771–782.

Delsing, Lars-Olof & Katarina Lundin Åkesson. 2005. *Håller språket ihop Norden? En forskningsrapport om ungdomars förståelse av danska, svenska och norska* [Does language hold the Nordic Countries together? A research report on the comprehension of Danish, Swedish, and Norvegian among young people.] [. Copenhagen: Nordiska ministerrådet.

EKG I 1995 = *Eesti keele grammatika I. Morfoloogia. Sõnamoodustus* [Estonian grammar I: Morphology. Word formation], edited by Mati Erelt, Tiiu Erelt, Henn Saari & Ülle Viks. Tallinn: Eesti Teaduste Akadeemia Eesti Keele Instituut.

Erelt, Mati (ed.). 2003. *Estonian Language*. Tallinn: Estonian Academy Publishers.

Field, Andy. 2009. *Discovering Statistics Using SPSS*. London: SAGE Publications.

Frinsel, Felicity, Anne Kingma, Femke Swarte & Charlotte Gooskens. 2015. Predicting the asymmetric intelligibility among spoken Danish and Swedish using conditional entropy. *Tijdschrift voor Scandinavistiek* 34(2), 120–138.

Gooskens, Charlotte. 2006. Linguistic and extra-linguistic predictors of inter-Scandinavian intelligibility. *Linguistics in the Netherlands 2006*, 101−113.

Gooskens, Charlotte & Wilbert Heeringa. 2014. The role of dialect exposure in receptive multilingualism. *Applied Linguistics Review* 5(1), 247–271.

Gray, Russell D. & Quentin D. Atkinson. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* 426, 435–439.

Härmävaara, Hanna-Ilona 2013. Kielten samankaltaisuus monikielisen virolais-suomalaisen vuorovaikutuksen resurssina [Cross-linguistic similarities as a resource of multilingual interaction between Finns and Estonians]. *Lähivõrdlusi – Lähivertailuja* 23, 60–88.

Härmävaara, Hanna-Ilona. 2014. Facilitating mutual understanding in everyday interaction between Finns and Estonians. In ■■■■ (eds.), *Receptive Multilingualism*: Special issue of *Applied Linguistics Review* 5(1), 211–245.

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

Heeringa, Wilbert & Frans Hinksens. 2011. The measurement of Dutch dialect change: Lexicon versus morphology versus sound components. In Gunther de Vogelaer & Wilbert Heeringa (eds.), *Talige en buitentalige factoren bij regiolectvorming* [factors of regiolectsformation■■■■]: Special issue of *Taal en Tongval* 63(1), 79–98.

Heeringa, Wilbert, Peter Kleiweg, Charlotte Gooskens & John Nerbonne. 2006. Evaluation of string distance algorithms for dialectology. In John Nerbonne & Erhard Hinrichs (eds.), *Linguistic Distances Workshop at the Joint Conference of International Committee on Computational Linguistics and the Association for Computational Linguistics*, *Sydney, July 2006*, 51–62.

Heeringa, Wilbert, Femke Swarte, Anja Schüppert & Charlotte Gooskens. 2014. Modeling intelligibility of written Germanic languages: Do we need to distinguish between orthographic stem and affix variation? *Journal of Germanic Linguistics* 26(4), 361–394.

Heeringa, Wilbert, Martijn Wieling, Boudewijn van den Berg & John Nerbonne. 2009. A quantitative examination of variation in Dutch Low Saxon Morphology. In Alexandra N. Lenz, Charlotte Gooskens & Siemon Reker (eds.), *Low Saxon Dialects across Borders – Niedersächsiche Dialekte über Grenzen hinweg, ZDL-Beiheft 138, Dedicated to Prof. Dr. Hermann Niebaum*, 195–216. Stuttgart: Franz Steiner Verlag.

Herlin, Ilona & Lari Kotilainen. 2004. External factors behind cross-linguistic similarities. In Olga Fischer, Muriel Norde & Harry Perridon (eds.), *Up and Down the Cline: The Nature of Grammaticalization*, 263–279. Amsterdam: John Benjamins.

Holman, Erik W., Søren Wichmann, Cecil. H. Brown, Viveka Velupillai, André Müller & Dik Bakker. 2008. Explorations in automated language comparison. *Folia Linguistica* 42, 331–354.

ISK= *Iso suomen kielioppi* [Descriptive grammar of Finnish], edited by Auli Hakulinen, Maria Vilkuna, Riitta Korhonen, Vesa Koivisto, Tarja Riitta Heinonen & Irja Alho. 2004. Helsinki: Finnish Literature Society.

Jarvis, Scott. 2016. On the combined effects of crosslinguistic similarity, structural complexity, and semantic complxity on word learnability. Presented in Eurosla 26 Conference in Jyväskylä.

Jarvis, Scott & Aneta Pavlenko. 2008. *Crosslinguistic Influence in Language and Cognition*. New York & London: Routledge.

Junttila, Jaakko. 2006. ‘Koomista lasten kieltä’. Jyväskyläläisten ja tarttolaisten opiskelijoiden asenteista sukukieleen. [‘Comical child language’: About the attitudes to related language by students in Jyväskylä and Tartu.] In Annekatrin Kaivapalu & Külvi Pruuli (eds.), *Lähivõrdlusi – Lähivertailuja* 17, (Jyväskylä Studies in Humanities 53), 135–154.

Kaivapalu, Annekatrin. 2005. *Lähdekieli kielenoppimisen apuna* [Contribution L1 to foreign language acquisition] (Jyväskylä Studies in Humanities 44). Jyväskylä: University of Jyväskylä.

Kaivapalu, Annekatrin & Maisa Martin. 2007. Morphology in transition: The plural inflection of Finnish nouns by Estonian and Russian learners. *Acta Linguistica Hungarica* 54(2), 129–156.

Kaivapalu, Annekatrin & Maisa Martin. 2014. Symmetry of similarity: Definition, perception, measurement: Finnish and Estonian noun morphology as a testing ground. In Heli Paulasto, Lea Meriläinen, Helka Riionheimo & Maria Kok (eds.), *Language Contacts at the Crossroads of Disciplines*, 283–318. Cambridge: Cambridge Scholars.

Kaivapalu, Annekatrin. 2015. Eesti ja soome keele vastastikune mõistmine üksiksõna- ja tekstitasandil: lingvistilised tegurid, mõistmisprotsess ja sümmeetria [Mutual intelligibility of Estonian and Finnish context-free words and texts: Linguistic determinants, comprehension process and symmetry]. *Estonian Papers in Applied Lin*guistics, 55–74.

Kaivapalu, Annekatrin & Maisa Martin. 2016. Detection of morphological similarity between closely related L2 and L3. In Vetter (ed.), 36−37.

Kaivapalu, Annekatrin & Pirkko Muikku-Werner. 2010. Reseptiivinen monikielisyys: miten suomenkielinen oppija ymmärtää viroa äidinkielensä pohjalta? [Receptive multilingualism: How does Finnish as a first language help learners to understand Estonian?]. *Lähivõrdlusi – Lähivertailuja* 20, 68−97.

Karlsson, Fred.1983. *Suomen kielen äänne- ja muotorakenne* [Phonological and morphological structure of Finnish]. Helsinki: WSOY.

Karlsson, Fred. 2015. *Finnish: An Essential Grammar*, 3rd edn. (Routledge Essential Grammars). London & New York: Routledge.

Kolehmainen, Leena, Matti Miestamo & Taru Nordlund (eds.). 2013. *Kielten vertailun metodiikka*

[Methodology of comparing languages]. Helsinki: Finnish Literature Society.

Letica Krevelji, Stela. 2016. The L2 status vs. psychotypology… and beyond. In Vetter (ed.), 43.

Martin, Maisa. 2006. Suomi ja viro oppijan mielessä. Näkökulmia taivutusmuotojen prosessointiin [Finnish and Estonian in the mind of the learner: Approaches to processing inflectional forms]. In Annekatrin Kaivapalu & Külvi Pruuli (eds.), *Lähivõrdlusi – Lähivertailuja* 17 (Jyväskylä Studies in Humanities 53), 43–60.. Jyväskylä: University of Jyväskylä.

Moberg, Jens, Charlotte Gooskens, John Nerbonne & Nathan Vaillette. 2006. Conditional entropy measures intelligibility among related languages. In Peter Dirix, Ineke Schuurman, Vincent Vandeghinste & Frank Van Eynde (eds.), *Computational Linguistics in the Netherlands 2006: Selected Papers from the 17th CLIN Meeting*, 51–66. Utrecht: LOT.

Muikku-Werner, Pirkko. 2013. Vironkielisen tekstin ymmärtäminen suomen kielen pohjalta [Understanding Estonian texts on a Finnish language base]. *Lähivõrdlusi – Lähivertailuja* 23, 210–237.

Muikku-Werner, Pirkko. 2014a. Koteksti ja viron ymmärtäminen lähisukukielen pohjalta [Co-text and intelligibility of Estonian language on the basis of a cognate language]. *Lähivõrdlusi – Lähivertailuja* 24, 100–124.

Muikku-Werner, Pirkko. 2014b. Co-text and receptive multilingualism: Finnish students comprehending Estonian. *Estonian Journal of Estonian and Finno-Ugric Linguistics* 3, 99−113.

Muikku-Werner, Pirkko & Maria Heinonen 2012. *Lumesadu* – ‘tarina’ vai ‘lumikasa’ vai ei kumpikaan? Suomalaiset lukiolaiset viron sanoja tunnistamassa [*Lumesadu* – ‘story’ or ‘wodge’ or something completely different? How Finnish senior high school students try to recognise Estonian words]. *Lähivõrdlusi – Lähivertailuja* 22, 157–187.

Paajanen, Ilona & Pirkko Muikku-Werner. 2012. *Tee on kitsas* – onko 'tee kitkerää' vai oletteko 'te

saita'? Suomalaiset opiskelijat viroa ymmärtämässä [*Tee on kitsas* – is ‘tea bitter’ or are ‘you penny-pinching’? Finnish students comprehending Estonian]. *Lähivõrdlusi – Lähivertailuja* 22, 219–258.

Paunonen, Heiki 1983. Allomorfien dynamiikka [Dynamics of allomorphs]. In Auli Hakulinen & Pentti Leino (eds.), *Suomen kielen rakenne ja kehitys* [Structure and development of Finnish language]. Helsinki: Finnish Literature Society.

Reis, Alexandra & Alexandre Castro-Caldas. 1997. Illiteracy: A cause for biased cognitive development. *Journal of the International Neuropsychological Society* 3(5), 444–450.

Remes, Hannu. 1995. *Suomen ja viron vertailevaa taivutustypologiaa* [A contrastive study of inflectional typology in Finnish and Estonian]. Oulun yliopiston suomen ja saamen kielen ja logopedian laitoksen julkaisuja 2. Oulu: Oulun yliopisto.

Remes, Hannu. 2009. *Muodot kontrastissa. Suomen ja viron vertailevaa taivutusmorfologiaa.* [Forms in contrast A contrastive study of inflectional morphology in Finnish and Estonian] (Acta Universitatis Ouluensis B Humaniora 90). Oulu: Oulun yliopisto.

Ringbom, Håkan 2007. *Cross-linguistic Similarity in Foreign Language Learning*. Clevedon: Multilingual Matters.

Ringbom, Håkan & Scott Jarvis. 2009. The importance of cross-linguistic similarity in foreign language learning. In Michael H. Long & Catherine. J. Doughty (eds.), *Handbook of Language Learning*, 106–118. Oxford: Blackwell.

Schepens, Job, Frans van der Slik & Roeland van Hout. 2013. The effect of linguistic distance across Indo-European mother tongues on learning Dutch as a second language. In Lars Borin & Anju Saxena (eds.), *Approaches to Measuring Linguistic Differences*, 199–230. Berlin: Mouton de Gruyter.

Tang Chaoju & Vincent J. van Heuven. 2015. Predicting mutual intelligibility of Chinese dialects from multiple objective linguistic distance measures. *Linguistics* 53(2), 285–312.

van Heuven, Vincent J., Charlotte Gooskens & Renée van Bezooijen. 2015. Introducing Micrela: Predicting mutual intelligibility between closely related languages in Europe. In Judit Navracsics & Szilvia Batyi (eds.), *First and Second Language: Interdisciplinary Approaches* (Studies in Psycholinguistics 6), 127–145. Budapest: Tinta konyvkiado.

Vetter, Eva (ed.). 2016. *Book of Abstracts: 10th International Conference on Multilingualism and Third Language Acquisition.Vienna September 1−3, 2016*. Vienna: University of Vienna.

Wichmann, Søren, Erik W. Holman, Dik Bakker & Cecil H. Brown. 2010. Evaluating linguistic distance measures. *Physica A: Statistical Mechanics and its Applications* 389, 3632–3639.

Viitso, Tiit-Rein. 2003a. Structure of the Estonian language. Phonology, morphology and word formation. In Mati Erelt (ed.), *Estonian Language*, 9–92. Tallinn: Estonian Academy Publishers.

Viitso, Tiit-Rein. 2003b. Rise and development of the Estonian language. In Mati Erelt (ed.), *Estonian Language*, 130–230. Tallinn: Estonian Academy Publishers.

<Figure captions>

Figure 1. Agglutination and fusion in Finnish and Estonian noun inflection.

Figure 2. An example of calculating Levenshtein Distance.

Figure 3. Correlations between the Levenshtein Distance (LD) of the morphological forms as a whole and the IPS of the Finns (Pearson), $r = -.751$, $p = .000$.

Figure 4. Correlations between the Levenshtein Distance (LD) of the morphological forms as a whole and the Index of Perceived Similarity (IPS) of the Estonians (Pearson), $r = .079$, $p = .595$.

<5 tables>

Group	Answers (total, balanced)			Index of Perceived Similarity (total)
	Similar	Somewhat similar	Different	
L1 Finnish (n = 115)	1684.3	1631.3	1424.3	5017.4
L1 Estonian (n = 109)	1125.7	2027.5	1550.5	4370.6

Table 1. Total numbers of answers by the Finns and Estonians and the Index of Perceived Similarity.

	Predictability of Levenshtein value for perceived similarity		
	Morphological form	Stem	Formative
Finns	56% IPS = 205.273 + (–2.723)(LD)	40% IPS = 159.285 + (–2.235)(LD)	21% IPS = 170.750 + (–0.997)(LD)

Table 2. The predictability of the Levenshtein value for the similarity perceived by the Finns.

Similarity category	Similarity category averages				
	IPS FIN	IPS EST	LD whole	LD stem	LD formative
Similar stem, similar formative	171.6	129.1	0.2	0.1	0.5
Similar stem, different formative	**92.3**	**90.1**	0.4	0.2	0.9
Different stem, similar formative	104.9	81.6	0.4	0.3	0.5
Different stem, different formative	43.8	**59.5**	0.5	0.3	0.8

Table 3. The average Index of Perceived Similarity (IPS) and Levenshtein Distance (LD) of the Finns (FIN) and Estonians (EST) by similarity categories. Bold indiactes ■■■.the categories where the Estonians see more similarities than the Finns or where the difference between similarity perceptions of the two participant groups is not significant.

Estonian – Finnish	IPS FIN	IPS EST	LD whole	LD stem	LD formative
Stem vs. agglutinative partitive plural					
lehti 'leaf. PAR.PL' – *leht-i-ä* 'leaf- PL-PAR'	66.1	132.1	0.17	0.20	1,00
noote 'note. PAR.PL'– *nuotte-j-a* 'note-PL-PART'	44.3	60.6	0.25	0.33	1,00
võlgu 'debt. PAR.PL'– *velko-j-a* 'debt- PL-PART'	21.7	31.2	0.71	0.60	1,00
heinu'hay. PAR.PL' – *hein-i-ä* 'hay-PL-PART'	45.2	87.2	0.33	0.20	1,00
nelju 'four. PAR.PL'– *nelj-i-ä* 'four- PL-PART'	48.7	108.3	0.33	0.20	1,00
lube 'permission. PAR.PL'– *lup-i-a* 'permission- PL-PAR'	32.2	38.5	0.60	0.43	0.50
Different stems and formatives					
*pere-**sid*** 'family-PAR.PL' – *perhe-**i-tä*** 'family-PL-PAR'	10.4	39.4	0.44	0.20	0.75
*lain-**te**-st* 'wave-PL-ELA' – *laine-**i**-sta* 'wave-PL-ELA'	87.0	112.8	0.30	0.20	0.67
*õnne**tu-t*** 'unhappy-PAR' – *onne**ton-ta*** 'unhappy-PAR'	47.8	60.6	0.44	0.43	0.50
*keskus-**te-sse*** 'center-PL-ILL' – *keskuksi-in* 'center-PL-ILL'	20.9	44.0	0.54	0.25	1,00
toa 'room.GEN.SG' – ***tuva**-n* 'room-GEN'	20.9	31.2	0.60	0.50	1,00
*us-**te*** 'door-PL' – *uks-**i-en*** 'door-PL-GEN'	31.3	62.4	0.57	0.33	0.75
*maa-**de-sse***'land-PL-ILL' – *ma-**i-hin*** 'land-PL-ILL'	11.3	18.3	0.78	0.33	1,00
*kooli-**de**-s* 'school-PL-INE'– *koulu-**i**-ssa* 'school-PL-INE'	53.9	71.6	0.50	0.40	0.80
*laeva-**de**-lt* 'ship-PL-ABL'– *laivo-**i**-lta* 'ship-PL-ABL'	62.6	69.7	0.50	0.40	0.50

Table 4. **Word form pairs with higher Index of Perceived Similarity (IPS) in the results of the Estonians (EST) and Levenshtein Distances (LDs). Bold in the example words emphasizes differences between Estonian and Finnish inflectional forms.**

■■■.

Estonian – Finnish	IPS FIN	IPS EST
sinis-te 'blue-PL' – *sinis-te-**n*** 'blue-PL-GEN'	176.5	138.5
suur-te 'big-PL' – *suur-te-**n*** 'big-PL-GEN'	170.4	129.4
keel-t 'tongue-PAR' – *kiel-t**ä*** 'tongue-PAR'	133.0	95.4
pu-i-d 'tree-PL-PAR' – *pu-i-t**a*** 'tree-PL-PAR'	156.5	123.9
hamba-i-d 'tooth-PL-PAR' – *hampa-i-t**a*** 'tooth-PL-PAR'	172.2	120.2
mustika-i-d 'blueberry-PL-PAR' – *mustiko-i-t**a*** 'blueberry-PL-PAR'	120.0	113.8
kuusiku-i-d 'spruce forest-PL-PAR' – *kuusiko-i-t**a*** 'spruce forest-PL-PAR'	127.8	91.7
murde-i-s 'dialect- PL-INE'– *murte-i-ss**a*** 'dialect- PL-INE'	168.7	120.2
silmi-s 'eye.PL-INE' – *silm-i-ss**ä*** 'eye- PL-INE'	185.2	123.9
pabere-i-s 'paper-PL-INE' – *papere-i-ss**a*** 'paper-PL-INE'	177.4	130.3
vangla-s 'prison-INE'– *vank**i**la-ss**a*** 'prison-INE'	122.6	75.2
raamat-ust 'book-ELA'– *raamatu-st**a*** 'bible-ELA'	188.7	150.5
herne-i-st 'pea-PL-ELA'– *herne-i-st**ä*** 'pea-PL-ELA'	186.1	137.6
hoone-i-st 'room-PL-ELA'– *hoone-i-st**a*** 'house/PL-ELA'	162.6	135.8
kuninga-i-l 'king-PL-ADE' – *kuninka-i-ll**a*** 'king-PL-ADE'	174.8	115.6
tiigre-i-lt 'tiger-PL-ABL' – *tiik**e**re-i-lt**a*** 'tiger-PL-ABL'	150.4	101.8
tütre-le 'daughter-ALL'– *tytt**ä**re-ll**e*** 'daughter-ALL'	158.3	142.2
lään-de 'west-ILL' – *länte-**en*** 'west-ILL'	109.6	108.3
*hõbe**da**-i-st* 'silver-PL-ELA'– *hope-i-st**a*** 'silver-PL-ELA'	76.5	59.6
järve 'lake.ILL.SG' – *järve-**en*** 'lake-ILL'	100.9	114.7
*nobe**da**-i-le* 'speedy-PL-ALL'– *nope-i-ll**e*** 'speedy-PL-ALL'	96.5	98.2
*sade**me**-i-s* rain-PL-INE'– *sate-i-ss**a*** 'rain-PL-INE'	48.7	93.4
pesa 'nest.PAR.SG'– *pesä-**ä*** 'nest-PAR'	73.9	154.1

Table 5. Additions and deletions in test pairs: Index of Perceived Similarity (IPS) of Finns (FIN) and Estonians (EST). Bold in the example words emphasizes differences between Estonian and Finnish inflectional forms.

■■■.

Estonian – Finnish	LD word	LD stem	LD formative	FIN balanced (100)	EST balanced (100)
Similar stem, similar formative					
raamat-ust 'book-ELA'– *raamatu-sta* 'bible-ELA'	0.10	0,00	0.33	188.7	150.5
kala-d 'fish-PL' – *kala-t* 'fish-PL'	0.20	0,00	1,00	178.3	157.8
sinis-te 'blue-PL' – *sinis-te-**n*** 'blue-PL-GEN'	0.13	0,00	0.33	176.5	138.5
keel-t 'tongue-PAR' – *kiel-tä* 'tongue-PAR'	0.33	0.25	0.50	133.0	95.4
pu-i-d 'tree-PL-PAR' – *pu-i-ta* 'tree-PL-PAR'	0.40	0,00	0.33	156.5	123.9
hamba-i-d 'tooth-PL-PAR' – *hampa-i-ta* 'tooth-PL-PAR'	0.25	0.20	0.75	172.2	120.2
murde-i-s 'dialect- PL-INE'– *murte-i-ssa* 'dialect- PL-INE'	0.30	0.20	0.50	168.7	120.2
kuninga-i-l 'king-PL-ADE' – *kuninka-i-lla* 'king-PL-ADE'	0.27	0.14	0.50	174.8	115.6
pabere-i-s 'paper-PL-INE' – *papere-i-ssa* 'paper-PL-INE'	0.30	0.17	0.50	177.4	130.3
hoone-i-st 'room-PL-ELA'– *hoone-i-sta* 'house/PL-ELA'	0.22	0.20	0.25	162.6	135.8
silmi-s 'eye.PL-INE' – *silm-i-ssä* 'eye- PL-INE'	0.25	0.20	0.75	185.2	123.9
suur-te 'big-PL' – *suur-te-n* 'big-PL-GEN'	0.14	0,00	0.33	170.4	129.4
herne-i-st 'pea-PL-ELA'– *herne-i-stä* 'pea-PL-ELA'	0.11	0,00	0.25	186.1	137.6
Average	0.20	0.10	0.50	171.6	129.1
Similar stem, different formative					
pere-sid 'family-PAR.PL' – *perhe-i-tä* 'family-PL-PAR'	0.44	0.20	0.75	10.4	39.4
saar-te 'island-PL' – *saar-i-en*'island-PL-GEN'	0.29	0,00	0.50	140.0	53.2
pesa 'nest.PAR.SG'– *pesä-ä* 'nest-PAR'	0.40	0.25	1,00	73.9	154.1
lehti 'leaf. PAR.PL' – *leht-i-ä* 'leaf-PL-PART'	0.17	0.20	1,00	66.1	132.1
noote 'note. PAR.PL'– *nuotte-j-a* 'note- PL-PAR'	0.25	0.33	1,00	44.3	60.6
võlgu 'debt. PAR.PL'– *velko-j-a* 'debt- PL-PAR'	0.71	0.60	1,00	21.7	31.2
järve 'lake.ILL.SG' – *järve-en* 'lake-ILL'	0.29	0,00	1,00	100.9	114.7
lään-de 'west-ILL' – *länte-en* 'west-ILL'	0.57	0.33	0.67	109.6	108.3

sedele-i-d ‘note-PL-PAR’– *setele-j-* ‘note-PL-PAR’	0.50	0.17	1,00	149.6	67.9
endisi ‘former.PAR.PL’– *entis-i-ä* ‘former-PL-PAR’	0.29	0.33	1,00	156.5	116.5
lain-te-st ‘wave-PL-ELA’ – *laine-**i**-sta* ‘wave-PL-ELA’	0.30	0.20	0.67	87.0	112.8
albume-i-d ‘album-PL-PAR’– *albume-j-a* ‘album-PL-PAR’	0.25	,00	1,00	147.8	89.9
Average	0.40	0.20	0.90	92.3	90.1
Different stem, similar formative					
harrastuse-st ‘hobby-ELA’– *harrastukse-sta* ‘hobby-ELA’	0.14	0.09	0.33	153.9	74.3
õnnetu-t ‘unhappy-PAR’ – *onneton-ta* ‘unhappy-PAR’	0.44	0.43	0.50	47.8	60.6
ve-te-l ‘water-PL-ADE’– *ves-i-llä* ‘water-PL-ADE’	0.57	0.50	0.67	51.3	40.4
tiigre-i-lt ‘tiger-PL-ABL’ – *tiikere-i-lta* ‘tiger-PL-ABL’	0.27	0.29	0.25	150.4	101.8
tütre-le ‘daughter-ALL’– *tyttäre-lle* ‘daughter-ALL’	0.30	0.43	0.33	158.3	142.2
vangla-s ‘prison-INE’– *vank**i**la-ssa* ‘prison-INE’	0.40	0.29	0.67	122.6	75.2
sademe-i-s rain-PL-INE’– *sate-i-ssa* ‘rain-PL-INE’	0.50	0.50	0.50	48.7	39.4
nobeda-i-le ‘speedy-PL-ALL’– *nope-i-lle* ‘speedy-PL-ALL’	0.40	0.50	0.25	96.5	98.2
hõbeda-i-st ‘silver-PL-ELA’– *hope-i-sta* ‘silver-PL-ELA’	0.50	0.50	0.25	76.5	59.6
mustika-i-d ‘blueberry-PL-PAR’ – *mustiko-i-ta* ‘blueberry-PL-PAR’	0.30	0.14	0.67	120.0	113.8
kuusiku-i-d ‘spruce forest-PL-PAR’ – *kuusiko-i-ta* ‘spruce forest-PL-PAR’	0.30	0.14	0.67	127.8	91.7
Average	0.40	0.30	0.50	104.9	81.6
Diferent stem, different formative					
kolmanda-te-s ‘the thirth-PL-INE’– *kolmans-i-ssa*‘the thirth-PL-INE’	0.46	0.25	0.67	60.9	52.3
keskus-te-sse ‘center-PL-ILL’ – *keskuks-i-in* ‘center-PL-ILL’	0.54	0.25	1,00	20.9	44.0
toa ‘room.GEN.SG’ – *tuva-n* ‘room-GEN’	0.60	0.50	1,00	20.9	31.2
us-te ‘door-PL’ – *uks-i-en* ‘door-PL-GEN’	0.57	0.33	0.75	31.3	62.4
maa-de-sse‘land-PL-ILL’ – *ma-i-hin* ‘land-PL-ILL’	0.78	0.33	1,00	11.3	18.3
linde ‘bird. PAR.PL’ – *lintu-j-a* ‘bird-PL-PART’	0.57	0.40	1,00	47.8	47.7

heinu‘hay. PAR.PL’ – *hein-i-ä* ‘hay-PL-PAR’	0.33	0.20	1,00	45.2	87.2
nelju ‘four. PAR.PL’– *nelj-i-ä* ‘four-PL-PART’	0.33	0.20	1,00	48.7	108.3
lube ‘permission. PAR.PL’– *lup-i-a* ‘permission- PL-PAR’	0.60	0.43	0.50	32.2	38.5
punase-i-d ‘red-PL-PAR’ – *punais-i-a* ‘red-PL-PAR’	0.40	0.29	0.67	89.6	82.6
kooli-de-s ‘school-PL-INE’– *koulu-i-ssa* ‘school-PL-INE’	0.50	0.40	0.80	53.9	71.6
laeva-de-lt ‘ship-PL-ABL’– *laivo-i-lta* ‘ship-PL-ABL’	0.50	0.40	0.50	62.6	69.7
Average	0.50	0.30	0.80	43.8	59.5

Table A1. The Levenhstein Distances (LDs) and Index of Perceived Similarity (IPS) scores of individual test items.

<ENDNOTE>

[1] The glossing abbreviations: 1PL = first person plural, ABL = ablative, ADE = adessive, ALL = allative, CLI = clitic, ELA = elative, GEN = genitive, ILL = illative, INE = inessive, NOM = nominative, PAR = partitive, PL = plural, POSS = possessive suffix.