

This paper is based on a talk presented at the *22nd Scandinavian Conference of Linguistics*, University of Aalborg, June 2006. It has been submitted for publication in the proceedings of SCL 22:

Kaiser, Elsi & Virve Vihman. (to appear) *On the referential properties of Estonian pronouns and demonstratives*. To appear in the *Proceedings of the 22nd Scandinavian Conference of Linguistics*.

On the referential properties of Estonian pronouns and demonstratives

Elsi Kaiser, University of Southern California
&
Virve-Anneli Vihman, University of Tartu

1. Introduction

It is commonly assumed that there exists a correlation between the salience/accessibility of a referent and the expressions used to refer to it, such that the most reduced referring expressions (e.g. pronouns) are used for the most accessible referents, and fuller expressions (e.g. full NPs) for less accessible referents (e.g. Ariel 1990, Gundel et al. 1993). Positing such a correlation brings up the question of what makes a referent a good candidate for subsequent reference with a reduced anaphoric form. A number of factors have been put forth in the literature, and in this paper we focus on two properties of the antecedent that have been claimed to influence reference resolution, namely grammatical role and word order. We address these issues from the perspective of Estonian, a highly inflected, flexible word order language with canonical SVO order and two kinds of third person anaphors available for singular, human referents: (1) the gender-neutral pronoun *ta* 's/he' and (2) the demonstrative *see* 'this'.

Based on the results of a sentence-completion experiment investigating the referential properties of these two forms, we conclude that the relation between referring expressions and accessibility is more complex than commonly assumed, and that the referential properties of different forms cannot always be mapped directly onto a single accessibility scale. More specifically, we claim that *ta* 's/he' and *see* 'this' exhibit different degrees of sensitivity to grammatical role and word order. Estonian thus provides support for the *form-specific multiple-constraint* approach to reference resolution (see Kaiser 2003, Kaiser 2005, Kaiser & Trueswell in press), which is built on the claim that a unified 'monolithic' notion of salience is insufficient, and that anaphoric forms can differ in their sensitivity to different kinds of information.

2. Background: Referent salience

Correlations have often been claimed to exist between different kinds of referential expressions (full NPs, pronouns, demonstratives etc.) and the level of salience/accessibility of their antecedents. There is a general consensus that the more reduced an anaphoric expression is, the more salient/accessible its antecedent has to be. For example, Arnold (1998) observes that "loosely speaking, all researchers have observed that pronouns are used most often when the referent is

represented in a prominent way in the minds of the discourse participants, but more fully specified forms are needed when the representation of the referent is less prominent” (Arnold 1998:4). Various accessibility-based hierarchies of referential forms that aim to represent these correlations have been proposed in the literature (e.g. Gundel, Hedberg and Zacharski 1993, Givón 1983 and Ariel 1990). A generalized hierarchy is shown in (1).

- (1) *Null* > *pronouns* > *demonstratives* > *full NPs...*
 more accessible referents less accessible referents

In this paper, we focus on two factors that have been claimed to influence referent salience and anaphor resolution: syntactic function and word order. As becomes clear in the later sections of this paper, our findings indicate that the mapping between accessibility/salience levels and referential forms is not straightforward. However, for ease of exposition, in our review of existing work we assume that referential forms can be ranked along an accessibility hierarchy.

The influence of syntactic function is supported by a large body of research suggesting that a tight connection exists between a referent’s grammatical role and the likelihood of subsequent pronominal reference. Specifically, researchers have found that subjects are more salient than objects (e.g. Brennan, Friedman & Pollard 1987, Crawley & Stevenson 1990, Gordon, Grosz & Gilliom 1993, *inter alia*).

However, for languages like English, with relatively rigid subject-object order, we cannot tell whether the special status of subjects is due to their position at the beginning of the sentence, or their grammatical role. In order to pull these factors apart, researchers have turned to languages with flexible word order, which allow us to look at the notions of subjecthood and linear order independently. Existing research has revealed different findings for different languages. For example, Rambow (1993) and Strube & Hahn (1996) claim that in German, word order correlates at least partially with the likelihood of being referred to later with a pronoun.¹ Rambow provides examples of sentences with subject-object and object-subject order in the middle field, where a subsequent pronoun prefers the linearly initial referent, regardless of whether it is a subject or an object.

In contrast, work by Turan (1998) and Hoffman (1998) suggests that in Turkish, salience correlates with the grammatical (or semantic) role of the discourse entity, and is not affected by word order. They found that subjects are more salient than objects (more likely to be referred to with a null pronoun) even in

¹ Strube & Hahn’s (1996, 1999) algorithm assumes that “hearer-old discourse entities are ranked higher than hearer-new discourse entities” (1999:320). It follows that if word order variation is guided by information structural factors – in particular, hearer-status – then it will influence salience ranking.

sentences with noncanonical word order where the object linearly precedes the subject.²

However, the assumption made above – that one factor determines referent salience – is not necessarily correct. In fact, while some researchers (e.g. Strube & Hahn 1996, Hoffman 1998) claim that salience is determined by a single factor, others regard it as a result of the interaction of multiple factors (e.g. Ariel 1990, Arnold 1998, Lappin & Leass 1994, and others). We argue for a slightly modified version of the multiple-factors approach, namely the form-specific multiple-constraints framework. This approach claims that multiple factors influence reference resolution and can be weighted differently for different anaphoric forms (see Kaiser 2003, Kaiser & Trueswell in press). Thus, anaphoric forms can differ in how sensitive they are to different kinds of information. For example, one form may be more sensitive to the linear position of the antecedent, while another form is more sensitive to the grammatical role.

Support for this approach comes from psycholinguistic experiments focusing on Finnish, a language closely related to Estonian (Kaiser 2003, Kaiser & Trueswell in press). Kaiser investigated Finnish sentences with SVO/OVS order, where both arguments are full NPs (see also Kaiser 2005, Järvikivi et al. 2005) and found that the pronoun *hän* ‘s/he’ is influenced primarily by grammatical role and prefers subjects over objects regardless of word order, whereas the demonstrative *tämä* ‘this’ (which can be used to refer to human referents) is sensitive to word order (correlated with given/new) and also exhibits some sensitivity to grammatical role; *tämä* especially prefers post-verbal non-subjects. Thus, these two forms are not in complementary distribution, and they differ in the degree of sensitivity they exhibit to different kinds of information. In the present paper, our aim is to take steps toward evaluating the crosslinguistic validity of this form-specific approach to reference resolution, by testing its application to Estonian.

3. Estonian

Estonian is well-suited for investigating the validity of the form-specific hypothesis and more specifically, the contributions of word order and syntactic role. Estonian has flexible word order; all six configurations of SVO are possible given the right context, though SVO is the least marked. In a corpus study, 25% of all sentences had this order (Tael 1988). However, even though Estonian has flexible word order, the variation is not random – rather, it is restricted by information-structural factors. As in many languages, old information tends to precede new information. Thus, if the subject is old or given information and the object is new information, SVO order is more likely to be used (ex.2). Similarly, if

² The question of whether the difference between overt pronouns (German) and null pro (Turkish) plays a role here is intriguing and merits further attention.

the object is old information and the subject is new, OVS order is used (3). If both arguments have the same information status (both old or both new), canonical SVO order is typically used.

(2) **SVO** (www.postimees.ee/191004/online_uudised/147754.php)
 Maailmapank hoiatab Läti majandust ülekuumenemise eest
 World Bank-NOM warns [Latvia-GEN economy]-PAR over-heating from
 ‘The World Bank warns the Latvian economy against overheating.’

(3) **OVS** (www.epl.ee, 18.06.06)
 Majandust ohustavad kinnisvara- ja laenuboom
 Economy-PAR threaten [real estate and loan boom]-NOM
 ‘A boom in real estate and bank loans is threatening the economy.’

Estonian also exhibits a preference for verb-second (V2) word order, although this is not a rigid rule as in Germanic, but rather a tendency. V2 operates in Estonian to the extent that, when it clashes with information-structural constraints, it can override them. Finnish, although similar to Estonian in the general given/new ordering patterns shown above, does not have a V2 bias. For more discussion of V2 in Estonian, see e.g. Viikuna (1998), Tael (1988).

The Estonian anaphoric paradigm offers multiple possibilities for referring back to human antecedents. In this paper we focus on the pronoun *ta* ‘s/he’ and the demonstrative *see* ‘this.’ Accessibility hierarchies such as the one exemplified in (1) predict that *ta* will refer to more accessible, salient referents than *see*. In a similar vein, Pajusalu (1995, 1997) describes *ta* ‘s/he’ as the most common way of referring to the entity that is currently in the focus of attention. According to Erelt *et al.* (1993:209) and Tauli (1983:323), if there are two human, third person referents in a clause, the short pronoun *ta* is used to refer to the first referent and the demonstrative *see* to the second (ex.4).

(4)
 Tüdruk₁ vilksas pois₂ poole; **ta₁/see₂** oli kahvatu.
 Girl-NOM glanced boy-GEN towards; ta/see-NOM was pale.
 ‘The girl glanced towards the boy; she/he was pale.’ (Erelt *et al.* 1993:209)

However, this discussion leaves open the crucial question of what happens when the word order is changed. In other words, we cannot tell whether the referential properties of *ta* and *see* are sensitive to grammatical role or word order, or some combination of both. To test this, Kaiser & Hietam (2004) investigated the referential properties of *ta* and *see* following SVO and OVS orders. Their results suggest that (i) the pronoun *ta* refers to subjects, regardless of word order

(SVO/OVS), and (ii) the demonstrative *see* refers to objects in SVO order, but is split between subject and object in OVS order. However, these findings are somewhat preliminary, because (i) the SVO/OVS sentence was presented out of context – which meant that both arguments were new information, a situation in which OVS order is infelicitous – and (ii) only one test sentence was presented, and thus the results were too small for reliable statistical analysis. Our aim in this paper is to test whether the form-specific approach to reference resolution applies to Estonian, with a more thorough and larger-scale investigation of how grammatical role and word order influence the referential properties of *ta* and *see*.

3.1 Predictions

This section sketches out the predictions made by single-factor and multiple-factor approaches regarding the antecedent preferences of *ta* and *see* when preceded by SVO or OVS order. Let us assume for now that *ta* is used to refer to highly accessible/salient referents, and *see* for less accessible/salient referents.

One-factor approaches assign full responsibility for antecedent choice to a single determining criterion – for example, grammatical role or linear position. According to the grammatical-role one-factor approach, subjecthood is the determining factor, as subjects are more salient than objects. This approach predicts that the pronoun *ta* will refer to subjects and the demonstrative *see* to objects, regardless of word order. In contrast, the word-order one-factor approach states that salience is determined by linear position, with constituents to the left being more salient than constituents to the right. This approach predicts that the preverbal element is referred to with *ta* and the postverbal element with *see*, regardless of grammatical role.

As discussed above, a multiple-factor approach offers another way to explain choice of antecedent. If salience is the result of the interaction of grammatical role and word order, different predictions arise for SVO and OVS order. In SVO order, the two factors are aligned, since the more salient grammatical role (subject) is also in the more salient linear position (leftmost). Thus, the predictions for SVO order are clear: *ta* will prefer the preverbal subject and *see* the postverbal object. In OVS order, however, the two factors contradict each other, with no clearly preferred antecedent predicted for either *ta* or *see*.

The multiple-factor description above assumes that (i) word order and grammatical role are equally influential relative to each other, and that (ii) whatever amount of influence they exert on pronouns is equal to the amount of influence they exert on demonstratives. Multiple-constraint analyses of other linguistic phenomena (e.g. MacDonald, Pearlmutter, & Seidenberg 1994, Trueswell & Tanenhaus 1994) suggest that assumption (i) may not be true: different kinds of information can be weighted differently. Furthermore, the form-specific multiple constraints approach goes against assumption (ii). According to

the form-specific approach, various factors can be weighted differently for different anaphoric forms. In particular, if we extend the Finnish findings of Kaiser (2003, 2005, see also Kaiser & Trueswell in press) to Estonian, we can predict that the pronoun *ta*, like its Finnish equivalent *hän* ‘s/he’, will be sensitive primarily to grammatical role, whereas *see* will exhibit susceptibility to the influence of both linear order and grammatical role, similar to Finnish *tämä* ‘this’. Thus, we predict that (i) *ta* will prefer subjects in both SVO and OVS order (see also Kaiser & Hiietam 2004), whereas (ii) *see* will strongly prefer the postverbal object with SVO sentences (due to the confluence of linear order and grammatical role) and will show a weaker preference for the postverbal subject in OVS order.

4. Sentence continuation experiment

To find out which hypothesis most accurately captures the referential properties of Estonian *ta* and *see*, we investigated the interpretation of these forms following SVO and OVS sentences. As shown in example (5), the critical SVO and OVS sentences were preceded by a brief context-setting text, which mentioned the preverbal argument of the target sentence (S in SVO, O in OVS). This was done to ensure that sentences with both orders were felicitous. Each target sentence was followed by a prompt word, either *ta* ‘s/he’ or *see* ‘this’. We thus created four conditions: [SVO.Ta...], [OVS.Ta...], [SVO.See....] and [OVS.See.....]. The participants’ (n=24) task was to write a natural-sounding continuation using the anaphoric prompt. The nouns used for the subject and object in the SVO/OVS sentences referred to professions (e.g. doctor, baker, reporter), in order to make the continuations easier to interpret for coding purposes. Verbs were agentive (see Stevenson et al. 1994). Participants’ continuations were coded according to which of the NPs in the preceding sentence (subject or object) the participants chose as the antecedent of the pronoun. In some cases, participants did not use the demonstrative *see* anaphorically to refer to the preceding subject or object, and rather opted for a deictic (e.g. ‘this made him laugh’) or modifier use (‘this man’). Such continuations were coded separately, as were those where the intended antecedent was unclear.

(5)

Juhan käis eile pargis.
 Juhan-NOM walked yesterday park-INE
 ‘Juhan went to the park yesterday.’

Ta märkas suure puu kõrval pruunis mantlis
 3sg-NOM noticed big-GEN tree-GENbeside brown-INE coat-INE
 fotograafi.
 photographer-PAR

‘He noticed a photographer in a brown coat next to a tree.’

Fotograaf aitas värviplekkidega kaetud **kunstnikku**.
 Photographer-NOM helped paint-spots-COM covered artist-PAR
 ‘The photographer was helping an artist covered with paint splatters.’

Ta/See ...

Ta-NOM/ See-NOM...

5. Results and discussion

Figure 1 shows the ‘subject-object difference score’ for each condition, calculated by subtracting the proportion of object continuations from the proportion of subject continuations.³ Positive numbers indicate a preference for the subject and negative numbers a preference for the object. As Fig. 1 illustrates, there is no overarching effect of either linear order or grammatical role. Rather, the results show that *ta* and *see* differ in which of these factors they are sensitive to, and to what degree.

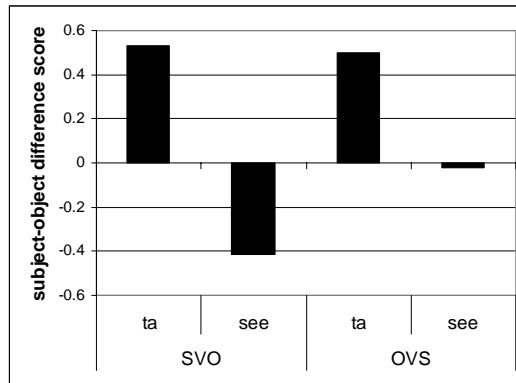


Figure 1. Preference for preceding subject vs. preceding object for *ta* and *see*. (Positive numbers reflect a preference for the subject. Negative numbers reflect a preference for the object.)

More specifically, in the *ta* conditions participants show a clear preference for interpreting the pronoun as referring to the preceding subject, regardless of word

³ The subject and object continuation proportions do not add up to 1 because there were some other continuation types as well, e.g. cases where the intended referent was unclear, or where the demonstrative was used as a prenominal modifier (e.g. ‘this situation’).

order. This shows that *ta* is sensitive to the grammatical role of its antecedent and prefers subjects, which confirms the findings of Kaiser & Hiietam (2004) and also resembles the referential properties of the Finnish pronoun *hän* 's/he'.

A different pattern arises in the *see* conditions. Following *SVO* sentences, *see* strongly prefers objects over subjects. However, a detailed analysis of the continuations reveals that demonstrative uses of *see* (e.g. using *see* discourse-deictically, as in 'this was a strange event') are also very common in this condition. In fact, in the *SVO/see* condition, the number of demonstrative uses (not shown in Fig. 1) is comparable to the number of object-reference uses. Turning now to *OVS* order, we see that *see* shows no preference for the subject or object. Data analysis shows that the majority of continuations in this condition use *see* demonstratively (>65%). The high proportion of non-anaphoric uses suggests that in *OVS* order, neither the subject nor the object is an ideal referent for *see*, presumably because the object is not post-verbal, and the subject, while post-verbal (and new information), has a dispreferred grammatical role. Thus, it appears that participants are using the demonstrative interpretation as a kind of referential 'escape hatch' in a context where using *see* anaphorically would result in ambiguity.

In sum, the continuation results indicate that *see* is sensitive to both grammatical role (prefers objects) and order-of-mention (prefers postverbal referents), whereas *ta* 's/he' shows a strong sensitivity to the grammatical role of its antecedent (prefers subjects). This finding cannot be captured by either of the single-factor approaches, nor by a multiple-constraint theory that assumes *ta* and *see* to show the same degree of sensitivity to linear order and grammatical role. However, the results are compatible with the form-specific approach, which allows for referential forms to differ in their degree of sensitivity to different kinds of information.

We do not have the space here for a detailed comparison of these Estonian results and the earlier Finnish findings. However, it is worth noting that in both languages, (i) the pronominal anaphors show a strong sensitivity to the grammatical role of the antecedent, and (ii) the demonstrative anaphors are not mirror images of the pronouns. Both languages provide support for the form-specific multiple-constraint approach. However, there are subtle differences in the referential properties of the Finnish demonstrative *tämä* and the Estonian demonstrative *see* which indicate that for a given anaphoric form, the relative contributions of word order and grammatical role can differ from language to language – an observation which can be readily captured within a system of weighted constraints.

6. Conclusions and implications

The results of the sentence completion experiment suggest that a unified, 'monolithic' notion of accessibility/salience is insufficient. The Estonian results

cannot be captured by single-factor models claiming that one factor determines referent salience, or by multiple-factor models assuming that several factors contribute equally for all anaphoric forms. The data fits the predictions of the form-specific multiple-constraint approach to reference resolution (see Kaiser 2003, 2005, Kaiser & Trueswell in press), which claims that different referential forms are sensitive to different kinds of information to different degrees. This approach can also be extended to deal with cross-linguistic variation, since various factors can be weighted differently for different anaphoric forms and across languages.

It should be noted that this work focuses on two kinds of informationally-impooverished anaphoric forms, and that our conclusions should not be interpreted as a claim that all referential forms, including full NPs and other more informative forms (e.g. *the little boy wearing a red t-shirt*) pattern in the same way.

Of course, a number of questions remain open. In future work, we plan to conduct a more detailed comparison of the referential properties of the Estonian demonstrative *see* and the Finnish demonstrative *tämä* in order to investigate further the nature of the interaction between word order and grammatical role. An investigation of word orders other than SVO/OVS, as well as a comparison of the information-structural properties of different word order configurations, are also important directions for future work; although many of the information-structural constraints guiding word order in Finnish and Estonian are very similar, they are not always the same (see also Huumo 1993). Furthermore, we hope to extend this work to other Finnish and Estonian pronominal patterns, including plural forms, the southern Estonian distal demonstrative *too* and colloquial Finnish.

References

- Ariel, M. 1990. *Accessing NP Antecedents*. Routledge/Croom Helm.
- Arnold, J.E. 1998. *Reference form and discourse patterns*. Ph.D. diss., Stanford University.
- Brennan, S. E., Friedman, M. A. & Pollard, C. J. 1987. A Centering approach to pronouns. In *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics*, 155-162. Stanford, CA.
- Crawley, R.J. and Stevenson, R.J. 1990. Reference in single sentences and in texts. *Journal of Psycholinguistic Research* 19(3):191-210.
- Erelt et.al. 1993. *Eesti keele grammatika II Süntaks. Lisa: Kiri*. Eesti Teaduste Akadeemia.
- Gernsbacher, M.A. 1990. *Language Comprehension as Structure Building*. LEA.
- Givón, T. 1983. *Topic continuity in discourse: A quantitative cross-language study*. Amsterdam: John Benjamins.
- Gordon, P.C., Grosz, B.J., & Gilliom, L.A. 1993. Pronouns, names, and the Centering of attention in discourse. *Cognitive Science* 17:311-347
- Gundel, J.K., Hedberg, N. & Zacharski, R. 1993. Cognitive Status and the Form of Referring Expressions in Discourse. *Language* 69:274-307.
- Hoffman, B. 1998. Word order, information structure and Centering in Turkish. In *Centering Theory in Discourse*, ed. M.A. Walker, A.K. Joshi & E.F. Prince, 251-272. Oxford: Clarendon Press.

- Huumo, T. 1993. Suomen ja viron kontrastiivista sanajärjestysvertailua. In *Studia Comparativa Linguarum Orbis Maris Baltici 1. Tutkimuksia syntaksin ja pragmasyntaksi alalta*, ed. V. Yli-Vakkuri, 97-158. Turku: Univ. of Turku.
- Järvikivi, J., van Gompel, R., Hyönä, J. & Bertram, R. 2005. Ambiguous pronoun resolution. Contrasting the First-Mention and Subject-Preference accounts. *Psychological Science* 16(4):260-264.
- Kaiser, E. 2003. *The Quest for a Referent: A Crosslinguistic Look at Reference Resolution*. Ph.D. dissertation, University of Pennsylvania.
- Kaiser, E. 2005. Different forms have different referential properties: Implications for the notion of 'salience'. In A. Branco, T. McEnery & R. Mitkov (eds), *Anaphora Processing: linguistic, cognitive and computational modeling*, 261-282. Philadelphia/Amsterdam: John Benjamins.
- Kaiser, E. & K. Hiietam 2004. A Typological Comparison of Third Person Pronouns in Finnish and Estonian. In A. Dahl, K. Bentzen & P. Svenonius (ed.), *Nordlyd* 31(4), Proceedings of the Workshop on Generative Approaches to Finnic Languages, 654-667.
- Kaiser, E. & Trueswell, J. (in press). Investigating the interpretation of pronouns and demonstratives in Finnish: Going beyond salience. In E. Gibson & N. Pearlmutter (eds), *The processing and acquisition of reference*. Cambridge, Mass.: MIT Press.
- Lappin S. & Leass H. 1994. An algorithm for pronominal anaphora resolution. *Computational Linguistics* 20(4):535-561.
- MacDonald, M., Pearlmutter, N. & Seidenberg, M. 1994. The lexical nature of syntactic ambiguity resolution. *Psychological Review* 101:676-703.
- Pajusalu, R. 1995. Pronominit *see, tema* ja *ta* viron puhekielessä. *Sananjalka* 37:81-93.
- Pajusalu, R. 1997. Eesti pronomeneid I. Ühiskeele *see, too* ja *tema/ta*. *Keel ja Kirjandus* 24-30, 106-115.
- Pajusalu, R., M. Hietaharju, V. Taro & K. Yallop. 1999. *Keelesild*. Helsinki: Otava.
- Rambow, O. 1993. *Pragmatic aspects of scrambling and topicalization in German*. Presented at Workshop on Naturally-Occurring Discourse, University of Pennsylvania.
- Stevenson, R. J., Crawley, Rosalind J. & Kleinman, D. 1994. Thematic roles, focus and the representation of events. *Language and Cognitive Processes* 9:519-548.
- Strube, M. & Hahn, U. 1996. Functional Centering. In *Proceedings of ACL '96*, 270-277.
- Strube, M. & Hahn, U. 1999. Functional Centering. *Computational Linguistics* 25:309-345.
- Tael, K. 1988. *Sõnajärgemallid eesti keeles (võrrelduna some keelega)*. Tallinn: Eesti NSV Teaduste Akadeemia Keele ja Kirjanduse Instituut. Preprint KK 1-56.
- Tauli, V. 1983. *Standard Estonian Grammar. Part II Syntax*. Acta Universitatis Upsaliensis, Studia Uralica et Altaica Upsaliensia. Uppsala.
- Tetreault, J. 2001. *A Corpus-Based Evaluation of Centering and Pronoun Resolution*. 27(4):507-520.
- Trueswell, J. C., & Tanenhaus, M. K. (1994). Towards a lexicalist framework for constraint-based syntactic ambiguity resolution. In C. Clifton, K. Rayner, & L. Frazier (Eds.), *Perspectives on sentence processing*. Hillsdale, NJ: Lawrence Erlbaum.
- Turan, Ü.D. 1998. Ranking forward-looking centers in Turkish. In *Centering Theory in Discourse*, ed. M.A. Walker, A.K. Joshi & E.F. Prince, 136-160. Oxford: Clarendon Press.
- Vilkuna, M. 1998. Word order in European Uralic. In *Constituent Order in the Languages of Europe*, ed. A. Siewierska. Berlin: Mouton de Gruyter, 173-233.