Gradient constraints on the use of Estonian possessive reflexives

Suzanne Lesage
Université de Paris, LLF, CNRS
suzanne.lesage.broyelle@gmail.com

Olivier Bonami
Université de Paris, LLF, CNRS
olivier.bonami@univ-paris-diderot.fr

Abstract

We report on a corpus study of the use of reflexive vs. nonreflexive possessives in Estonian sentences headed by verbs taking an allative argument. We parsed the Estonian National Corpus using UDPipe trained with the Estonian Dependency Corpus, extracted relevant data automatically, eliminated false positives and annotated the data by hand. This allowed us to document effects of grammatical functions, word order and person on the choice of a reflexive vs. nonreflexive, using generalized linear mixed models. We hypothesize that the documented effects are due to the combined effects of grammatical relations, information structure, and ambiguity avoidance.

1 Introduction

Estonian allows two ways of referring to the possessor of a noun: adnominal genitive pronouns, which agree in person and number with their antecedent (1), and two reflexive forms *oma* and *enda*, which do not agree (2). In the remainder of this paper we call relevant uses of genitive pronouns NONREFLEXIVE POSSESSIVES, and relevant uses of *oma* REFLEXIVE POSSESSIVES. We leave aside *enda* for brevity.

(1) a. Peeter vii-s mind$_i$ minu$_i$ vanema-te juurde.
Peeter.NOM lead-PST 1SG.PART 1SG.GEN parent-PL.GEN at
Peeter led me to my parents' place.

b. Peeter$_j$ vii-s Jaani$_j$ tema$_{j/*i}$ vanema-te juurde.
Peeter.NOM lead-PST JaanSG.PART 3SG.GEN parent-PL.GEN at
Peeter$_i$ led Jaan$_j$ to his$_{j/*i}$ parents' place.

(2) a. Ma$_i$ vii-si-n Jaan-i$_j$ oma$_{i/*j}$ vanema-te juurde.
1SG.NOM lead-PST-1SG Jaan-GEN POSS.REFL parent-PL.GEN at
I led Jaan to my parents' place.

b. Peeter$_j$ vii-s Jaani$_j$ oma$_{i/*j}$ vanema-te juurde.
Peeter.NOM lead-PST JaanSG.PART 3SG.GEN parent-PL.GEN at
Peeter$_i$ led Jaan$_j$ to his$_{i/*j}$ parents' place.

In canonical constructions, reflexive possessives are bound by the local subject (2), while nonreflexives can either be bound by a local non-subject, as in (1), or be locally unbound. The complementary distribution between reflexive and nonreflexive possessives collapses in some constructions, notably when the head verb has a noncanonical argument structures. Of particular interest here are bivalent verbs taking a subject and an allative argument. Most of these verbs, including those whose use is illustrated in (3), are psych verbs expressing the stimulus as a subject and the experiencer as an allative. As illustrated below, with such verbs, both reflexive and nonreflexive possessives can bind either the subject or the allative argument.

(3) a. Mu-lle meeldi-vad kassi-d$_i$ nende$_i$/oma$_i$ iseloomu pärast.
1SG-ALL please-3PL.PRS cat-PL.NOM 3PL.GEN/REFL/POSS temper-GEN because
'I like cats because of their temper'.

b. Lille-de-le$_i$ sobi-b taeva-st alla sadanud vesi oma$_i$/nende$_i$ pehmuse
flower-PL-ALL be.suitable-3SG.PRS sky-ELA. adown fallen water.NOM POSS.REFL /3PL
tõttu väga hästi.
fragility because very well
'Rain water is suitable to flowers because of their fragility.'

(Lesage, accepted) reports the results of two psycholinguistic experiments on this noncanonical construction, showing *inter alia* that: (i) Speakers do not exhibit a categorical preference for reflexives being bound by the surface subject (resp. nonreflexives being bound by the allative argument); (ii) Binding preferences are modulated by word order, with reflexives showing a preference for an initial antecedent irrespective of its grammatical function. In the present paper, we set out to explore whether these results from comprehension experiments are confirmed in production, on the basis of a corpus study. We first train a dependency treebank on a large web corpus to help select relevant examples, which were then all validated by hand. We then annotate the examples for various syntactic and semantic properties, and run a number of logistic regression models to establish which factors influence the choice of a reflexive or nonreflexive form of the possessive. Finally we hypothesize that the observed preferences for possessive choice follow from syntactic and pragmatic constraints.

2 Data collection and annotation

The main challenge for our study is that the combination of factors we set out to investigate is too rare for data to be easily available: as is usual in languages without articles, there is no mandatory overt expression of possession in Estonian, which makes possessive forms comparatively infrequent; in addition, the construction of interest is found only with a handful of verbs.

For this reason, we relied on resources from the Universal Dependencies community to parse a large web corpus and use it for initial data selection. Specifically, we trained UDPipe (Straka and Straková, 2017) on the Estonian UD v2.4 treebank, the Universal Dependencies version of the Estonian Dependency Treebank (Muischnek et al., 2014). We then used this to parse the 1.1 billion token Estonian National Corpus (Kallas and Koppel, 2018). We relied on the morphological, POS and dependency annotation to select all sentences satisfying the following criteria:

- The sentence contains a token *v* of one of eight verbs taking an allative argument: *meeldima* 'please', *sobima* or *kõlbama* 'be suitable for', *meenuma* 'come to one's mind', *võimaldama* 'make possible', *kuuluma* 'belong', *jätkuma* 'be enough', *maitsma* 'please by its taste'.
- The sentence contains a token *p* of a reflexive or nonreflexive possessive.
- The possessive word *p* is the possessor of some noun that has *v* on its head path – that is, the noun is a direct or indirect dependent of *v*.
- The verb *v* has an allative dependent.
- The person-number features expressed on *p*, if any, are compatible with the person-number features expressed either on the verb (if any), the subject (if it is overt), or the allative dependent.

This search allowed us to retrieve 5,593 candidate examples of a use of a possessive referring to the surface subject or allative argument of a verb in the relevant construction. We then sorted through the examples by hand to eliminate the numerous false positives due to parsing errors, other uses of the form *oma*, and/or possessives with antecedents other than the two co-arguments of the verb under examination. This narrowed down the dataset to 1,307 sentences. We then classified these examples in 5 groups as indicated in Table 1. Note that what we call the surface subject is that argument which may trigger agreement on the verb. This argument, when overtly expressed, is either in the nominative or partitive case, depending on factors orthogonal to our concerns. By design, all our examples include an allative argument,[1] whereas the subject is sometimes unexpressed. Direct objects are rare in our corpus, since only the verbs *meenutama* 'remind' and *võimaldama* 'make possible' takes a direct object. There are various interesting cases where the possessive is embedded within a direct dependent of the verb; we

[1] In principle, the grammar allows for another allative dependent with the status of an adjunct, but no such case is found in our data.

Type of relation	Count
Surface subject	415
Allative argument	285
Direct object	86
Other oblique dependent	366
Embedded within a dependent	155

Table 1: Syntactic relation between the possessed noun and the head verb of the antecedent.

leave these examples aside for purposes of this paper, as they do not form a uniform class and there is not enough data for a more fine-grained classification to be informative. Hence we will focus on the 1,152 sentences corresponding to the first four row in Table 1. Since the number of possessed direct objects is low, and all cases where the possessed noun is neither the subject nor the allative argument are structurally similar, we grouped together possessed noun under 'Direct object' and 'Other oblique dependent' under a single value 'other', with 452 data points.

Each example was then annotated using a combination of information collected from dependency parses and manual work. We annotated the following:

- The type of possessive (reflexive or nonreflexive)
- The grammatical function of the antecedent (surface subject or allative argument).
- The grammatical function of the possessed noun.
- The person and number, and animacy of each argument.
- The volitional involvement in the event of the participant realized as the subject.
- The relative order of the two arguments and the relative order of the possessive and its antecedent.

3 Results

Following (Bresnan et al., 2007) and many other studies of binary alternatives in corpus data, we fitted mixed effects logit models to our data, using the *lme4* (Bates et al., 2014) and *lmerTest* (Kuznetsova et al., 2017) R packages. The dependent variable was the type of `possessive`. All candidate models treated the identity of the verb as a random effect. As for fixed effects, two important independent variables of interest here are the grammatical function of the possessed noun and that of the antecedent. It is important to note however that the values of these two variables are not independent: if the possessed noun is the allative argument (resp. subject), by design, the antecedent can only be the subject (resp. the allative argument); if the possessed noun is another dependent of the verb, then it can take either the subject or the allative argument as its antecedent. Because of this, we combined the two variables into one, whose values are noted `f1->f2`, where `f1` is the function of the possessed noun and `f2` that of the antecedent. Figure 1 shows the proportions of use of a possessive reflexive for each pair of grammatical functions. As the figure highlights, it is not obvious that differences between all 4 levels are statistically significant. Hence we used forward difference coding of the 4 values of that variable to be able to assess the significance of differences between adjacent levels.

We examined various combinations of this variable with other fixed effects, and report only on the best fit. The model parameters are shown in Table 2.

As the table indicates, we found a significant effect of the combined choice of a function for the possessed noun and a function for the antecedent for all pairs of adjacent conditions except `other->sbj` and `all->sbj`. Overall, situations where the antecedent is the surface subject favor using the reflexive form, whereas situations where the antecedent is the allative argument favor the nonreflexive. Note that the clearest effect opposes allative vs. subject antecedents, but that there is still a significant difference among examples with allative antecedents depending on the function of the possessed noun.

Non-third person antecedents comparatively disfavor using the reflexive. Contrary to our expectations, none of our various schemes for integrating an effect of animacy or volitional involvement turned out to be significant[2]. We tried various ways of taking into account word order. A binary variable indicating

[2]It would have been relevant to observe the role of case marking of the subject, reflecting the defiteness among other features,

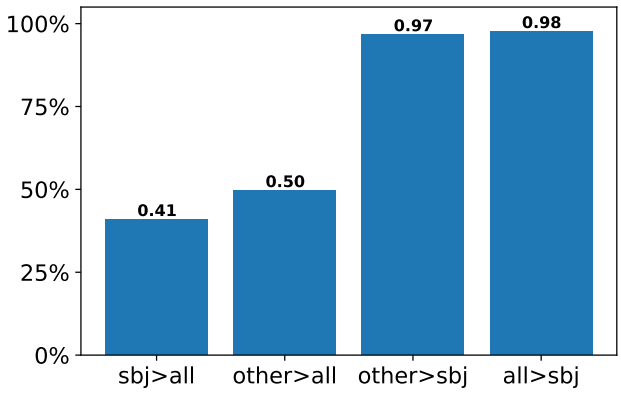

Figure 1: Proportion of use of the reflexive possessive (vs. nonreflexive possessive) for each combination of grammatical functions of the possessed noun and antecedent.

	Estimate	Std. Error	z value	p-value	
(Intercept)	0.005898	0.582356	0.010	0.991919	
`sbj->all vs.other->all`	-1.181075	0.274058	-4.310	1.64e-05	***
`other->sbj vs.all->sbj`	-2.271744	0.426328	-5.329	9.90e-08	***
`other->sbj vs.all->sbj`	-0.612193	1.331760	-0.460	0.645741	
`person=1`	-2.230623	0.296334	-7.527	5.18e-14	***
`person=2`	-1.308191	0.485689	-2.693	0.007071	**
`order=ant_first`	0.998497	0.262998	3.797	0.000147	***

Table 2: Parameters of GLMM modelling the proportion of use of a reflexive possessive.

whether the antecedent is realized before the other dependents turned out to be most relevant and lead to a significant effect: possessives that follow their antecedent are comparatively more likely to be reflexive. Finally, no significant interaction is documented for any combination of our dependent variables.

4 Discussion

Our model confirms that binding constraints on possessives are not categorical: in the constructions under examination, *oma* is often bound by a nonsubject argument, and nonreflexive possessives may (although they rarely are) be bound by the subject. We thus are in the familiar situation where one and the same constraint (reflexives tend be bound by the subject) that is categorical in some language/some part of the grammar is gradient in another language/another part of the grammar (Bresnan et al., 2001; Sorace and Keller, 2005). More importantly, binding preferences are modulated by the dependency configuration relating the possessive and its antecedent, their relative word order, and their shared person feature.

4.1 Grammatical functions

The most obvious effect is that subject antecedents show a higher preference for reflexive possessives than allative antecedents. This is just a gradient reflex of the well-known observation that relative obliqueness constrains binding (Pollard and Sag, 1992): reflexive dependents of a verb tend to be used when they are bound by a less oblique dependent, where obliqueness may be characterized using a hierarchy such as the one in (4). In the case of reflexive possessives, we submit that relative obliqueness of the possessed noun and the antecedent likewise constrains binding of the possessive.

(4) Subject < Direct object < Oblique argument < Adjunct

More subtle are the differences among contexts with allative antecedents (`sbj->all` vs. `other->all`). A likely reason for the observed difference has to do with how far one stands from a

as suggested by one reviewer, but in the construction under scrutiny, the subject is mostly nominative. We found few examples with a partitive subject that we had to remove for a more homogeneous data set.

prototypical reflexive binding situation. As we observed in the introduction, Estonian possessive reflexives tend to be bound by the subject of the local clause they occur in, which is by definition the least oblique dependent of the verb. In the `sbj->all` condition, we are departing maximally from that situation. Not only is the antecedent not a subject, it is also strongly *more oblique* than the possessed noun. Hence we have a strong expectation that a nonreflexive rather than a reflexive possessive be used. In the `other->all` condition, the situation is different. Remember that, in this condition, the possessed noun is either a nonargument oblique or a direct object; obliques are more common, and make up 65% of the data. It follows that, in a clear majority of examples, the antecedent is more oblique than the possessed noun. Hence, on average, only the strongest expectation that the antecedent be a subject, but not the weaker expactation that it be less oblique than the possessed noun, is violated in the `other->all` condition. Closer examination confirms that oblique possessed nouns are indeed driving the difference between the `sbj->all` and the `other->all` condition: whereas the proportion of reflexives is 60% for oblique possessed nouns, it drops to 22% for direct object possessed nouns.

4.2 Word order

We turn briefly to the effects of word order. As noted above, our model shows that, all other things being equal, possessives preceding their antecedents are less likely to be expressed as a reflexive than possessives that follow their antecedent.

This generalization is likely to be linked to information structure, given the tight link between word order and information structure in Estonian. The main relevant generalization here is that the dependent of a main clause verb that is realized first in linear order strongly tends to be topical (Lindström, 2005; Tael, 1988). (Bickel, 2004) suggests that, in Himalayan languages, reflexives are topic-oriented rather than subject-oriented: the reflexive tends to take the topic as its antecedent, which will coincide with the subject in most situations, but is likely not to in sentences with an experiencer expressed as an oblique. Our data supports the idea that Estonian reflexives are *both* subject-oriented and topic-oriented. While a subject antecedent favors the use of a reflexive, a topical (and hence initial) antecedent also favors such a use. Hence, where topicality and obliqueness do not align, we expect that conflicting constraints on the use of reflexives will lead to a somewhat balanced distribution of reflexives and nonreflexives.

This hypothesis helps explain the striking fact, apparent in Figure 1, that proportions of use of a reflexive reach much more extreme values when the antecedent is the subject than when it is the allative argument. Note that, unlike what happens in the canonical transitive construction, where subjects overwhelmingly precede objects, verbs with an allative argument tolerate much more easily realization of that argument before the subject (Metslang, 2013). In our data, this is true in 48% of the cases where the subject is overt. Importantly, antecedents tend to precede possessives: this is the case for 70% of allative antecedents and 84% of subject antecedents. Hence, when the antecedent is a subject, both obliqueness and topicality (as manifested in word order) favor the choice of a reflexive possessive, leading to a very high proportion of reflexives. Where the antecedent is an allative though, more often than not, obliqueness and topicality pose conflicting constraints on the choice of the possessive form: the obliqueness relation between possessive and antecedent favors a nonreflexive, while the topicality of the antecedent favors a reflexive. We conjecture that this is why, while nonreflexives are more common, reflexives are still a relevant option in most cases where the antecedent is the allative argument.

4.3 Person

We finally turn to the effect of person. As noted above, first and second person antecedents comparatively disfavor the use of a reflexive possessive. We submit that this may be due to speakers optimizing their speech for ambiguity avoidance, in accordance with Grice's maxim of manner (Grice, 1975).

To see how this plays out, let us reason first about cases in which the antecedent and possessed noun are co-arguments—that is, the `sbj->all` and `all->sbj` conditions. Example (5a) exhibits a situation where the antecedent is first person. In this situation, neither choice of pronoun form leads to ambiguity: reflexive *oma* has to corefer with the allative argument, as it is the only other referring expression in the local clause; but nonreflexive *minu* does not carry any ambiguity either, because it is explicitly 1st person singular. Now consider (5b). Using reflexive *oma* again does not lead to ambiguity. However,

if the speaker were to choose instead nonreflexive *tema*, this would lead to ambiguity between a local antecedent (namely the allative argument) and an extra-sentential antecedent. This line of reasoning should push a rational speaker to comparatively favor the use of a reflexive with third person antecedents as compared to first and second person antecedents.

(5) a. Mu-lle$_i$ meeldi-b minu/oma$_i$ naine.
1SG-ALL please-3SG.PRS 1SG.GEN/POSS.REFL wife.NOM
'I like my wife.'

b. Peetri-le$_i$ meeldi-b tema$_{?i/j}$/oma$_i$ naine.
Peeter-ALL please-3SG.PRS 3SG.GEN/POSS.REFL wife.NOM
'Peeter likes his wife.'

Note that exactly the same reasoning is valid, *mutatis mutandis*, in the `all->subj` condition. Overall then, when the possessive, possessed noun and antecedent are the only three referential expressions in the clause, pragmatic reasoning predicts a higher proportion of use of the reflexive with third person antecedents. A model identical to that above but trained on only the `sbj->all` and `all->sbj` conditions does confirm that this prediction is borne out.

If we now turn to the remaining `other->all` and `other->sbj` conditions, things are less clear, both conceptually and empirically. Conceptually, ambiguity avoidance does not make sharp predictions in such configurations. Table 3 lists all relevant configurations of person of the two co-arguments of the verb, and indicates what the ambiguity potential is depending on the choice of a possessive form; here local ambiguity is ambiguity with a clause-local antecedent, while global ambiguity is ambiguity with an extra-sentential antecedent. The clear prediction here is that, all other things being equal, reflexives should be rarest where both arguments are nonthird person (because using a nonreflexive avoids local ambiguity) and most frequent where both arguments are third person (because using a reflexive avoids global ambiguity). Where the other two situations should stand between these two extremes is unclear, in the absence of a hypothesis on the relative costs of local and global ambiguity.

Person of antecedent	Person of other arg.	Possessive type	Local ambiguity	Global ambiguity	# of observations	% of reflexives
non-3rd	non-3rd	non-reflexive	no	no	4	33%
non-3rd	non-3rd	reflexive	yes	no	2	
non-3rd	3rd	non-reflexive	no	no	43	51%
non-3rd	3rd	reflexive	yes	no	45	
3rd	non-3rd	non-reflexive	no	yes	4	95%
3rd	non-3rd	reflexive	yes	no	79	
3rd	3rd	non-reflexive	yes	yes	51	81%
3rd	3rd	reflexive	yes	no	224	

Table 3: Potential ambiguity of possessives in sentences with two candidate antecedents.

Empirical results are unconclusive. The raw counts in Table 3 clearly indicate that there is not enough data to conclude anything in the first condition, and proportions go against predictions when comparing the two last conditions. Be that as it may, a GLMM predicting possessive type on the basis of person configurations as indicated in Table 3 and grammatical function of the antecedent revealed no significant effect of person on this subset of the data. This in itself does not invalidate the idea that ambiguity avoidance constrains the choice of possessive forms: it could be that the preferences at play here are too small to be documented on a dataset of this size, or that some other factors counteract the effects of ambiguity avoidance; but it may also be that our hypothesis does not hold, and that the person effect documented in the co-argument conditions is due to some other factor.

5 Conclusion

In this paper we have used corpus evidence to explore constraints on the choice of reflexive vs. non-reflexive forms of possessives in one particular construction of Estonian. Our main empirical findings

are (i) that such these constraints are not categorical, and (ii) that separate influences of relative obliqueness, word order, and person can be documented. These results are in line with previous observations in comprehension experiments. We explored two separate but complementary lines of explanation for these findings: an interplay of grammatical relations and information structure on the one hand, and an influence of ambiguity avoidance.

On a methodological level, this paper highlights how useful the availability of dependency treebanks and parsing resources is for the linguistic study of rare syntactic phenomena in understudied languages.

References

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2014. Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Balthasar Bickel. 2004. The syntax of experiencers in the Himalayas. In Peri Bhaskararao and Karumuri V. Subbarao, editors, *Non-nominative Subjects*, volume 1, pages 77–112. John Benjamins, Amsterdam.

Joan Bresnan, Shipra Dingare, and Christopher D. Manning. 2001. Soft constraints mirror hard constraints: Voice and person in English and Lummi. In *Proceedings of the LFG01 Conference*, pages 13–32.

Joan Bresnan, Anna Cueni, Tatiana Nikitina, and Harald Baayen. 2007. Predicting the dative alternation. In Gerlof Bouma, Irene Kramer, and Joost Zwarts, editors, *Cognitive Foundations of Interpretation*, pages 69–94. Royal Netherlands Academy of Sciences, Amsterdam.

Paul Grice. 1975. Logic and conversation. In Donald Davidson and Gilbert H. Harman, editors, *The logic of grammar*. Dickenson, Ensino.

Jelena Kallas and Kristina. Koppel. 2018. Eesti keele ühendkorpus 2017.

Alexandra Kuznetsova, Per B Brockhoff, and Rune Haubo Bojesen Christensen. 2017. lmertest package: tests in linear mixed effects models. *Journal of Statistical Software*, 82(13).

Suzanne Lesage. accepted. Liage du réfléchi possessif en estonien : une approche expérimentale. *Études finno-ougriennes*.

Liina Lindström. 2005. *Finiitverbi asend lauses: sõnajärg ja seda mõjutavad tegurid suulises eesti keeles*, volume 16. Tartu Ülikooli kirjastus.

Helena Metslang. 2013. Coding and behaviour of Estonian subjects. *Eesti ja soome-ugri keeleteaduse ajakiri. Journal of Estonian and Finno-Ugric Linguistics*, 4(2):217–293.

Kadri Muischnek, Kaili Müürisep, Tiina Puolakainen, Eleri Aedmaa, Riin Kirt, and Dage Särg. 2014. Estonian dependency treebank and its annotation scheme. In *Proceedings of the 13th Workshop on Treebanks and Linguistic Theories*.

Carl Pollard and Ivan A. Sag. 1992. Anaphors in English and the scope of Binding Theory. *Linguistic Inquiry*, pages 261–303.

Antonella Sorace and Frank Keller. 2005. Gradience in linguistic data. *Lingua*, 115(1497–1524).

Milan Straka and Jana Straková. 2017. Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada, August. Association for Computational Linguistics.

Kaja Tael. 1988. Infostruktuur ja lauseliigendus. *Keel ja Kirjandus*, pages 133–143.