The competition between certain constructions involving grammatical case in Estonian

Kaarel Hänni

24.914 Spring 2021

1 Introduction

Estonian has a grammatical case system that is commonly analyzed as having 14 cases. There are a number of situations in which one meaning can be expressed by a few alternatives, each involving the same stem with a different case marker, possibly accompanied by some other postpositions. Here are 3 examples of what seems to be a very common format for this: various cases vs genitive plus a postposition. (I think that this might be expected from the fact that all cases other than nominative and partitive look like the genitive case with a merged postposition.)

(1) a. *kotis*
bag.inessive

'in the bag'

b. *koti* *sees*
bag.genitive in

'in the bag'

(2) a. *laual*
table.adessive

'on the table'

b. *laua* *peal*
table.genitive on

'on the table'

(3) a. *õpin* *eksamiks*
study exam.translative

'I'm studying for the exam'

b. *õpin* *eksami* *jaoks*
study exam.genitive for

'I'm studying for the exam'

My goal in this paper is to look at the competition within pairs of this form over time. The hypothesis that I will be testing is that this competition is regular in the following senses:

1. The competition in analogous pairs involving different nouns is similar. For example, the relative frequencies of the options meaning 'on the chair' should behave similarly over time as the relative frequencies of the two options meaning 'on the table'.

2. The competition between pairs involving different cases is also similar, in the sense that if the genitive construction is gaining ground against one case, then it is also gaining ground against other cases.

3. The fraction of words in each case over time matches the results from looking at individual pairs. That is, if the ratio of the number of word tokens in a particular case (inessive, adessive, or translative) to the number of word tokens in the genitive case (from some time period) is increasing, then the relative frequency of the genitive case version of the above pairs is decreasing.

I will later refer to these three statements as parts 1, 2 and 3 of "the regularity hypothesis".

2 The competition within particular pairs

2.1 Discussion of what exactly to study and how to study it

For the competition between particular pairs such as the ones presented in the introduction, it would be great to look at data from many centuries. Unfortunately, while there is a corpus of written Estonian available that spans around 4 centuries (which will, in fact, be discussed more and used in the next section), this corpus only has 2 million words (and what's worse, these words are not evenly distributed between centuries), and some searching suggests that this might be too few to find sufficiently many instances of the above constructions

to arrive at meaningful results. There is also a second problem: Estonian orthography before the last few centuries seems to vary wildly, and this presents complications when searching the corpus.[1]

For these reasons, I will instead be looking at a larger corpus of literary Estonian from 1890 to 2000, available here: `https://www.cl.ut.ee/korpused/baaskorpus/`. In an effort to have a decent amount of data in each time period, I will be comparing the language of 1890-1940 to the language of 1990-2000 in the analysis here. The corpus has newspapers and fiction, but I will only be looking at fiction, in order to avoid effects coming from there being a different fraction of fiction to newspapers in the corpus from the two time periods. (And I am choosing fiction instead of newspapers because that leaves me with more data.) This leads to a corpus size of about slightly less than 1 million words for 1890-1940, and slightly more than 6 million words for 1990-2000.

I encountered some problems in coming up with search strings that are specific enough to really test a competing pair, but that also give more than 0 or maybe a few results. To elaborate, the first desired condition here is that all (or at least a majority of) sentences that come up upon searching for one of the search strings in a pair should be such that both members of the competing searched pair would have the correct meaning to be substituted into the sentence. For instance, if one tries to emulate the pair (3) by searching for the string 'eksamiks' or for the string 'eksami jaoks', then we might run into trouble if all found instances of 'eksamiks' are actually part of something like 'me jäime eksamiks sõpradeks', meaning *we remained friends for the duration of the exam*, as it turns out to not be grammatical to say 'me jäime eksami jaoks sõpradeks' (with the same meaning). If the search returns many examples like this one, then the count of 'eksamiks' does not give great information about the count of 'eksamiks' that means the same and is an alternative to 'eksami jaoks', so 'eksamiks' and 'eksami jaoks' would perhaps not be competing in the right way in this case.[2] That said, as long as the total number of search results remains low, it is feasible to manually check and only count the valid ones after searching.

Unfortunately, even after a decent amount of trying different things in the corpus, I was unable to find any words to search for with a sufficient frequency of actual competing alternatives analogous to (3), 'eksami jaoks', i.e. for the translative case. There were a number of phrases analogous to 'eksami jaoks' appearing in the corpus, but all of them seemed to have token frequencies that were too low (like 0 or 1 tokens from at least one of the two time periods) after excluding the non-competing examples. So I was unfortunately not able to get

[1] To give one particularly amusing example of erratic spelling, the word that is now spelled 'ähvardus' appears as 'Echffarduß', 'æffwarduß', 'ewardus', and 'Effartuß' in the writings of **the same** 17th century author. Individual words in the corpus are marked with contemporary spellings and one can find various spellings of one word that way, but I did not figure out how to easily search the corpus for longer strings of words inputted with contemporary spellings (one can search for exact strings of characters, but the existence of many different spellings makes that approach inefficient).

[2] Actually, this did not turn out to be relevant for 'eksami jaoks', as I found no tokens of that in the corpus anyway, but it did turn out to be relevant for some other nouns.

reasonable data for this case.[3]

The story is a bit better for examples with the inessive case, such as (1). While I was not able to find exactly the example in (1), even though I initially guessed that it would be quite common, I was able to instead identify two examples with at least a few tokens from[4] each time period. Here they are:

(4) a. *metsas*
forest.inessive

'in the forest'

b. *metsa* *sees*
forest.genitive in

'in the forest'

(5) a. *elus*
life.inessive

'during life'

b. *elu* *sees*
life.genitive in

'during life'

There are still some problem with these examples. Firstly, (4b) appears in what might be the most popular children's song, whereas (4a) appears in what might be the second most popular children's song – I am not sure whether random changes in the relative popularities of these songs could be affecting things. Secondly, (5a) also means *alive*, so some manual filtering is required here (the number of results is kind of large though, 1500 for the later time period, so I used some reasonable (but crude) statistical estimate from only manually sorting through a small sample), and I would characterize (5b) as semi-idiomatic. But even with all these problems, I think the case of the inessive case turned out to be slightly better than the case of the translative case.

For the adessive case, everything was much simpler than for the inessive and translative cases, as my first two guesses for frequent words turned out to both pass my minimal frequency requirement, although the second one only barely

[3]It might have been a good idea to instead combine the counts of a small number of such constructions with different nouns, but by the time I realized this, I had already finished almost all of the writeup and the paper had already become quite unwieldy.

[4]I found these examples by searching for the Estonian word 'sees' (meaning in), and for each phrase of the form 'X sees' that was outputted and appeared plausibly frequent, I tried searching to see if it was indeed frequent enough (and my standards were low, so frequent enough meant appearing at least twice in each time period). I fear that this method for choosing the pair to compare might cause some bias in the results, but this seems like a smaller problem than the small number of tokens anyway.

passed it for the 1890-1940 time period. (There were exactly two tokens of (7b) in this period.) Here they are.

(6) a. *laual*
table.adessive

'on the table'

b. *laua* *peal*
table.genitive on

'on the table'

(7) a. *toolil*
chair.adessive

'on the chair'

b. *tooli* *peal*
chair.genitive on

'on the chair'

For each pair and each time period, I computed the ratio of the number of tokens of one member of the pair to the number of tokens of the other member of the pair. For instance, for the pair (6) and the time period $1890-1940$, there were 8 tokens of 'laua peal' and 33 tokens of 'laual', so the ratio I computed is $8/33 \approx 0.242$

2.2 Results

The results are shown in Table 1. I think the token frequencies are generally so low that we are not justified in coming to any solid conclusions from this data alone, but if something can be claimed here, it is that generally, it looks like the relative frequency of the genitive forms has decreased, as this happens for all but one pair, for which there is only a slight increase (and this increase is likely to not be statistically significant either – adding just one token to the first period would turn it into a tiny decrease). I think this can count as some very weak evidence in favor of parts 1 and 2 of the regularity hypothesis.

year	metsa sees / metsas	elu sees / elus	laua peal / laual	tooli peal / toolil
1890-1940	$2/31 \approx 0.065$	$5/106 \approx 0.047$	$8/33 \approx 0.242$	$2/11 \approx 0.182$
1990-2000	$12/506 \approx 0.024$	$56/1000 \approx 0.056$	$47/350 \approx 0.134$	$10/91 \approx 0.110$

Table 1: Ratios of frequencies of the two competing options in each pair for each of the two time periods.

3 Ratios of numbers of tokens in each case over time

3.1 Discussion of what exactly to study and how to study it

In this section, I will be describing what I did to get information about the number of noun tokens in each case over time. I will only be considering the genitive, inessive, adessive, and translative cases. These are the cases that appear in the pairs considered before.

Fortunately, there are Estonian corpora available in which words are annotated with information about which case they are in. I am aware of two such corpora. Firstly, there is a morphologically disambiguated corpus of written Estonian from after 1980, available here: `https://www.cl.ut.ee/korpused/morfkorpus/index.php?lang=en`. I will call this the *modern* corpus from now on. Secondly, there is a corpus of written Estonian from the 15th to 19th centuries that can be conveniently searched by century, available here: `https://vakk.ut.ee/`; unfortunately the interface seems to only be available in Estonian. I will call this the *old* corpus from now on. There are a number of other Estonian corpora, but unfortunately the words in these are not annotated by case. In fairness, there are some pieces of code available that should be doing the grammatical case identification automatically which could be used on these other corpora, but I did not manage to become sufficiently familiar with these tools yet, and I am also not sure if the possible errors could cause problems with the analysis (without manually checking a sample of results, which would take additional time). So I am using these two annotated corpora for the analysis to come. One possible problem that I will note but that I will not do anything to mitigate is that it could be that the two corpora contain a different selection of kinds of texts (or it could even be that the selection of texts in the old corpus could differ by century in some relevant way), and that this could be affecting the data and contributing to some of the observed differences.

I will only be looking at singular common nouns. The numbers I will report are the number of tokens in a case divided by the number of tokens in the genitive case for each time period that will be considered. There are at least two reasons for choosing this particular ratio to report, instead of e.g. the ratios to the number of all nouns. The first one is that in this paper, I am interested in the competition between various cases and the genitive case, and if we want to capture this competition in one number, this seems like a reasonable thing to consider. The second reason is that computing these numbers required fewer corpus queries. One problem with this choice is that it would also be interesting to know whether, say, the ratio of inessive to genitive is large at a certain time because inessive is gaining ground or because genitive is losing ground, but I will leave figuring this out to another paper.

Here are two other caveats that I will mention but will not do anything about. The first is the fact that among the 18th and 19th century texts in

the old corpus, only some have been annotated by case. My hope is that this is roughly a random selection of texts, and hopefully this does not introduce some additional systematic error to the analysis. The second caveat is that a significant fraction of the earlier texts seem to be written by people whose first language might not have been Estonian, e.g. there seem to be a number of religious texts translated by Baltic German priests. It is not clear if this makes the Estonian used in these texts more artificial or German-like than the Estonian that would have been written by a native speaker. (The texts sure seem strange to me, but the strangeness might well have nothing to do with the text being written in "artificial" Estonian. It could instead just be coming from old Estonian seeming strange to me.)

3.2 Results

The numerical results are presented in Table 2 and Table 3, and the same data is presented graphically in Figure 1. Looking at the plot, all three cases (inessive, adessive, translative) seem to be increasing in frequency compared to the genitive case. Most importantly for our analysis, there is an increase from the 19th century to > 1980. Together with the results from looking at particular pairs also suggesting that genitive constructions are losing ground, this provides some support to part 3 of the regularity hypothesis. However, I think this should again be treated as rather weak evidence, mainly because the number of tokens of particular pairs was small in Section 2 of this paper. It would also be better to have the years here match up more closely with the time periods in Section 2. I think that this is a limitation of the corpora I used.

year	inessive/genitive	adessive/genitive	translative/genitive
< 1600	$\frac{31}{1034} \approx 0.02998$	$\frac{91}{1034} \approx 0.08801$	$\frac{53}{1034} \approx 0.05126$
17th century	$\frac{1450}{47246} \approx 0.03069$	$\frac{3318}{47246} \approx 0.07023$	$\frac{1645}{47246} \approx 0.03482$
18th century	$\frac{318}{4160} \approx 0.07644$	$\frac{335}{4160} \approx 0.08053$	$\frac{360}{4160} \approx 0.08654$
19th century	$\frac{2658}{27473} \approx 0.09675$	$\frac{3003}{27473} \approx 0.10931$	$\frac{946}{27473} \approx 0.03443$
>1980	$\frac{8140}{39871} \approx 0.20416$	$\frac{8360}{39871} \approx 0.20968$	$\frac{4591}{39871} \approx 0.11515$

Table 2: The number of noun tokens in various cases divided by the number of noun tokens in the genitive case, from corpora restricted to particular times. The last row is calculated from data from the modern corpus; the first four rows are calculated from data in the old corpus.

year	inessive/genitive	adessive/genitive	translative/genitive
< 1600	0.02998	0.08801	0.05126
17th century	0.03069	0.07023	0.03482
18th century	0.07644	0.08053	0.08654
19th century	0.09675	0.10931	0.03443
>1980	0.20416	0.20968	0.11515

Table 3: Like Table 2 but with less clutter

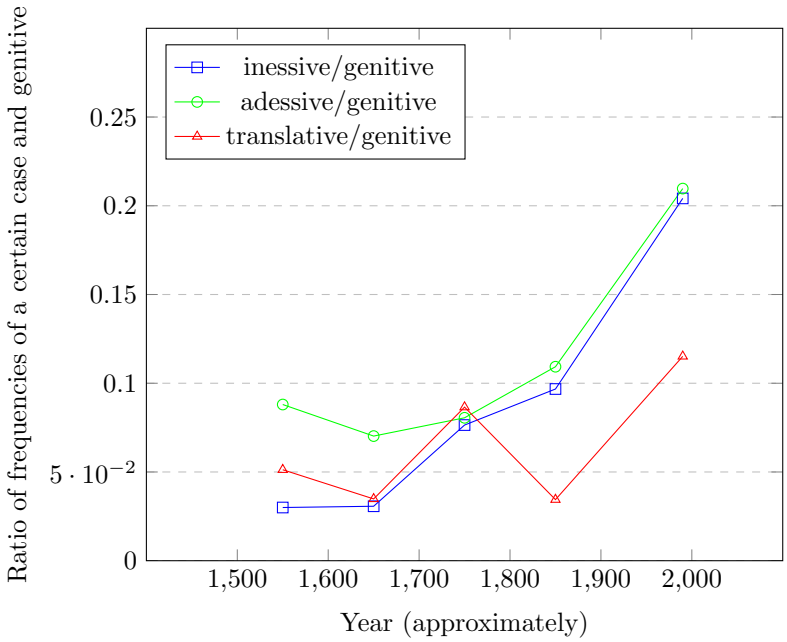

Figure 1: Ratios of frequencies of the inessive, adessive, translative cases to the frequency of the genitive case over time

4 Directions for further inquiry

It would also be insightful to analyze the number of times the Estonian word 'sees' (meaning *in*) appears in corpora at different times, as some searching indicates that a significant fraction of its appearances is in constructions like (1b) (by "like (1b)", I mean (1b) or the same construction except with a different noun), so comparing this with the frequency of the inessive case could maybe also tell us about the competition in pairs like 1. But this is complicated by the fact that whereas for early texts, a cursory search suggests that a large majority of instances of 'sees' are in such constructions, a search of later texts suggests instead that the word 'sees' is appearing in other constructions a fair amount, as well. It would be good to distinguish between constructions like (1b) and other

appearances of the word ‘sees’; one idea would be to do that by conditioning on the case of the previous word, but I have not tried this yet. With even less of a proper check, I would expect a similar story to hold for the word ‘peal’ (meaning on) and constructions like (6), but not for the word ‘jaoks’ (meaning for) and constructions like (3). The problem with the latter is that I would guess that there are many other uses of ‘jaoks’, and I find it likely that the examples we want to find out about will only contribute a minority of all instances of ‘jaoks’.

There are also other pairs of competing alternative constructions involving different cases:

(8) a. *lapsena*
child.essive

‘as a child’

b. *kui laps*
as child.nominative

‘as a child’

c. *laps olles*
child.nominative being

‘as a child’

(9) a. *lähen arstile*
go doctor.allative

‘I’m going to see a doctor’

b. *lähen arsti juurde*
go doctor.genitive to/at/around

‘I’m going to see a doctor’

(10) a. *õpin arstiks*
study doctor.translative

‘I’m studying medicine’

b. *õpin arsti*
study doctor.partitive

‘I’m studying medicine’

One complication with studying (10) is that it is hard to get data on it, as searches suggest that (10b) might be a recent innovation and as it is also quite

colloquial. Also, while other similar sentences seem more or less possible to me (e.g. replacing 'doctor' with 'lawyer' or 'veterinarian'), the version with 'doctor' seems to be the only one that comes up when googling.

Example (9a) is also essentially the unique example of that sort. In fact, I think it is not acceptable to substitute 'doctor' with 'lawyer' or 'hairdresser' in that example[5]. Maybe 'arst' is a fairly frequent noun and the fact that interesting things are happening to it is related to that. Maybe it also has something to do with the possibility that there could be many ways to interact with doctor-related things, compared to e.g. the number of ways in which one usually interacts with plumber-related things. Or maybe I just randomly happened to think about a few examples involving 'arst', and maybe there are really many other exciting constructions involving other nouns, as well.

Finally, as always, it would be good to repeat this analysis, and especially the consideration of competition within particular pairs in Section 2, with larger corpora.

5 Errata

I noticed later that I had forgotten a space when searching the corpus and one entry in the top left cell of Table 1 is off by 1: it should be 1/31. Unfortunately, fixing this would require a decent amount of rewriting because of my self-imposed requirement of having at least 2 tokens in each time period...

[5] although thinking about this for multiple days has almost begun to convince me that one can also say it with 'hairdresser'