ESUKA – JEFUL 2023, 14–3: 5–44

A CORPUS STUDY OF GRAMMATICAL CASE FORMS IN WRITTEN AND SPOKEN ESTONIAN: FREQUENCY, DISTRIBUTION AND GRAMMATICAL ROLE

Merilin Miljan, Virve Vihman
University of Tartu, EE
merilin.miljan@gmail.com, virve.vihman@ut.ee

Abstract. In this paper, we present the results of a corpus study investigating the distribution of the three grammatical cases in Estonian (nominative, genitive, partitive) and the factors affecting the interpretation of syntactic role for nouns marked in these cases. Unlike previous studies, which have focussed on the properties of grammatical relations, we take the perspective of morphological case, and investigate the relative frequency of each case in both written and spoken corpora, according to the encoded grammatical roles, referential properties (animacy, number, countability) and syntactic context (word order, transitivity), as well as probing the differences according to register. We find that each case is prototypically, but not reliably, associated with a particular grammatical role, and that a cluster of features are available to assist speakers in identifying the function of a case-marked noun.

Keywords: corpus analysis, nominative, partitive, genitive, grammatical relations, Estonian, written language, spoken language

DOI: https://doi.org/10.12697/jeful.2023.14.3.01

1. Introduction

In this paper, we pursue the relationship between case and grammatical relations. More specifically, we investigate to what extent Estonian morphological case-marking serves to encode particular grammatical functions. Unlike previous studies in Estonian linguistics, which focus on the properties of *grammatical relations* (e.g., Rajandi & Metslang 1979; Metslang 2013; Ogren 2018), we take the perspective of *morphological case* as a starting point; moreover, we examine its distribution by grammatical relations and referential properties in

both written and spoken corpora. Our aim is, first, to map the distribution of case usage in the corpora, and second, to examine the degree to which morphological case-marking can be said to signal grammatical functions.

In the literature, case is often equated with grammatical relations, e.g., nominative as subject case and accusative as object case.[1] The classification of languages by their argument case-marking patterns also derives from this assumption: when the subject of a transitive clause and the subject of an intransitive are both encoded with nominative case, and differentiated from accusative objects, we classify these fundamentally as accusative languages, contrasted with ergative languages.

While morphological case-marking helps establish the syntactic relationship of the case-marked noun to its head (e.g., the verb), often no straightforward one-to-one correspondence holds between grammatical functions and morphological case-marking (Blake 2001). This is illustrated for Estonian in Figure 1, showing that grammatical relations cannot be defined in terms of case marking only: nominative case cannot always be used to identify the grammatical roles of subject and object, as nominative objects exist in the language as well as nominative subjects. Partitive case, too, marks both subject and object arguments. Genitive encodes possessors, postpositional complements, and other functions, in addition to the core argument functions shown in Figure 1 (see Table 2 in section 2.2).

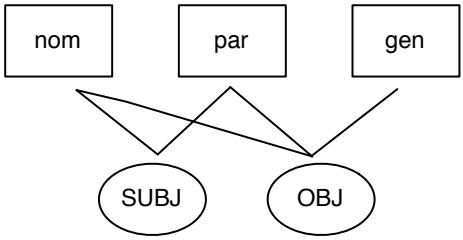

Figure 1. The coding of grammatical functions in Estonian.

1 See, e.g., Hiietam (2003, 2004) with regard to Estonian, or consider the concept of abstract structural case in Chomsky's Government and Binding theory.

This naturally raises the question of what role is played by morphological case in the grammar: how are nominative, partitive, and genitive distributed with regard to grammatical functions in language use? How informative is case form for grammatical role interpretation, especially in a language without a neat, one-to-one correlation between morphological case and grammatical functions? In Estonian, speakers identify these functions with ease, despite the complex case form to function mappings and flexible word order. Another related question, then, concerns what other factors assist in identifying grammatical relations. We look at the distribution of case-marked nouns in Estonian according to their grammatical functions, as well as the interaction of the cases with other factors (word order, animacy, polarity, countability and number), and to what extent these determine the mapping to grammatical functions. Morphological case is important for differentiating argument roles. Since differentiation is most relevant in clauses with more than one argument, we also examine the relation between transitivity and case-marking.

Our aim is to identify what type of information is associated with morphological case in Estonian: which grammatical roles are most frequently associated with each of the three grammatical cases, and what other factors are relevant in the use of nominal case-marking. Turning this inside-out, we wish to assess what sort of information may be associated with a case-marked noun. We therefore include *all* nominative, genitive and partitive nouns in our sample of clauses, drawn from an equivalent number of clauses in the written and spoken corpora. We are not investigating particular lexemes or particular grammatical constructions, and the full set of nouns in both corpora is of a similar size, allowing us to draw inferences about the use of each of the three case forms without conflating frequency with proportional preference (Biber 2012). In this paper, we are mostly interested in painting a descriptive picture of the use and interpretation of nouns in the three most frequent cases in Estonian. In the following section, we describe the nominal case system. Section 3 presents background on the factors which may affect the use and interpretation of the three cases: animacy, number, word order and register. We describe our method in section 4, and we present the results in section 5. Section 6 discusses theoretical implications of the findings and concludes.

2. Overview of the Estonian case system

This section provides an overview of the basic properties of the Estonian nominal case system. The main focus is on the three grammatical cases – their form, functions, and case syncretism.

2.1. Nominal paradigm

Estonian distinguishes 14 cases in singular and plural. Traditionally, these are divided into grammatical (nominative, genitive and partitive) and semantic cases, although some of the latter also bear various grammatical functions. In this paper, we are concerned only with the three core grammatical cases. The partial declension paradigm is given in Table 1, exemplified with two nouns representing different declension patterns, *raamat* 'book' and *sõber* 'friend'.

Table 1. Partial case paradigm: the three grammatical case forms of two Estonian nouns.

CASE	NUMBER			
	EX.1: 'book'		EX. 2: 'friend'	
	SG	PL	SG	PL
NOMINATIVE	*raamat*	*raamatu-d*	*sõber*	*sõbra-d*
GENITIVE	*raamatu*	*raamatu-te*	*sõbra*	*sõpra-de*
PARTITIVE	*raamatu-t*	*raamatu-id*	*sõpra*	*sõpru*

Nominative is morphologically unmarked; genitive and partitive are distinguished from the other cases by affixation, phoneme deletion/substitution, stem gradation or other stem changes. The genitive form is used as the stem in the formation of other morphological cases, in both singular and plural. As case formation is not the subject of this paper, we will not discuss the declension classes or paradigms further (for more details, see Blevins 2008; Kaalep 2010, 2012; Viht & Habicht 2019).

2.2. Paradigmatic case vs case syncretism in Estonian

Typologically, Estonian is predominantly a dependent-marking language, which indicates grammatical functions with nominal and pronominal case forms, and is classified as a nominative-accusative

language. This classification turns out to be more complicated, however, as discussed below.

All three grammatical cases in Estonian are versatile. They can appear in a variety of syntactic functions (see Table 2), marking core grammatical relations such as subject and object (ex. 1 from the written corpus,[2] WRI), as well as adjuncts, as shown in examples (2)–(4) below. Note that examples (2)–(4) derive from Internet searches rather than the corpus from which all other examples in the paper are extracted, indicating that the adjunct uses are comparatively infrequent. In Table 2, 'semantic effect' refers to the interpretation of a case-marked noun as bounded or unbounded; this is explained below.

Table 2. Functions of grammatical cases in Estonian.

CASE	GRAMMATICAL FUNCTION	SEMANTIC EFFECT
NOMINATIVE	subject/ object/ predicative/ adjunct	none/ bounded
GENITIVE	adnominal/ possessor/ object/ adjunct/ complement of adpositions	none/ bounded
PARTITIVE	object/ subject/ complement of adpositions/ adjunct	none/ unbounded

(1) ***Superkangelane*** *kehita-s* *ükskõikselt* ***õlgu***. (WRI)
superhero.NOM.SG shrug-PST.3SG indifferently shoulder.PL.PAR
'The superhero shrugged [his/her] shoulders indifferently.'

(2) Nominative NP as adjunct:
Vaikne *habetunud* *mees,* *kes* *tööta-b*
quiet bearded man.NOM.SG who work-PRS.3SG
[pika-d ***päeva-d]*** *metsa-s...* [3]
[long-NOM.PL day-NOM.PL] forest-SG.INE
'A quiet, bearded man, who works for long days in the woods…'

2 All examples in the paper, apart from (2)–(4), derive from the written (WRI) and spoken (SPO) corpora, as marked (see section 4). Examples (2–4) are taken from Google searches and their sources are noted in footnotes.

3 kaupokikkas.eestifoto.ee/blog/toeline-metsamees (Accessed 27.09.2022).

(3) Genitive NP as adjunct:
Murelik *ema* *oota-s* ***tunni****,*
worried mother.NOM.SG wait-PST.3SG hour.GEN.SG
oota-s *kaks,* *kolm (...)*[4]
wait-PST.3SG two three
'The worried mother waited [for an] hour, waited two, three...'

(4) Partitive NP as adjunct:
Vene *spioon* *esine-s* ***[pikki*** ***aastaid]***
Russian spy.NOM.SG perform-PST.3SG [long.PAR.PL year.PAR.PL]
glamuurse *ehte-disaineri-na*[5]
glamorous.GEN.SG jewelry-designer-ESSIVE.SG
'A Russian spy pretended for many years to be a glamorous jewelry designer.'

The subject function can be indicated by nominative or partitive (Erelt, Metslang & Plado 2017; Lindström 2017b); nominative subjects occur in examples (1)–(4), and partitive subjects are shown in (5a–b). Note that the partitive-marked subject does not agree with the verb in number. Lindström (2017b) found that partitive subjects occur most frequently with negated verbs, as in (5c), although, as a rule, the alternation between nominative and partitive is taken to express 'bounded' vs 'unbounded' interpretation of the case-marked noun, respectively (see Table 2 above). Thus, effectively, both (5a) and (5b) would also be acceptable with nominative marking instead of partitive.

(5) a. *tolle-ga* *on* *`ka* *jälle* ***prob`leem-e*** (SPO)
that-COM be.3PRS also again problem-PAR.PL
'There were problems with that one again too.'

b. *Kohe* *hakka-s* *kuldse* *paviljoni*
immediately begin-3SG.PST golden.GEN.SG pavilion.GEN.SG
katuse-st *kerkima* ***suitsu*** (WRI)
roof-ELA.SG rise.INF smoke.PAR.SG
'Smoke began to rise right away from the roof of the golden pavilion.'

4 www.ohtuleht.ee/133649/ootasin-kiirabi-ule-kuue-tunni (5 Jan. 2003), Accessed 27.09.2022.

5 www.delfi.ee, 30.08.2022

c. *Mingi-t* ***korda*** *tema* *juhis-te-s*
some-PAR.SG order.PAR.SG 3SG.GEN direction-PL-INESSIVE
ei *tundunud* *ole-vat.* (WRI)
NEG seem.PST.PTCP be-PRS.QUOTATIVE
'There appeared to be no order in his/her instructions.'

The object function can be signalled by all three grammatical cases. Objects marked by genitive singular or nominative plural, as in (6a–b), are referred to as 'total objects' in the linguistic tradition. Genitive is not used to mark plural objects (6a is from the written corpus, WRI, while 6b is an example constructed, CONS, to demonstrate the contrast between singular and plural total objects). Objects marked by partitive case, as in (6c) and (7), are known as 'partial objects' (again, SPO is attested in the spoken corpus; CONS is constructed).

(6) a. *Me* *valluta-me* ***[poliitilise*** ***maastiku]***. (WRI)
1PL.NOM conquer-PRS.1PL [political.GEN.SG landscape.GEN.SG]
'We [will] conquer the political landscape.'

b. *...valluta-me* ***[poliitilised*** ***maastikud]***. (CONS)
…conquer-PRS.1PL [political.NOM.PL landscape.NOM.PL]
'We [will] conquer the political landscapes.'

c. *...valluta-me* ***[poliitilisi*** ***maastikke]***. (CONS)
…conquer-PRS.1PL [political.PAR.PL landscape.PAR.PL]
'We are conquering (the) political landscapes.'

(7) a. *`mina=sis* *sõrm* *`suu-s […]* *`kuula-sin*
1SG.NOM then finger mouth-INE listen-PST.1SG
[seda ***juttu]***. (SPO)
[this.PAR story.PAR.SG]
'So then I was listening to this story with my finger in my mouth.'

b. *... kuula-sin* ***[neid*** ***jutte]***. (CONS)
listen-PST.1SG [those.PAR story.PAR.PL]
'…[I] was listening to those stories.'

Objects in partitive case, as in (6c), are associated with an unbounded interpretation, applying to the case-marked noun or the predicate (verb phrase), hence the reading of 'partial' affectedness and the term 'partial

object'. In contrast to this 'partial' interpretation, objects that are marked by the other cases (genitive, nominative) are taken to have a 'bounded' interpretation: a non-partial reading indicating affectedness of the whole object and completedness and perfectivity of the predicate, hence the label 'total object' (see Metslang 2017; Tamm 2007). While aspectual transitive verbs allow object alternation between 'partial' and 'total objects', 'partitive verbs', as in (7), denote a class of transitive verbs which only occur with partitive objects, or have restrictions on occurrence with total objects (see Tamm & Vaiss 2019). The use of the terms 'total' and 'partial' object is an example of how *encoding* by case-marking can reify syntactic functions: two separate syntactic categories ('total' and 'partial' object) are assumed instead of one object relation. Some researchers argue that both the cases marking 'total' object (genitive in singular, nominative in plural) are realizations of a dedicated object case, i.e., accusative (Hiietam 2004; Norris 2018). In this paper we do not discuss the accusative analysis of Estonian objects.

It is worth highlighting the fact that nominative occurs with the subject, some adjuncts (2) and plural objects (6b), as well as the singular object of imperative verbs (8), objects of some non-finite verbs (9) and impersonal constructions (10). In all of these contexts, nominative alternates with partitive, and genitive case is judged to be ungrammatical.

(8) *pane* ***`uks /*** ***(\* ukse)*** *`kinni.* (SPO)
put.IMP.SG door.NOM.SG door.GEN.SG closed
'Close the door.'

(9) *ükskord* *tule-b* ***[see tühi purk] /*** ***(*[selle tühja purgi])***
sometime come-3SG [this empty jar].NOM.SG / GEN.SG
ära *koristada.* (WRI)
PRF clean-up.INF
'Sometime someone will have to clean up the empty jar.'

(10) ***operatsi`ooni-aeg /*** ***(* operatsiooniaja)*** *pan-di.* (SPO)
surgery-appointment.NOM.SG appointment.GEN.SG put-IMPERS.PST
'[They] scheduled an appointment for the operation.'

Thus, *from the perspective of morphological case*, nominative occurs in a variety of grammatical functions. In contrast, from *the perspective of grammatical relations*, case homonymy (syncretism) is often used

to describe cases which encode both subject and object: for example, nominative case encoding a subject function is perceived as the 'true' nominative and its occurrence in the 'total' object function is taken as an instance of accusative marking, homophonous with nominative case (see, e.g., Hiietam 2004).

Likewise, *from the perspective of morphological case*, genitive case in Estonian occurs on possessives, (semantically agentive) modifiers of non-finite verb forms (e.g. ***[isa** [praetud]] pannkoogid* 'father.GEN.SG fried-PTCP pancakes': pancakes fried by father), modifiers of adjectives, complements of postpositions, the singular 'total' object, and adverbials. *From the perspective of grammatical functions*, again an implicit distinction is often made between the adnominal genitive and the genitive as a form of accusative encoding the object function (Hiietam 2004; Norris 2018). When authors assume identity between a case and the grammatical function it marks, they tend to assume case syncretism: the starting point is the grammatical functions, which are signalled by cases, and so comparisons are drawn between properties not of the case-marked nouns but of their grammatical roles. Thus, the genitive on modifiers is considered the 'true' genitive, whereas the genitive on the object is taken to be a form syncretic with 'true' object case (accusative). Our stance is that the grammatical cases in Estonian are polyfunctional, as is characteristic of morphological case (Blake 2001).

Previous studies have either taken the perspective of grammatical function, or else investigated case form frequency without mapping them to function. Kaalep (2010), an example of the latter, investigates case forms in the morphologically tagged corpus of written Estonian; the aim of that study was to obtain an overview of the relative frequency of case forms. Hence, distinct word forms were counted, but not the usage of those forms to encode syntactic functions; some frequencies from that study are given in section 3.5. Hennoste (2004) focuses on case form frequency in spoken Estonian and lists the most dominant grammatical roles for *singular* nouns in four cases (nominative, genitive, partitive, and adessive) in the subcorpus of colloquial Estonian; while some grammatical functions are listed, no counts of form-function mappings are provided there either (Hennoste 2004: 23).

In-depth descriptions of the properties of particular arguments in various constructions in Estonian abound. For example, the frequency of 'total' objects in Estonian literary texts was investigated as early as Tauli

(1968) (see below); a qualitative analysis of genitive-marked modifiers of non-finite verb forms has been conducted by Sahkai (2011); the frequency of use of partitive subjects in Estonian dialects was investigated by Lindström (2017b). Metslang (2013) examined factors determining the category of subject, and compared the encoding of the core arguments in written Estonian according to the grammatical relations of transitive subject, intransitive subject, object, and the single argument of an existential clause.

We take the case-marked forms as a starting point and ask how reliably or variably the cases realise grammatical relations. We investigate how they are used in the written and spoken corpora to express various grammatical relations, and their co-occurrence patterns with particular semantic and morphosyntactic properties. To our knowledge, no such study has been carried out on grammatical cases in Estonian.

3. Factors potentially affecting case marking

We know from decades of work on grammatical models and the role of frequency in language use (see Divjak 2019 for an overview) that speakers use probabilistic information to guide their language usage, and that the frequency of occurrence of linguistic elements affects how language is represented and processed. The relative occurrence frequencies of word forms or constructions in a corpus can be informative regarding the relative ease of recognising or producing them. Speakers are likely to generalise information regarding the frequency of co-occurrence, not only of words and structures, but also of the factors involved in choosing one form or another. Over time, speakers' internal probabilistic models of language use will feed back into the grammar of the language itself. In this study, we attempt to capture a snapshot of the frequencies of the three grammatical noun case forms, with the understanding that this underlies their processing and use.

Case-marking, especially case-marking of the core arguments, is often sensitive to referential properties of the noun (such as animacy), and other categories marked on the noun (such as number). In this study, we look at the relative importance of these factors for the use of the three morphological cases under investigation. We also look at how the morphological encoding of grammatical functions interacts with word

order, or more specifically, the relative order of the verb and its arguments, and the number of arguments expressed in a sentence (in section 5.2), in order to determine the effect of these two clause-level properties on case-marking. The effect of register is included throughout. For context, in this section we briefly discuss these factors and their effects on case-marking, as reported in the literature. Note that we also include polarity in section 5.3, since object arguments in the scope of negation invariably receive partitive case-marking. We do not discuss polarity any further here; the only variability found in the context of negative polarity is in subject marking in existential or presentational sentences (for more detail, see Lindström 2017b; Tamm 2015).

3.1. Animacy

In Estonian, case-marking is obligatory for all nouns, irrespective of the animacy of their referents. There is no animacy-based case-marking of arguments, despite the existence of case alternation shown above in (6), and animacy-based voice alternations (see, e.g., Torn-Leesik 2009; Nelson & Vihman 2018). First and second person pronouns follow different object case-marking patterns; person marking is subsumed under some versions of the animacy hierarchy, but we will not pursue this further here (see, e.g., Foley & Van Valin 1985; Siewierska 1999).

Although animacy is a referential property of nouns, it functions as a relational feature in argument interpretation (Dahl 2008). That is, animacy becomes decisive when we need to map more than one argument to grammatical relations in a clause (e.g., Wang et al. 2013). For example, arguments referring to human or animate entities are often assigned the subject function by default in clauses with two arguments differing in animacy (e.g., Primus 2013). Specific mechanisms for encoding argument roles become crucial when both the subject and object arguments are animate, as these qualify equally well for the subject role (Meir et al. 2017). This may result in differential object marking (DOM, as in Spanish), but the DOM system in Estonian is not sensitive to animacy.

Animacy has also been shown to play a key role in determining the syntactic function of an argument with ambiguous case-marking (as, e.g., nominative, genitive and partitive in Estonian are, see Figure 1). For example, Schlesewsky & Bornkessel (2004), examine German constructions with nominative and dative arguments. Based on

case-marking, either of these arguments could be interpreted as the actor (subject) or the undergoer (object). They found that animate entities are preferred as actor (subject) arguments more often than inanimate ones, regardless of the case-marking (nominative vs dative). It is important to note that animacy also interacts with word order, or argument alignment. Specifically, it has been shown that the first argument of a sentence is typically animate and the second one inanimate (e.g., Bader & Häussler 2010); this is discussed further in section 3.3. We expect to see a similar result in our study with regard to two-argument clauses.

3.2. Number and countability

Morphological number in Estonian involves the overt marking of plural nouns. This marking for plural does not interact with animacy, unlike languages where number marking is restricted to animate NPs (see, e.g., Comrie 1989). As for the relationship between number, grammatical role assignment, and case-marking, it has not received much attention in the literature, where the main focus is on subject-verb agreement effects (see Lago et al. 2015 for a review). In Estonian, the number distinction is also relevant in assigning object case. The alternation between partial and total objects discussed in section 2 is realised differentially depending on number: nominative case is assigned for (total, affected) *plural* objects, and genitive for (total, affected) *singular* objects, as shown in (6a–b) above. Nominative objects (either singular or plural) are also used in subjectless constructions, see examples (8)–(10) above. The choice is less categorical, however, than is sometimes claimed by descriptive grammars: Ogren (2015a,b) reports great variability in object-marking in certain infinitival constructions, but no studies have examined whether objects which ostensibly require nominative are also variably case-marked in other contexts, such as plural number.

A semantic factor which is closely related to number in Estonian and plays an important role in the use of partitive case is countability. Partitive case can normally be used on mass or plural nouns that allow readings of partitivity or unboundedness. Singular count nouns are often pragmatically odd under the reading of partitivity without a supporting context, so partitive-marked singular nouns tend to enter into a predicate modifying relationship as the object argument; the interaction with verb

semantics yields the interpretation of unboundedness or imperfectivity at the level of the verb phrase (see, e.g., Verkuyl 1993). Nevertheless, a subset of verbs (e.g., *leidma* 'to find') avoid partitive-marked singular count nouns as their object complement in affirmative contexts, and select for plural or mass nouns in partitive. On the other hand, as mentioned in section 2.2, there is a subset of verbs (e.g., *nägema* 'to see', *armastama* 'to love', *kuulama* 'to listen to') which always select for partitive objects, regardless of their number and countability properties, as in (7) above. These 'partitive verbs' are a prominent group in Estonian (Vaiss 2004) and, after negation, contribute most to the occurrence of partitive on the object argument (Metslang 2014).

3.3. Word order

Estonian has flexible word order at the clause level, as expected, considering its rich case-marking system. Although all six combinations of S, V and O are possible, the neutral word order is SVX (including SVO; Lindström 2017a: 547). The flexibility of word order in Estonian pertains only to the clause level, whereas at the phrase level, word order is rigid.

To our knowledge, there are no studies which look at animacy effects on *case-marking* and argument alignment in Estonian (but see Lindström 2002: 94, who relates animacy to subject-verb inversion; also see production studies by Miljan, Kaiser & Vihman 2017; Kaiser, Miljan & Vihman 2020). Cross-linguistically, it has been established that animacy (and specifically, humanness, cf. Helasvuo 2001: 80 on Finnish) is a central factor in determining word order in speech production. Estonian follows the general pattern of SVO and SOV languages showing human/animate-first order. Higher animacy nouns in the first argument position are typically aligned with the grammatical role of subject (as mentioned in section 3.1). As Meir et al. (2017) argue, the human-first preference is not reducible to the grammatical or semantic role of the arguments, but to the conceptual salience of the participants: more salient (human) entities in an event are more likely to occur before less salient (inanimate) ones, and thus they are more likely to occur in early sentence positions. Lindström (2002: 95) and Huumo (1995) indicate that in Estonian and Finnish, animacy has a greater effect on word order in spoken than written language.

In this study, we examine the interaction of argument order and animacy in Estonian, to establish whether clauses with two arguments exhibit animate-first order, meaning that OS sentences (clauses with the object before the subject) would be more likely to have animate rather than inanimate objects, just as the more common SO is more likely to have animate subjects.

3.4. NP form: Personal pronouns vs lexical nouns

Although case-marking in Estonian is not overtly dependent on nominal categories, the grammatical case paradigm for personal pronouns differs from lexical nouns. In contexts where object nouns are in nominative case, first and second person singular pronouns receive partitive marking; object pronouns in plural always receive partitive marking (Metslang 2017: 272–273). That is, first and second person pronouns do not occur as *nominative* objects like other nominals do, shown in (11a–b). In contexts where lexical nouns as objects receive only genitive case, first and second person singular pronouns show variation between genitive and partitive (11c).

(11) a. *Oli* *mis* *oli,* *Balti* *Kett* *liit-is*
be.3SG.PST what be.3SG.PST Baltic Chain join-3SG.PST
meid/ (*meie) *jäädavalt.* (WRI)
1PL.PAR/ NOM permanently
'Be that as it may, the Baltic Chain unified us for good.'

b. *aga* *kuidas* ***mind/ (*mina)*** *ikka* *maha*
but how 1SG.PAR/ NOM continually down
jäeti? (WRI)
leave.IMPERS.PST
'But how was it that I was always abandoned?'

c. *Nad* *jät-sid* ***mind ~ mu / sind ~ su / sõbra/ *sõpra***
3PL.NOM leave-3PL.PST 1SG.PAR~GEN/2SG.PAR~GEN/ friend.GEN/PAR
maha.
down
'They abandoned me / you / [their] friend.' (adapted from Metslang 2017: 273)

Thus, in contexts where objects take nominative marking (plural, imperatives and impersonals), the distribution of the three grammatical cases may be 'biased' toward partitive object marking in the case of pronominal objects.

3.5. Register

Register is known to affect language usage, but is often overlooked as a factor in corpus studies (Biber 2012; Biber & Conrad 2019; Szmrecsanyi & Hinrichs 2008). Biber (2012) details how conflating results from spoken and written corpora may lead to two possible problematic results: veiling the differences between the registers, and constructing a model which corresponds to neither register.

Differences in register may well explain some of the variance we see in the reported proportional usage of partitive versus genitive nouns in earlier empirical studies of Estonian, for instance, although different authors approach the question from different angles. Tauli (1968), in a quantitative study of objects in literary texts, found that the vast majority of 2,252 direct objects were in partitive case, with only 29% marked as 'total' objects (n = 644).

More recently, Kaalep (2010) examined the relative overall frequencies of noun cases in an automatically parsed written corpus, and found much higher frequencies for genitive than partitive nouns. As shown in Table 3, the distribution of cases is more balanced for plural than singular nouns.[6] The greater proportion of *plural* partitives may result from an association between partitive case, plurality, partitivity and unboundedness.

6 It is possible that this difference between singular and plural may derive in part from inaccurate automatic parsing, as singular paradigms involve syncretism in many declension classes, which can lead to misanalysis, whereas this is less of a risk for the more agglutinative plural forms. On the other hand, both the functions and semantics of the noun cases are different in singular and plural, and the differing distribution may also be related to syntactic or semantic difference in use.

Table 3. Frequency of noun cases in an automatically parsed corpus from Kaalep (2010).

CASE	NUMBER OF TOKENS (% OF ALL CASE FORMS)		
	SG	PL	SG & PL
NOMINATIVE	10,686 (28%)	3,502 (26%)	14,188 (28%)
GENITIVE	7,654 (20%)	2,744 (21%)	10,398 (20%)
PARTITIVE	4,711 (13%)	2,587 (19%)	7,298 (14%)
OTHER	14,682 (39%)	4,503 (34%)	19,185 (38%)
TOTAL	37,733 (100%)	13,336 (100%)	51,069 (100%)

Kaalep's results are suggestive, but they do not include spoken data. Granlund et al. (2019) examined the distribution of case-marked nouns in corpora of child-directed speech in three languages, including Estonian; their sample was limited to singular nouns. To address the issue of syncretism across case forms in certain declension classes, they interpolated frequencies for noun forms which were potentially ambiguous. They found, for this subregister of spoken language, 26% nominative, 20% genitive, and 18% partitive *singular noun* forms: these are similar proportions to Kaalep's (2010) findings for *plural* nouns.

Our study intends to give a fuller picture of the distribution of noun cases by: (i) including both spoken and written registers, and explicitly comparing the two; (ii) analysing nouns coded manually in context, thus reducing the risk of erroneous parsing and sidestepping the potential ambiguity caused by syncretic case forms; (iii) directly addressing the interaction between case form and the potentially relevant factors of: animacy, number and countability, word order, and transitivity (only for arguments); and finally, (iv) coding grammatical roles.

4. Method

We compiled a sample including an equal number of clauses from written and spoken corpora. Most of the syncretism found in the Estonian nominal paradigm is precisely among the three grammatical cases of interest here, and so any frequency analysis based on automatically parsed text is prone to error. For this reason, we chose to extract and manually code two language samples, one of written fiction and one of spoken language. This may prove to be a useful sample with which

to compare automatically parsed (and interpolated or otherwise automatically disambiguated) text. The coded data is openly accessible in the University of Tartu's data repository at (https://doi.org/10.23673/re-429).

The written data were extracted randomly from the Fiction subcorpus of the University of Tartu's Balanced Corpus of Written Estonian (5 million words in Fiction), using an online search engine (https://cl.ut.ee/korpused). The spoken data were also randomly drawn from the University of Tartu's Corpus of Spoken Estonian, maintained by the research group of Spoken Estonian (not publicly available at the time of coding). Our spoken language selection derives from a subset of everyday (face-to-face and telephone) conversations. The written corpus includes 751 clauses, and the spoken corpus includes 758 clauses.

We included all clauses which contained a finite verb and at least one nominal in one of the three grammatical cases under investigation. We coded each nominative, partitive and genitive NP within those clauses, to capture the full range of grammatical roles assigned to these nouns, rather than limiting our sample to a set of typical grammatical roles. Each clause was then manually coded for: clause type (declarative, interrogative, imperative, exclamative), constituent order, and polarity (negative, affirmative). All nouns in any one of the target cases under investigation were included in the noun sample. Each nominal was coded for referential form (lexical noun,[7] pronoun), case (nominative, genitive, partitive), number (singular, plural), countability (count, mass), animacy (animate, inanimate, other), and grammatical role (30 grammatical roles were initially distinguished, with the most frequent roles being: subject, object, adjunct (adverbial, gerund), adpositional complement, verb complement, and possessor). In the overview given below of the distribution of nominals in the three cases, we present tables and visual figures where appropriate; additional tables and raw figures can also be found in the data repository (https://doi.org/10.23673/re-429). Where relevant, we test the significance of differences in distribution according to the independent variables by applying chi-squared tests. In section 5.3, we apply Multiple Correspondence Analysis (MCA)

7 Quantity and number phrases like *mitu päeva* 'several day.par.pl' and *kaks bandiiti* 'two bandit.par.sg' were initially coded separately, but were grouped with lexical nouns in the analysis. Our data included 21 of these.

to the entire dataset, to visualise frequency-based associations within the dataset. MCA is an exploratory technique for identifying relationships in a large set of categorical variables. This technique produces a two-dimensional plot of the co-occurrence frequencies converted to (Euclidean) distances, allowing us to analyse interactions between two or more categorical variables. We used the *factoextra* package and the *fviz* function in R.

After coding all nouns in the corpus samples, our data include 2,370 nominals: 1,331 from the written corpus (WRI) and 1,039 from the spoken corpus (SPO). Overall, across all clauses in both corpora, the proportional use of each case was: 65% nominative, 18% genitive and 17% partitive. The predominance of nominative NPs was expected (see Table 3 from Kaalep 2010 above). Interestingly, however, the distribution of cases shows an effect of register, meaning that the overall percentages might be construed as misleading (as per Biber 2012).

Table 4. Relative frequency of nouns in three grammatical cases in our sample of spoken (SPO) and written (WRI) corpora.

CASE	CORPORA				
	SPOKEN		WRITTEN		SPO + WRI
	SG	PL	SG	PL	
NOMINATIVE	630 (73%)	138 (78%)	658 (58%)	119 (60%)	1545 (65%)
GENITIVE	98 (11%)	7 (4%)	283 (25%)	32 (16%)	420 (18%)
PARTITIVE	134 (16%)	32 (18%)	191 (17%)	48 (24%)	405 (17%)
TOTAL by number and register	862 (100%)	177 (100%)	1132 (100%)	199 (100%)	
TOTAL by register	1039		1331		2370

While partitive and genitive together account for less than half of all the nominals in our sample, we nevertheless have a total of 825 non-nominative NPs, a large enough sample of each case to examine differences in factors potentially affecting case-marking. Across all the data, the sample includes very similar proportions of genitive (420) and partitive (405) NPs. Yet we see a difference emerge when we examine the distribution of noun case by register, as shown in Figure 2, which

gives an overview of case use by corpora. The proportion of nominative NPs overall in the spoken corpus is much larger than in the written corpus (74% compared to 58%), and the distribution of non-nominative NPs also differs. Whereas the written data includes fewer partitive than genitive nouns (239 PAR to 315 GEN, or 18% vs 24%), the proportion is reversed in the spoken corpus (166 PAR and 105 GEN, or 16% to 10%); this is a significant difference in proportional distribution of the cases *by register* ($\chi^2(1) = 23.17$, $p < 0.001$), as well as a significant difference *between partitive and genitive case within each register* (WRI: $\chi^2(1) = 10.43$, $p = 0.001$, SPO: $\chi^2(1) = 13.7$, $p < 0.001$).

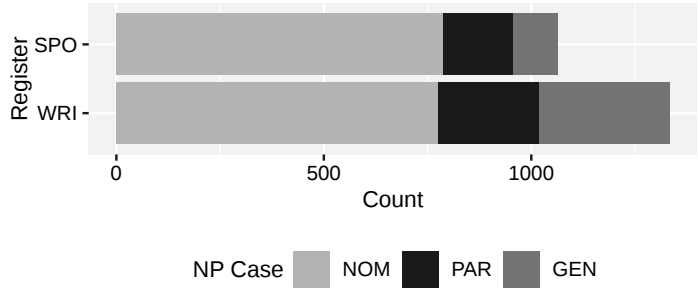

Figure 2. Overall distribution of noun case by register.

We explore these register differences, as well as what other factors are associated with the distribution of each case, in section 5.

5. Results

5.1. All nominals in the three grammatical cases

In this section, we present an overview of the distribution of nominative, partitive and genitive nouns according to the criteria included in our study.

Table 5. [8] Grammatical role distribution by case and register (spoken, written). Grammatical roles are subject, object, comp (predicate complement), possess (possessor), obj-pp (object of pre- or postposition), obj-inf (infinitival complement), n-comp (noun complement), other (various less frequent categories).

GRAM. ROLE	CASE					
	SPOKEN			WRITTEN		
	NOM	PAR	GEN	NOM	PAR	GEN
SUBJECT	**656** (85%)	26 (16%)	0	**638** (82%)	24 (10%)	1 (< 1%)
OBJECT	50 (7%)	**115** (69%)	29 (28%)	25 (3%)	**138** (58%)	47 (15%)
COMP	45 (6%)	0	0	49 (6%)	3 (1%)	0
POSSESS	0	0	**39** (37%)	0	0	**140** (44%)
OBJ-PP	0	0	28 (27%)	0	12 (5%)	87 (28%)
OBJ-INF	1 (< 1%)	12 (7%)	0	0	26 (11%)	3 (< 1%)
N-COMP	0	12 (7%)	2 (2%)	0	15 (6%)	2 (< 1%)
OTHER	16 (2%)	1 (< 1%)	7 (7%)	65 (8%)	21 (9%)	35 (11%)
TOTAL	768 (100%)	166 (100%)	105 (100%)	777 (100%)	239 (100%)	315 (100%)

Table 5 shows the distribution of nouns in nominative, partitive and genitive case, according to their grammatical roles in each corpus, looking at their frequency of occurrence without considering any other semantic or morphosyntactic information. Nominative is strongly associated with the role of subject, as expected, in both the written and spoken datasets. However, nominative shows more diverse functions in the written than the spoken corpus, as shown by greater proportions of the 'other' category, which groups thirteen less frequent grammatical roles (those with fewer than 25 tokens across written and spoken data). The 'other' category makes up 8% of nominatives in the written corpus, as compared to 2% in the spoken corpus. For instance, a total of only 9 nominative nouns are coded as adjuncts, of which 8 are in the written corpus. Subjects make up 82% of nominative NPs in the written corpus, similar to 85% in the spoken corpus. Partitive nominals are clearly associated with the object role across both corpora.

Polarity plays an interesting role here: of all the partitive nominals in our sample, nearly one quarter (23%) occur in negative clauses (SPO: 28%, WRI: 19%), compared with 13% negative clauses overall

8 Note that percentages are rounded. As such, they do not always add up to 100%.

(N = 306); for comparison, only 7% of genitive nominals occur in negative clauses. Looking more closely at only the negative clauses, 56% of the partitive nominals are objects; 27% are subjects. For genitive nominals in negative clauses, 87% are either complements of adpositions or possessors (compared to 69% in affirmative clauses).

Genitive, on the other hand, is distributed across three primary grammatical roles: possessor (SPO: 37%, WRI: 44%), postpositional complement (obj-pp, SPO: 27%, WRI: 28%) and object (SPO: 28%, WRI: 15%). All three cases show more diverse functions in the written corpus. We can also see that the much higher proportion of genitives in the written corpus does not derive from any one grammatical role being used more often than in the spoken data: genitive nominals do not occur in the possessor role significantly more frequently in the written than the spoken data ($\chi^2(1) = 1.43$, $p = 0.23$). General proportions of usage across the corpora are similar. Although the written corpus shows more diversity in terms of attested grammatical roles, these uses are infrequent.

The form of nominals also differs across corpora (see Table 6), and this is likely to interact with the mapping between morphological case and syntactic function. Overall, *pronouns* are used proportionally more in the spoken than written corpus ($\chi^2(1) = 1241.6$, $p < 0.001$). In the written corpus, we find a predominance of *lexical nouns* (68%). This is reversed in the spoken data, in which *pronouns* make up 57% of the nominals in the sample.

Table 6. The distribution of NP form across corpora by register.

NOMINALS	REGISTER	
	SPOKEN	WRITTEN
LEXICAL NOUNS	449 (43%)	901 (68%)
PRONOUNS	590 (57%)	430 (32%)
TOTAL	1039 (100%)	1331 (100%)

This difference in NP form is especially noticeable with nouns in nominative, as shown in Figure 3. In the spoken language, nominative *pronouns* are more frequent than nominative *lexical* nouns. Partitive and genitive case occur more frequently with *lexical* nouns, regardless of register (partitive: 62% lexical nouns in spoken, 77% in written; genitive: 68% lexical nouns in spoken, 75% in the written data).

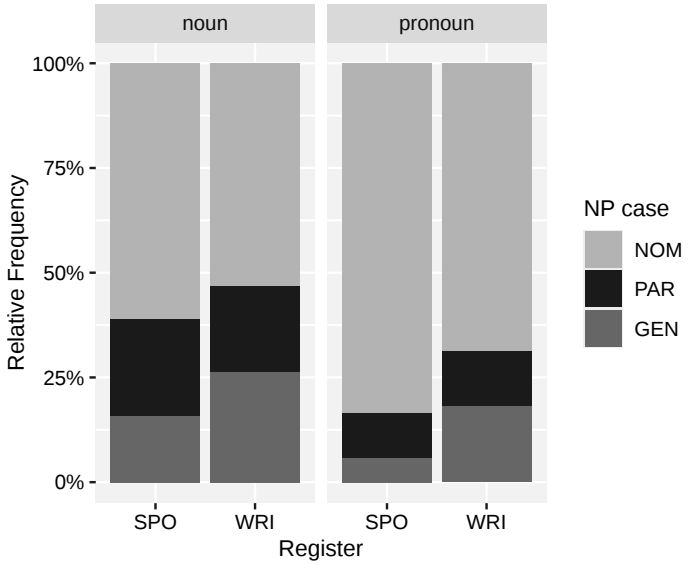

Figure 3. The distribution of NP case by form (noun, pronoun) and register (spoken, written).

It is worth highlighting that pronouns predominate *only* as nominative subjects: 734 nominative subject pronouns across both corpora, compared to 560 nominative subject nouns (see Table III at https://doi.org/10.23673/re-429 – Raw_data.pdf). This is in keeping with what has been noted cross-linguistically regarding Preferred Argument Structure (Du Bois et al. 2003): grammatical function, form and information structure redundantly support one another. Pronouns typically express given information and are used as subject-topics, hence occurring more frequently in nominative case. In other syntactic contexts where nominative occurs, lexical nouns predominate.

When looking at the possible interactions between case-marking, nominal form and number, we see that singular nominals far outnumber plurals (Table 7). Yet in terms of noun form, pronouns and lexical nouns have almost identical proportions in singular and plural: pronouns comprise 43% of all singular nominals, across corpora, and 42% of plural nominals.

Table 7. The distribution of NP form across corpora by number.

NOMINALS	NUMBER	
	SINGULAR	PLURAL
LEXICAL NOUNS	1133 (57%)	217 (58%)
PRONOUNS	861 (43%)	159 (42%)
TOTAL	1994 (100%)	376 (100%)

Figure 4 plots the relation between NP case, form and number. The most frequent nominals in nominative are plural pronouns. The proportion of singular and plural lexical nouns in nominative is nearly identical. A striking difference is observed between the proportion of plurals among partitive lexical nouns (24%) and pronouns (10%; $\chi^2(1) = 9.09$, $p < 0.003$). In genitive, the most predominant form is singular among lexical nouns. We do not find many differences here between registers; the only notable difference is that genitive plurals are even less frequent in spoken (7% of all genitive nominals) than written corpus data (10%).

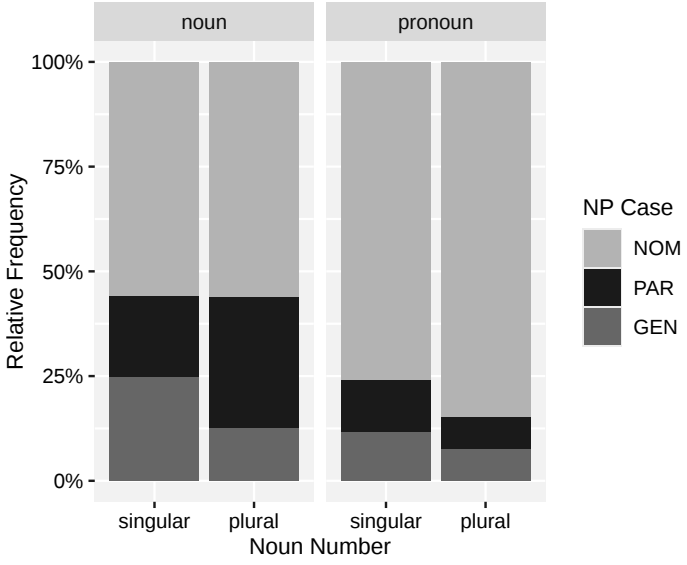

Figure 4. The distribution of NP case by form (noun, pronoun) and number (singular, plural).

We also coded for countability and found mass nouns, at very low rates, among nominative subjects, partitive objects and, more rarely, partitive subjects. The proportion of mass nouns in genitive case overall is negligible. The distribution of mass nouns does not show differences between registers, even with partitive case.

Turning to animacy, we find an effect of animacy on the distribution of lexical nouns and pronouns across corpora: a large proportion of lexical nouns refer to inanimates (see Table 8), and large proportion of pronouns refer to animates.

Table 8. The distribution of NP form across corpora by animacy.

ANIMACY	NOUN FORM	
	LEXICAL NOUNS	PRONOUNS
ANIMATE	403 (30%)	727 (71%)
INANIMATE	947 (70%)	293 (29%)
TOTAL	1350 (100%)	1020 (100%)

Figure 5 shows that animate pronouns are most likely to be in nominative and least likely to occur in partitive case. Overall, both lexical nouns and pronouns predominantly occur in nominative when referring to animates. Lexical nouns in all three cases are more likely to refer to inanimates. For pronouns, this is reversed in nominative and genitive (where more pronouns have animate referents), but not in partitive case: even pronouns are more likely to refer to inanimates when they are marked by partitive. This includes personal, demonstrative and interrogative pronouns without differentiation, and is true despite the partitive bias among first and second person pronouns (noted in section 3.4).

Further, animacy and case interact in determining the grammatical role of a nominal, shown in Table 9. Nominative prefers animate referents when functioning as the grammatical subject, but not in other grammatical functions. Partitive prefers nominals with inanimate referents overall, but genitive reveals clear differences by grammatical function: we see more *animate* genitive nouns as possessors, while *inanimates* tend toward the functions of object and adpositional complement (OBJ-PP).

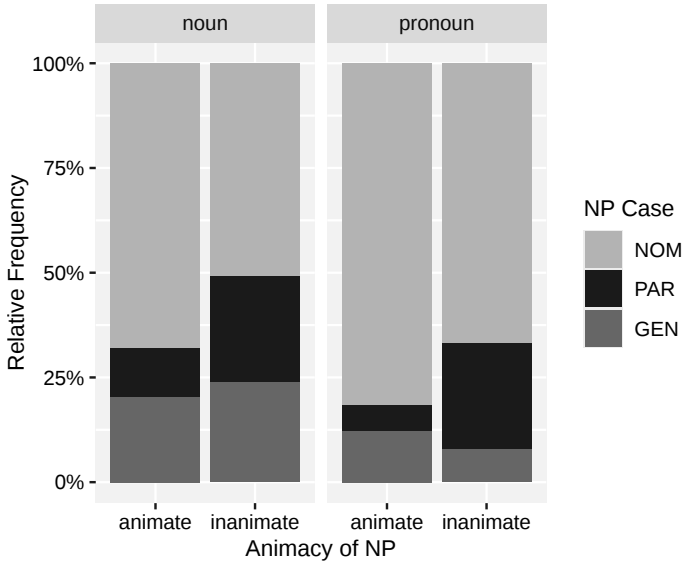

Figure 5. The distribution of NP case by NP form (noun, pronoun) and animacy (animate, inanimate).

Table 9. Animacy of referent, by case and grammatical role.[9]

GRAM. ROLE	CASE					
	ANIMATE			INANIMATE		
	NOM	PAR	GEN	NOM	PAR	GEN
SUBJECT	**803** (93%)	6 (7%)	1 (< 1%)	**491** (72%)	44 (14%)	0
OBJECT	5 (< 1%)	**60** (65%)	10 (6%)	70 (10%)	**193** (62%)	66 (27%)
COMP	24 (3%)	1 (1%)	0	70 (10%)	2 (< 1%)	0
POSSESS	0	0	**106** (62%)	0	0	73 (29%)
OBJ-PP	0	3 (3%)	39 (23%)	0	9 (3%)	**76** (31%)
OBJ-INF	0	15 (16%)	0	1 (< 1%)	23 (7%)	3 (1%)
N-COMP	0	2 (2%)	1 (< 1%)	0	25 (8%)	3 (1%)
OTHER	35 (4%)	5 (5%)	14 (8%)	46 (7%)	17 (5%)	28 (11%)
TOTAL	867 (100%)	92 (100%)	171 (100%)	678 (100%)	313 (100%)	249 (100%)

9 Note that percentages are rounded. As such, they do not always add up to 100%.

We also checked for differences in animacy between registers. With all cases, the spoken data is more biased toward inanimate referents than the written data: For nominals in nominative case, 59% are animate in the written data, 53% are animate in the spoken data. For partitive, animate referents account for 28% in the written data but only 14% in the spoken data, and for genitive case, animate referents account for 43% of the tokens in written data and 34% in spoken data.

5.2 Argument case-marking

One primary function of case-marking is to distinguish between arguments, and this might lead to differences in usage in transitive and intransitive contexts. Moreover, in Estonian, the grammatical options available in one- and two-argument clauses differ. In impersonal or imperative subjectless clauses, no genitive arguments are used (as shown in section 2.2, ex-s 8–10). Partitive subjects appear only in intransitive contexts (at least in our dataset[10]), see ex (5). In this section, we turn our attention to the questions of how the distribution of case-marked nouns differs in one- and two-argument clauses. Hence, we include here only noun phrases which function as arguments of the verb. One- and two-argument clauses have been extracted automatically, meaning that any clause with one overt argument is marked as a one-argument context, even if the predicate is transitive and one argument is implicit. One-argument clauses include intransitive clauses, transitive clauses with only one overt subject or object, as well as subjectless clauses such as impersonals and imperatives. For an overall picture of the differences, Table 10 shows the numbers of tokens in each context. Note that the one-argument clauses are divided into those with only a subject (S-only) and those with only an object argument (O-only); these are independent of each other.

10 However, in the closely related language Finnish, partitive subjects have been noted to occur in transitive clauses as well as intransitives. According to Huumo (2018), these partitive subjects of transitive verbs prefer animate, human referents, and tend to be in plural or quantificational phrases, along the lines of *Palju inimesi ootas vihmas bussi* ‘There were many people waiting for the bus in the rain’.)

Table 10. Core argument structure: Summary of the distribution of grammatical cases by argument number in the clause, argument role and register.

CORPUS	CASE	ARGUMENT STRUCTURE			
		S-only	O-only	S&O	
				SUBJ	OBJ
SPO	NOM	**530**	29	**115**	20
	PAR	26	**36**	0	**75**
	GEN	0	8	0	20
WRI	NOM	**435**	14	**159**	5
	PAR	21	**33**	0	**110**
	GEN	0	8	0	36

Note that even though the one-argument data may include transitive clauses, the distribution of cases by roles is markedly different from two-argument clauses (S&O). No partitive subjects are attested in transitive contexts. Genitive marks very few objects overall, but much less in one-argument than two-argument clauses. Genitive is ungrammatical in impersonals and imperatives, but recall that the data also includes clauses with implicit subjects. Proportionally more nominative objects are found in the spoken data (27% of all objects) than in the written data (8%), an imbalance which holds in both one- and two-argument clauses.

From Table 10, we find that nominative dominates in one-argument clauses far more than in two-argument clauses, but it is strongly associated with subjects. Among two-argument clauses, nominative subjects outnumber objects of any case, but partitive objects are a close second, and are much more frequent in two-argument than one-argument clauses. The preference for nominative subjects over objects is significantly greater overall in the written data (χ^2 (1) = 9.67, p = 0.002).

Hence, when we limit our data to only core verb arguments, partitive is second in frequency by far. In transitive contexts, partitive only marks objects, never subjects. In one-argument clauses, partitive signals objecthood more than subjecthood, yet the proportion of partitive subjects is substantial (39% of partitive nominals in the written corpus, and 45% in the spoken corpus). As for genitive arguments, these can only be objects, and they appear much less frequently than genitive in other functions: only 5% of verb arguments are in genitive case in the written corpus, and 3% in the spoken data.

Figure 6 shows the distribution of NP case by the animacy of subject and object arguments and linear position relative to the verb (SV – preverbal subject, VS – postverbal subject, OV – object preceding a finite verb, VO – object following a finite verb); note that the dataset includes arguments of negated verbs. Preverbal, nominative subjects are more likely to be animate than postverbal nominative arguments. Animate subjects are hardly in partitive in the preverbal position. In partitive, subjects are most likely to be inanimate and postverbal. As for the object argument, we see that *animate* objects are predominantly partitive in the preverbal position, and occasionally in nominative (one occurrence), whereas genitive stands out as a case which does not mark *animate* nouns in the preverbal position as the object, preferring inanimate nouns and in the postverbal position.

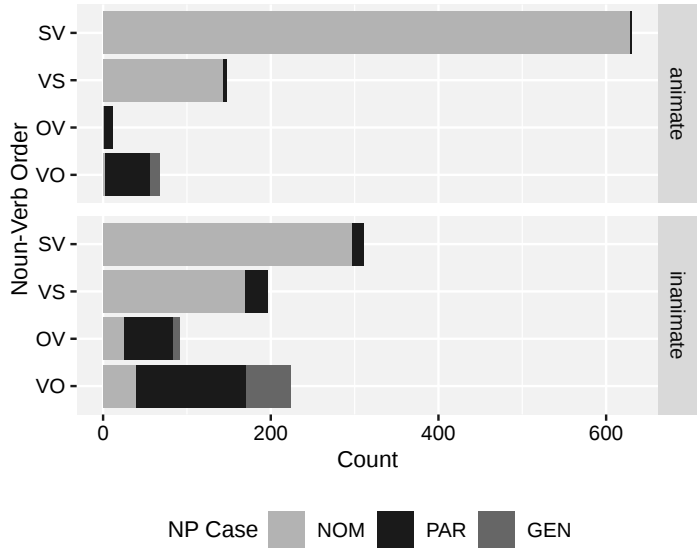

Figure 6. The distribution of case by animacy and word order: subject and object arguments only.

We can see from Figure 7 that nominative subjects dominate in single-argument (S-only) clauses and are more likely to be animate than inanimate; this difference in animacy is particularly visible in two-argument clauses (S&O), where animate nominative subjects are clearly preferred. Partitive shows a clear tendency to occur with inanimate

nominals, both in the object and subject function. In one-argument clauses with only an object argument, nominative is more likely to mark the object than genitive, although partitive is the most frequent object case. Genitive tends to occur with inanimate objects, but is more likely in two-argument than one-argument clauses.

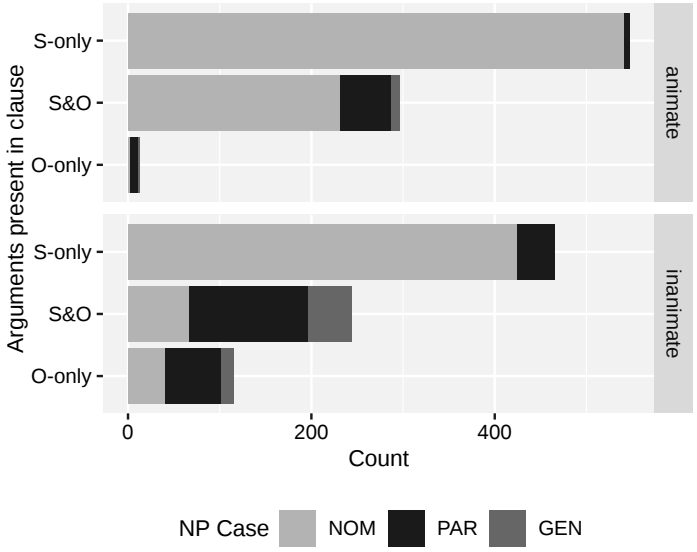

Figure 7. The distribution of case by animacy and number of overt arguments: subject and object arguments only.

Figure 8 focusses on the case distribution according to grammatical role and number of overt arguments in the clause (subject, object or both), as well as linear position of the argument with respect to the verb. It shows that two-argument clauses include a significant proportion of objects in nominative which precede the finite verb. More importantly, we can see that, despite flexibility in word order, the two-argument clauses exhibit proportionally less inversion than clauses with only the subject. VS order is used more in single-argument clauses than two-argument transitive clauses (see Vihman & Walkden 2021, who note that subject inversion is infrequent, and that postverbal S appears mostly in V-first clauses rather than XVS).

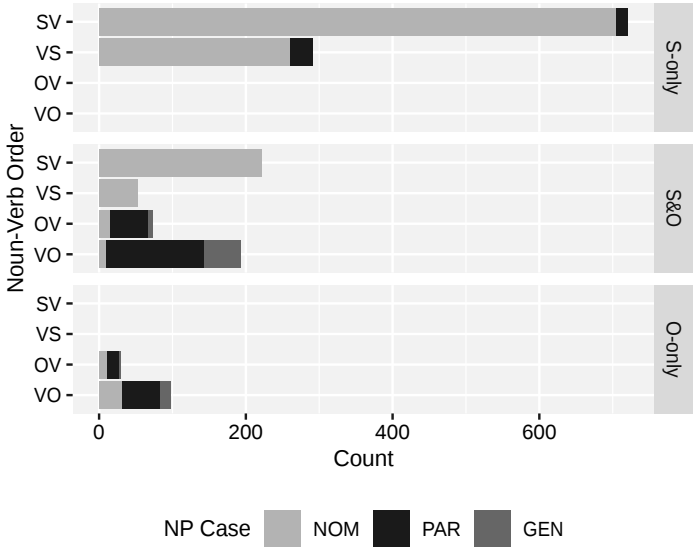

Figure 8. The distribution of case by number of overt arguments and word order: subject and object arguments only.

5.3. Correspondence Analysis

In sections 5.1–5.2, we probed an array of features affecting the distribution of cases. We looked at some relations between register, case, form, number (including mass and count nouns), animacy, polarity and grammatical role. In this section, we ask: What is the relationship between *all the features* in our dataset? That is, how relevant is each feature for the specific case label and what are the key relationships between each case and the features? Multiple Correspondence Analysis (MCA) is a statistical technique enabling us to address these questions by establishing relationships between a large number of categorical factors and plotting a two-dimensional graphic of the co-occurrence frequencies found in the corpus data converted to (Euclidean) distances. These distances can be viewed on the plot by judging the relative distances between the numbers –2 to 2 on the y-axis and –1 to 2 on the x-axis.

Figure 9 shows an MCA analysis of all the features under study, treating them as regular variables, including NP case. The scatterplot is a visualisation of the frequency-based relations found in our corpus

data, produced in a two-dimensional plot, where Dim1 stands for the first, horizontal dimension (x-axis), and Dim2 for the second, vertical dimension (y-axis). The horizontal and vertical dimensions show the proportion of variance explained (given in brackets in the labels: 14.3% and 12%); such proportions are expected with a large number of factors. As noted, the analysis treats all the variables as equally important, and displays relationships between variables only in instances of a very strong relation. The contribution of each variable is indicated by the darkness of its label.

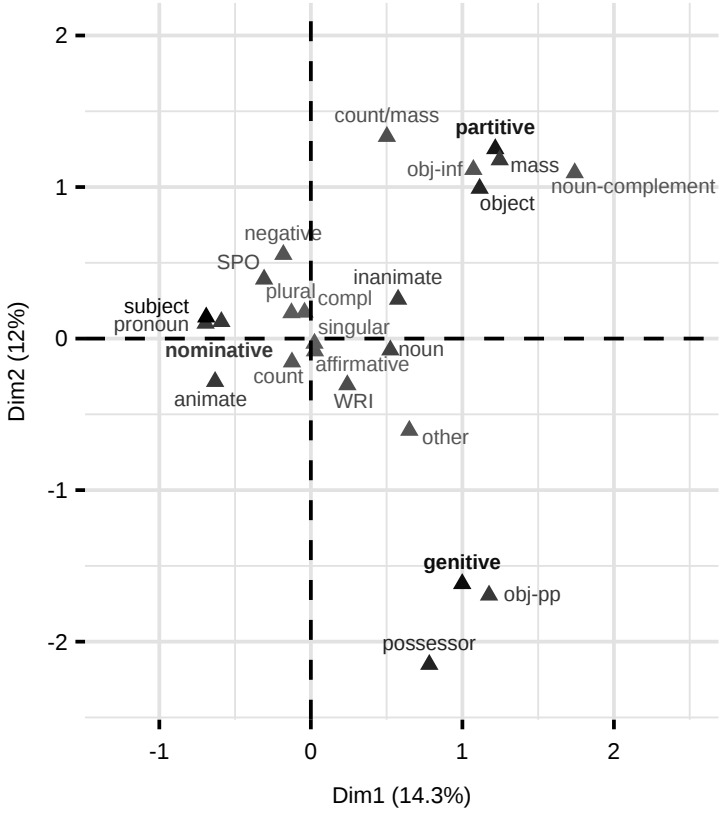

Figure 9. Multiple Correspondence Analysis of all the datapoints.

In Figure 9, partitive and genitive are shown to be clearly distinguished from each other, as they are placed far apart along the vertical dimension (Dim2). They are also distinguished from nominative, which is near the origin, but clearly at a distance from both

partitive and genitive along both the vertical and the horizontal dimensions (Dim1). Nominative is close to the origin along with many other features, indicating a lack of differentiation from these.[11] An examination of the contribution of each variable (indicated by the darkness of its label), and quality of representation of the factor levels allows us to confirm that there is a strong tendency for grammatical subjects to occur in nominative (and not in genitive or partitive).

The vertical dimension (Dim2) reflects the fact that the grammatical roles of possessor and object of adposition (obj-pp) are strongly connected to genitive case (i.e., the angles formed by connecting the point 'genitive' to the origin and back to the points 'possessor' and 'obj-pp' are very small). We can also see in the MCA plot that genitive has a stronger affinity for the written corpus than partitive does.

The grammatical roles of object, noun complement (e.g. *tükk leiba* 'piece [of] bread-PAR.SG') and object of infinitive are shown to be strongly associated with the partitive case. In the same cluster, a feature related to noun countability turns up, marked by the label 'count/mass'. This label (marked in the data file as NA) stands for nouns whose countability cannot be determined without further context (e.g., *see* 'this' as in ***See*** *muserdas mind lõplikult* 'This devasted me completely'; or *midagi* 'something' in *Lõpuks ostad sa ikka* ***midagi*** 'In the end you still end up buying something'.) Both these indeterminate nouns and mass nouns display a preference for partitive case.

Overall, in this analysis, grammatical roles and their strong preferences with regard to specific cases clearly dominate, backgrounding the other variables.

6. Discussion and conclusions

6.1. Nominative and its distribution

Nominative is, unsurprisingly, by far the most frequent form used in both written and spoken corpus data, and this is even more pronounced in the spoken data (as also noted by Hennoste 2004, on singular nouns in the colloquial spoken register). This aligns with the fact that the

11 In technical terms, this means that fairly average results are measured on the variables, hence centred.

spoken language has more single-argument clauses, where nominative dominates.

Observing how nominative is distributed with regard to grammatical functions in language use, we see that it is mapped to the most frequent grammatical role, the subject; hence its status as the most frequent case form. Yet, it is worth highlighting that a morphological case marker is only as informative as the syntactic context in which it occurs, at least in Estonian, where each of the grammatical cases may signal more than one grammatical role. While nominative is most frequent as a marker of subject case, it is not an entirely reliable signal of subjecthood. Considering the frequency with which objects occur in nominative case in one-argument clauses, probabilistic models and online language comprehension cannot take nominative to be a clear indication of the subject role. The MCA model, too, shows a lack of differentiation of nominative from the origin, or the other features in the dataset, rather than any strong associations with any of these. The data overall seem to support an approach which takes nominative to be an unmarked form – a lack of case-marking rather than a case form (see Kaiser, Miljan & Vihman 2020). Subject-verb agreement, on the other hand, only occurs with nominative subject arguments, hence it always picks out subjects. Partitive subjects do not trigger verb agreement; they are marked as atypical subjects. Hence, we may conclude that verb agreement uniquely signals the unmarked subject role, while partitive case-marking signals either objects or marked subjects. Nominative on its own does not carry information about grammatical relations.

Moreover, the marking of personal pronouns corroborates the status of nominative as ‘no case’: first and second person pronouns never occur as nominative (i.e. unmarked) objects, while lexical items do occur in nominative in these contexts (see section 3.4). These personal pronouns always take case-marking (either genitive or partitive) in the object function, as they are highly likely candidates for subject rather than object. Indeed, as our data show, speakers can also draw probabilistic evidence from the form of the NP (pronouns occurring more frequently as subjects than nouns), the animacy of the referent (nominals with animate referents being more frequent subjects), and word order (preverbal nominative nominals are highly likely subject candidates).

6.2. Genitive and its distribution

Genitive is the second most frequent case in our data, but it accounts for less than a fifth (18%) of all the nominals included in the study, and varies by register, occurring in the written data much more frequently than the spoken data (as also found by Hennoste 2004). This is most likely due to lengthier, syntactically more complex clauses, as indicated by the distribution of genitive nominals by grammatical role.

Although the genitive is described by some as a realisation of accusative case, it occurs far more frequently in contexts which show nominal rather than verbal dependencies. In our study, genitive is strongly associated with the possessor role (considerably more frequent in the written than spoken data), as well as with adpositional complements, and is less likely to be related to the object function. When genitive does mark the object, it is most frequently in two-argument clauses, as nominative is more likely than genitive in single-argument, null subject clauses (genitive marks 12% of object NPs in clauses with one argument, compared to 33% nominative). This merits further investigation, as it may be that genitive only occurs as the object case in a particular, limited set of constructions. So far, studies on object case alternation and the conditions for the genitive vs partitive marking on the object have focussed on lexical and semantic features (see sections 2.2 and 3.2), which may have created expectations for higher frequency of genitive objects than are actually used in the modern language.

6.3. Partitive and its distribution

Partitive occurs with nearly identical frequency to genitive in the corpus overall, constituting 17% of all nominals. Unlike the genitive, its distribution is not sensitive to register, occurring with similar frequency in the written and spoken data. As with genitive, partitive prefers a specific environment: it is most likely to occur as the object marker in two-argument transitive clauses. We found that the object role is twice as frequent in two-argument clauses (with an overtly expressed subject) than in subjectless, one-argument clauses.

As noted, partitive predominantly marks the object role, which is the second most frequent grammatical role in our data. Yet we found a difference between object marking in one- and two-argument clauses. While in two-argument clauses, the partitive is much more likely to

mark the object role than either genitive or nominative, in one-argument clauses the difference between the object alternatives (here, primarily partitive and nominative) is much smaller; in the spoken language, they are nearly equivalent. Thus, when a direct object is not marked by partitive case, it is *equally* likely to be marked by either nominative or genitive. Genitive is more likely than nominative to mark the object in two-argument clauses, whereas in one-argument clauses the nominative marks the object twice as often as genitive.

In our data, partitive never occurs on the subject in transitive clauses, but in single-argument clauses, 41% of partitive nominals are subjects, 59% are objects. The subject role marked by partitive in our data overwhelmingly comprises nominals with inanimate referents.

6.4. Each grammatical case has its own preferred factors

We observed a sensitivity to animacy in case-marking in our corpus data. This is not entirely expected, as animacy of the referent is not usually discussed as playing a role in case-marking in Estonian (though it was found to be relevant by Miljan, Kaiser & Vihman 2017). In our data, the partitive case showed a tendency to occur with inanimate nominals. For partitive-marked nominals, not only lexical nouns, but even pronouns are predominantly inanimate. The subject role marked by partitive pervasively comprises inanimates, overriding the strong tendency for subjects to be animate.

Nominative and genitive show a finer-grained animacy preference than partitive. Lexical nouns in any case are more likely to be inanimate, while pronouns in nominative and genitive tend to be animate. Animacy, in turn, interacts with word order: when an animate nominal precedes the verb, it is most likely the subject if it is nominative, and possessor if it is genitive.

As for word order, we asked whether clauses with OS order in Estonian would be more likely to have animate rather than inanimate objects – that is, whether these clauses follow the animate-first principle (Bader & Häussler 2010). We found that our data do not bear out this hypothesis. Clauses with OS order do not exhibit a reversal of the general tendency to assign animate referents to subjects and inanimate referents to object status (see Dahl 2008: 142); instead, the OS sample displayed more extreme canonicity, with greater proportions of animate

S and inanimate O. Thus, here we confirm the well-known complementary tendencies of S and O: animate referents are greatly preferred as subjects, especially in two-argument clauses, and inanimate referents are preferred as objects, even when in non-canonical, inverted order.

Ogren (2015b: 200–201) highlights the role of polarity in word order in Estonian, specifying that canonical VO order dominates in affirmative clauses, but in negative clauses VO and OV occur in similar proportions, with OV usually including a partitive object. Objects in partitive are more frequent preverbally than nominative or genitive objects. We found that genitive object arguments are nearly non-existent preverbally: genitive nominals may function as possessors in preverbal position, but are extremely rare as preverbal objects. Nominative objects, however, occur in one-argument clauses in equal proportions in VO and OV order. This contrasts with Ogren's observation that word order boils down to polarity. Perhaps this is due to the written register analysed by Ogren, as well as the specifics of the infinitival constructions he focussed on. In general, our study shows that more factors are at play here than polarity: the single-argument clauses and the spoken register make nominative objects more likely, although even in the preverbal position, partitive is preferred.

6.5. Conclusion

We set out to investigate what type of information is associated with the morphological cases nominative, genitive and partitive in Estonian. We can conclude that each of the grammatical cases is prototypically, but not reliably, associated with a particular grammatical role. We have shown that a cluster of features is available to indicate the function of a case-marked noun, in the absence of entirely reliable cues from morphological case. Thus, the interpretation of a nominal in nominative, genitive or partitive case is likely to draw on all the available semantic and syntactic information in online processing. Of the semantic information, animacy was found to have a greater impact than previously highlighted. Of the syntactic factors, the role of the grammatical context should not be underestimated: the number of syntactic arguments expressed in the clause and word order help constrain the ambiguous syntactic information provided by a case marker to pick out a specific grammatical function. In other words, a morphological case marker is only as informative as the syntactic context in which it occurs.

Acknowledgments

We gratefully acknowledge funding from the University of Tartu's base funding grants in humanities, awarded to each author, which supported the data coding and analysis for this study. Initial manual coding of the data was performed by Carl Eric Simmul and Merilyn Muru, for which we express heartfelt thanks. We would also like to acknowledge the Spoken Estonian research group for access to the spoken corpus and Maarja-Liisa Pilvik and Piia Taremaa for consultation. All remaining mistakes are, needless to say, entirely our own.

References

Bader, Markus & Jana Häussler. 2010. Word order in German: A corpus study. *Lingua* 120(3). 717−762. https://doi.org/10.1016/j.lingua.2009.05.007.

Biber, Douglas. 2012. Register as a predictor of linguistic variation. *Corpus Linguistics and Linguistic Theory* 8(1). 9–37. https://doi.org/10.1515/cllt-2012-0002.

Biber, Douglas & Susan Conrad. 2019. *Register, genre, and style.* Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108686136.

Blake, Barry J. 2001. *Case*, 2nd edn. Cambridge: Cambridge University Press.

Blevins, James P. 2008. Declension classes in Estonian. *Linguistica Uralica* 44(4). 241–267. https://doi.org/10.3176/lu.2008.4.01.

Comrie, Bernard. 1989. *Language universals and linguistic typology: Syntax and morphology*. Chicago: University of Chicago Press.

Dahl, Östen. 2008. Animacy and egophoricity: Grammar, ontology and phylogeny. *Lingua* 118(2). 141–150. https://doi.org/10.1016/j.lingua.2007.02.008.

Divjak, Dagmar. 2019. *Frequency in language: Memory, attention and learning*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781316084410.

Du Bois, John W., William J. Ashby & Lorraine E. Kumpf. 2003. *Preferred Argument Structure*. Amsterdam: John Benjamins. https://doi.org/10.1075/sidag.14.

Erelt, Mati, Helle Metslang & Helen Plado. 2017. Alus. In Mati Erelt & Helle Metslang (eds.), *Eesti keele süntaks*, 240–257. Tartu: University of Tartu Press.

Foley, William A. & Robert D. Van Valin. 1985. Information packaging in the clause. In Timothy Shopen (ed.), *Language Typology and Syntactic Description,* vol. 1: Clause structure, 282–354. Cambridge: Cambridge University Press.

Granlund, Sonja, Joanna Kołak, Virve-Anneli Vihman, Felix Engelmann, Elena Lieven, Julian Pine, Anna Theakston & Ben Ambridge. 2019. Language-general and language-specific phenomena in the acquisition of inflectional noun morphology: A cross-linguistic elicited-production study of Polish, Finnish and Estonian. *Journal of Memory and Language* 107. 169–194. https://doi.org/10.1016/j.jml.2019.04.004.

Helasvuo, Marja-Liisa. 2001. *Syntax in the Making: The Emergence of Syntactic Units in Finnish Conversation.* Studies in Discourse and Grammar 9. Amsterdam/Philadelphia: Benjamins. https://doi.org/10.1075/sidag.9.

Hennoste, Tiit. 2004. Mõnede käänete sagedus ja lauseliikmelisus suulises kõnes. In Liina Lindström (ed.), *Lauseliikmeist eesti keeles*. Tartu Ülikooli eesti keele õppetooli preprindid 1, 16–25. Tartu: Tartu Ülikooli Kirjastus.

Hiietam, Katrin. 2003. Definiteness and grammatical relations in Estonian. University of Manchester, Ph.D. thesis.

Hiietam, Katrin. 2004. Accusative – why not? *Proceedings of the 11th Postgraduate Conference in Linguistics*. Manchester: University of Manchester.

Huumo, Tuomas. 1995. On the position of subject in Finnish and in Estonian. In R. Kataja & K. Suikkari (eds.), *XXI kielitieteen päivät Oulussa 6.–7.5.1994*. Acta Universitatis Ouluensis Series B Humaniora 19, 79–87. Oulu.

Huumo, Tuomas. 2018. The partitive A: On uses of the Finnish partitive subject in transitive clauses. In Ilja A. Seržant & Alena Witzlack-Makarevich (eds.), *Diachrony of differential argument marking*, 383–411. Berlin: Language Science Press. https://doi.org/10.5281/zenodo.1228271.

Kaalep, Heiki-Jaan. 2010. Mitmuse osastav eesti keele käändesüsteemis. *Keel ja Kirjandus* 53(2). 94–111.

Kaalep, Heiki-Jaan. 2012. Käänamissüsteemi seaduspärasused. *Keel ja Kirjandus* 55(6). 418−449. https://doi.org/10.54013/kk655a2.

Kaiser, Elsi, Merilin Miljan & Virve-Anneli Vihman. 2020. Estonian speakers' interpretations of morphological case: Implications for Case/Agree. In András Bárány & Laura Kalin (eds.), *Case, Agreement, and their Interactions: New perspectives on differential argument marking* (Linguistische Arbeiten), 301−347. Berlin: Mouton de Gruyter. https://doi.org/10.1515/9783110666137-010.

Lago, Sol, Diego E. Shalom, Mariano Sigman, Ellen F. Lau & Colin Phillips. 2015. Agreement attraction in Spanish comprehension. *Journal of Memory and Language* 82. 133–149. https://doi.org/10.1016/j.jml.2015.02.002.

Lindström, Liina. 2002. Veel kord subjekti ja predikaadi vastastikusest asendist laiendi järel. *Emakeele Seltsi aastaraamat* 47. 87–106.

Lindström, Liina. 2017a. Lause infostruktuur ja sõnajärg. In Mati Erelt & Helle Metslang (eds.), *Eesti keele süntaks*, 537–565. Tartu: University of Tartu Press.

Lindström, Liina. 2017b. Partitive subjects in Estonian dialects. *Eesti ja soome-ugri keeleteaduse ajakiri. Journal of Estonian and Finno-Ugric Linguistics*, 8(2). 191–231. https://doi.org/10.12697/jeful.2017.8.2.07.

Meir, Irit, Mark Aronoff, Carl Börstell, So-One Hwang, Deniz Ilkbasaran, Itamar Kastner, Ryan Lepic, Adi Lifshitz Ben-Basat, Carol Padden & Wendy Sandler. 2017. The effect of being human and the basis of grammatical word order: Insights from novel communication systems and young sign languages. *Cognition* 158. 189–207. https://doi.org/10.1016/j.cognition.2016.10.011.

Metslang, Helena. 2013. *Grammatical relations in Estonian: Subject, object and beyond*. University of Tartu. Ph.D. thesis.

Metslang, Helena. 2014. Partitive noun phrases in the Estonian core argument system. In Silvia Luraghi & Tuomas Huumo (eds.), *Partitive Cases and Related Categories,* 177–256. Berlin, Boston: De Gruyter Mouton. https://doi.org/10.1515/9783110346060.177.

Metslang, Helle. 2017. Sihitis. In Mati Erelt & Helle Metslang (eds.), *Eesti keele süntaks*, 258–277. Tartu: University of Tartu Press.

Miljan, Merilin, Elsi Kaiser & Virve-Anneli Vihman. 2017. Interplay between case, animacy and number: Interpretations of grammatical role in Estonian. *Finno-Ugric Languages and Linguistics* 6(1). 55–77. https://full.btk.ppke.hu/index.php/FULL/article/view/54/63.

Nelson, Diane & Virve Vihman. 2018. Shifting perspective: noun classes, voice and animacy type shifts. *Theoretical Linguistics* 44(1–2). 57–69. https://doi.org/10.1515/tl-2018-0005.

Norris, Mark. 2018. Non-autonomous accusative case in Estonian. *Finno-Ugric Languages and Linguistics* 7(2). 7–38. https://doi.org/10.15763/11244/320359.

Ogren, David. 2015a. Differential Object Marking in Estonian: proto-types, variation, and construction-specificity. *SKY Journal of Linguistics 28*. 277–312.

Ogren, David. 2015b. Sõnajärg, infostruktuur ja objekti kääne eesti keeles. *Eesti ja soome-ugri keeleteaduse ajakiri*. *Journal of Estonian and Finno-Ugric Linguistics* 6(3). 197–213. https://doi.org/10.12697/jeful.2015.6.3.08.

Ogren, David. 2018. Object case variation in Estonian *da*-infinitive constructions. *Dissertationes philologiae Estonicae universitatis Tartuensis* 41. Tartu: University of Tartu Press.

Primus, Beatrice. 2013. Animacy, generalized semantic roles, and differential object marking. In Monica Lamers & Peter de Swart (eds.), *Case, word order and prominence: Interacting cues in language production and comprehension*, 65–90. Dordrecht, Heidelberg, London, New York: Springer. https://doi.org/10.1007/978-94-007-1463-2_4.

Rajandi, Henno & Helle Metslang. 1979. *Määramata ja määratud objekt*. Tallinn: Valgus.

Sahkai, Heete. 2011. Eesti genitiivse agendifraasi süntaks. *Keel ja Kirjandus* 54(1). 12–30.

Schlesewsky, Matthias & Ina Bornkessel. 2004. On incremental interpretation: degrees of meaning accessed during sentence comprehension. *Lingua* 114(9–10). 1213–1234. https://doi.org/10.1016/j.lingua.2003.07.006.

Siewierska, Anna. 1999. From anaphoric pronoun to grammatical agreement marker: why objects don't make it. *Folia Linguistica* 33(2). 225–252. https://doi.org/10.1515/flin.1999.33.1-2.225.

Szmrecsanyi, Benedikt & Lars Hinrichs. 2008. Probabilistic determinants of genitive variation in spoken and written English. In Terttu Nevalainen (ed.) *The Dynamics of Linguistic Variation: Corpus evidence on English past and present,* vol. 2, 291–309. Amsterdam: John Benjamins. https://doi.org/10.1075/silv.2.22szm.

Tamm, Anne. 2007. Perfectivity, telicity, and Estonian verbs. *Nordic Journal of Linguistics* 30(2). 229–255. https://doi.org/10.1017/S0332586507001746.

Tamm, Anne. 2015. Negation in Estonian. In Matti Miestamo, Anne Tamm & Beáta Wagner-Nagy (eds.), *Negation in Uralic languages* (Typological studies in a language 108), 399–432. Amsterdam; Philadelphia: John Benjamins Publishing Company. https://doi.org/10.1075/tsl.108.15tam.

Tamm, Anne & Natalia Vaiss. 2019. Setting the boundaries: Partitive verbs in Estonian verb classifications. *Eesti Rakenduslingvistika Ühingu aastaraamat. Estonian Papers in Applied Linguistics* 15. 159–181. https://doi.org/10.5128/ERYa15.09.

Tauli, Valter. 1968. Totaalobjekt eesti kirjakeeles. *Suomalais-ugrilaisen Seuran Toimituksia*. 216–224.

Torn-Leesik, Reeli. 2009. The voice system of Estonian. *Language Typology and Universals* 62(1–2). 72–90. https://doi.org/10.1524/stuf.2009.0005.

Vaiss, Natalia. 2004. Eesti keele aspekti väljendusvõimalusi vene keele taustal. Unpublished Master's Thesis. Tallinn University.

Verkuyl, Henk J. 1993. *A theory of aspectuality: The interaction between temporal and atemporal structure.* Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511597848.

Vihman, Virve & George Walkden. 2021. Verb-second in spoken and written Estonian. *Glossa: A Journal of General Linguistics* 6(1): 15. 1–23. https://doi.org/10.5334/gjgl.1404.

Viht, Annika & Külli Habicht. 2019. *Eesti keele sõnamuutmine*. Tartu University Press.

Wang, Luming, Matthias Schlesewsky, Markus Philipp & Ina Bornkessel-Schlesewsky. 2013. The role of animacy in online argument interpretation in Mandarin Chinese. In Monica Lamers & Peter de Swart (eds.), *Case, word order and prominence: Interacting cues in language production and comprehension*, 91–119. Dordrecht, Heidelberg, London, New York: Springer. https://doi.org/10.1007/978-94-007-1463-2_5.

Kokkuvõte. Merilin Miljan, Virve Vihman: Eesti keele grammatilised käänded kirjalikus ja suulises korpuses: sagedus, jaotumus ning süntaktilised rollid. Artiklis esitatakse tulemused korpusuuringust, mille eesmärgiks oli välja selgitada eesti keele grammatiliste käänete (nominatiiv, partitiiv, genitiiv) jaotumus ning tegurid, mis mõjutavad nende käänetega markeeritud nimisõnade süntaktiliste rollide tõlgendamist. Erinevalt varasematest töödest, mis keskenduvad grammatilistele suhetele ja (seejärel) nende käändetähistusele, lähtub selle uurimuse fookus eelkõige morfoloogiast, st käändest endast. N-ö käände perspektiivist vaatleme iga grammatilise käände esinemise sagedust kirjalikus ja suulises korpuses: milliseid süntaktilisi funktsioone see markeerib, markeeritava nimisõna omadusi (elusus, arv, loendatavus), süntaktilist konteksti (sõnajärg, transitiivsus) ning registri erinevusi. Leiame, et kuigi üldpildina eristub iga grammatilise käände puhul sagedaim süntaktiline põhiroll, toob detailsem analüüs välja nimisõna omaduste ja süntaktilise konteksti olulisuse nende rollide jaotumuses.

Võtmesõnad: korpusanalüüs, nominatiiv, partitiiv, genitiiv, süntaktilised suhted, eesti keel, kirjalik keel, suuline keel