FINNISH JOURNAL OF LINGUISTICS

VOLUME 37:2024

EDITED BY
Mariann Bernhardt
Emmi Lahti
Lauri Marjamäki
Iira Rautiainen
Olli O. Silvennoinen

Finnish Journal of Linguistics is published by the Linguistic Association of Finland (one issue per year).

Notes for Contributors

Policy: *Finnish Journal of Linguistics* welcomes unpublished original works from authors of all nationalities and theoretical persuasions. Every manuscript is reviewed by at least two anonymous referees. In addition to full-length articles, the journal also accepts short (3–9 pages) 'squibs' as well as book reviews.

Language of Publication: Contributions should be written in English, French, German, Finnish, or Swedish. If the article is not written in the native language of the author, the language should be checked by a professional in that language.

Style Sheet: *Finnish Journal of Linguistics* follows the Generic Style Rules for Linguistics, with complementary house rules, which can be found at https://journal.fi/finjol/about/submissions.

Abstracts: Abstracts of the published papers are included in *Linguistics Abstracts* and *Cambridge Scientific Abstracts*. The published papers are included in *EBSCO Communication & Mass Media Complete*. *Finnish Journal of Linguistics* is also indexed in the *MLA Bibliography*.

Publisher:
The Linguistic Association of Finland
Suomalais-ugrilainen kielentutkimus
Hämeenkatu 1
FI-20014 Turun yliopisto
Finland
http://www.linguistics.fi, http://journal.fi/finjol

The Linguistic Association of Finland was founded in 1977 to promote linguistic research in Finland by offering a forum for the discussion and dissemination of research in linguistics, both in Finland and abroad. Membership is open to anyone interested in linguistics. The membership fee in 2023 was EUR 30 (EUR 15 for students and unemployed members).

Finnish Journal of Linguistics

36

Suomen kielitieteellisen yhdistyksen aikakauskirja
Tidskrift för den Språkvetenskapliga föreningen i Finland
Journal of the Linguistic Association of Finland

2024

ISSN: 2814-4376

Contents

Book reviews:

Reviewers of *Finnish Journal of Linguistics* 37 (2024)

Márton A. Baló (Hungarian Research Centre for Linguistics)
Richard Compton (Université du Québec à Montréal)
Viktor Elšík (Charles University)
Ylva Falk (Stockholm University)
Francesca Di Garbo (Aix-Marseille Université)
Ann-Kristin Helland Gujord (University of Bergen)
Lars Hellan (Norwegian University of Science and Technology)
Johanna Isosävi (University of Helsinki)
Anni Jääskeläinen (University of Helsinki)
Terje Lohndal (Norwegian University of Science and Technology)
Olga Lovick (University of Saskatchewan)
Ora Matushansky (Centre national de la recherche scientifique)
Andrew McKenzie (University of Kansas)
Renate Pajusalu (University of Tartu)
Malin Roitman (Stockholm University)
Kaius Sinnemäki (University of Helsinki)
Edward Vajda (Western Washington University)
Ulla Vanhatalo (University of Helsinki)

The editors are most grateful to all the scholars who have acted as reviewers for *Finnish Journal of Linguistics* in 2024, including those who wish to remain anonymous and are therefore not listed here.

Gender, meaning and arbitrariness: Evidence from Norwegian (and Swedish)

Hans-Olav Enger
University of Oslo

Abstract

Despite some claims in the literature, even NP-internal agreement can be meaningful in Norwegian and Swedish, and the arbitrariness of lexical gender (also known as 'formal gender' or 'syntactic gender') in these two languages has been overstated. This shows up also in homonyms of different genders, which pattern in a way linked to animacy. Furthermore, not all pronominal gender agreement in these languages is meaningful, either. Although there are differences between pronominal gender agreement and other kinds of agreement, this is a difference in degree, not in kind, so we should not draw a sharp distinction between pronominal gender agreement and other kinds of gender agreement. The paper also contributes to the long-standing discussion on the redundancy and usefulness of gender: gender is not as outlandish and different from other grammatical categories (such as tense and number) as it may seem, since no grammatical category correlates directly with conceptual distinctions.

Keywords: gender, arbitrariness, agreement, animacy, homonyms, Norwegian, Swedish

1 Introduction

This paper will present arguments in favour of the view that even NP/DP-internal agreement can be meaningful. More specifically, lexical gender in Norwegian (and Swedish) is not as meaningless as it is sometimes considered to be – not even on the indefinite article (determiner). The paper also shows, in less detail, that pronominal agreement is not always as meaningful as is usually thought. Finally, the paper shows that an emphasis on the arbitrariness of gender can be misplaced and heuristically unhelpful and that gender is not so dramatically different from such categories as tense and number as it has seemed, since neither category mirrors conceptual categories directly.

Agreement in general and gender in particular are often seen as meaningless. This view usually entails setting pronouns aside, because at least some of them clearly have meaning. On this view, agreement is often seen as mere copying. While this view has come in for criticism (see e.g., Corbett 2006: 114ff for a summary), it remains influential, and even those who reject it will often talk of agreement in terms of 'controllers' and 'targets'; this 'implies a fairly mechanical, monodirectional syntactic process of feature replication' (Haig & Forker 2018: 715). The view of agreement as mechanical copying is particularly tempting within the NP. Therefore, several scholars of different persuasions, discussing different languages (e.g., Papazian 1978; Palmer 1984; Lehmann 1988; Carstens 2000), draw a sharp line between agreement inside the NP and outside of it.

VERTAISARVIOITU
KOLLEGIALT GRANSKAD
PEER-REVIEWED
www.tsv.fi/tunnus

Finnish Journal of Linguistics 37 (2024), 9–32
DOI: 10.61197/fjl.146830

In this paper, we shall see that at *least some* agreement is meaningful – even inside the NP. Thus, drawing an absolute, sharp distinction between NP-internal and NP-external agreement can be positively unhelpful (cf. also Barlow 1999; Corbett 2006; Landau 2016). Despite appearances, semantic arguments do not give full support to a sharp distinction – at least not in Norwegian and Swedish. Thus, the arbitrariness of NP-internal agreement has sometimes been exaggerated in the Scandinavian literature.

The plan for this paper is as follows. After this introduction (Section 1), basic facts about gender systems in Norwegian and Swedish are laid out in Section 2. Section 3 outlines a widespread analysis of Scandinavian gender. This part draws on Teleman (1969; 1987), Josefsson (2013; 2014) and Åfarli, Nygård & Riksem (2022) in particular; they are clear and explicit representatives of a widespread view.

Section 4 presents arguments against the view outlined in Section 3, more specifically against the idea that lexical gender is arbitrary and meaningless in Scandinavian. The argument draws, amongst other things, on Bobrova's (2013) study of homonym pairs in Norwegian. Section 5 is devoted to the implications this has for our view of agreement and gender. I suggest that NP-internal agreement is sometimes meaningful and that there is some meaning in the gender of the determiner (5.1), that pronominal agreement is not invariably meaningful (5.2), and that gender is not quite as different from other categories as we may have thought (5.3). Section 6 summarises the paper.

2 Scandinavian gender systems

2.1 The many systems of Scandinavian

In this paper, we look at varieties of Mainland Scandinavian (North Germanic), primarily Norwegian, secondarily Swedish. These are languages of Norway, Sweden and Finland, but these languages – or dialects, depending on the definition – are not easily delimited from other varieties of Scandinavian (even if the written standards are).

There are many different dialects of Norwegian and two written standards, Bokmål and Nynorsk, and plenty of variation in both. In Bokmål, there are at least two somewhat different gender systems. In this paper, we focus on one only, which is close to the Swedish one, and thus relevant for sections 3–5. We shall also consider the 'classical' Nynorsk gender system, which is closer to the older stages with a traditional Germanic three-gender system. The Nynorsk system is described in Table 1. The system described in Table 2 may be the most common in written Bokmål, and it is close to the dialect of younger speakers in Oslo, for example.

Genders are defined as classes of nouns, reflected in the behaviour of associated words (Corbett 1991). In Norwegian, the associated words that are particularly relevant are adjectives, determiners, and pronouns, and so these are shown in the following Tables 1 and 2. In the varieties described in the following tables (as in most other Scandinavian varieties), the gender distinction is only relevant in the singular. Nouns inflect for definiteness, so that there is an indefinite and a definite form of each. The Nynorsk system in Table 1 shows a three-way gender distinction; masculine, feminine and neuter. In the Bokmål system in Table 2, masculine and feminine have merged, as it were, into a common gender. Compare Tables 1 (Nynorsk) and 2 (Bokmål)[1]. Tables 1 and 2 show some possible noun phrases with relevant agreement 'targets', and then the pronoun that would normally be used to refer to these phrases.

[1] Tables 1 and 2 illustrate attributive adjectives only. Problems with predicative adjectives are only briefly mentioned in the text below.

Table 1. Associated words showing gender in one variety of Norwegian Nynorsk

	determiner	adjective	noun	pronoun, typically
Masculine, inanimate, indefinite	*ein* a.M 'a fine car'	*fin* fine.INDF.MF	*bil* car	*han* 'he'
Masculine, inanimate, definite	*denne* this.MF 'this fine car'	*fine* fine.DEF	*bilen* car.DEF.SG[M]	*han* 'he'
Masculine, animate, indefinite	*ein* a.M 'a fine man'	*fin* fine.INDF.MF	*mann* man.INDF.SG	*han* 'he'
Masculine, animate, definite	*denne* this.MF 'this fine man'	*fine* fine.DEF	*mannen* man.DEF.SG[M]	*han* 'he'
Feminine, inanimate, indefinite	*ei* a.F 'a fine file'	*fin* fine.INDF.MF	*fil* file.INDF.SG	*ho* 'she'
Feminine, inanimate, definite	*denne* this.MF 'this fine file'	*fine* fine.DEF	*fila* file.DEF.SG[F]	*ho* 'she'
Feminine, animate, indefinite	*ei* a.F 'a fine woman'	*fin* fine.INDF.MF	*dame* woman.INDF.SG	*ho* 'she'
Feminine, animate, definite	*denne* this.MF 'this fine woman'	*fine* fine.DEF	*dama* woman.DEF.SG[F]	*ho* 'she'
Neuter, inanimate, indefinite	*eit* a.N 'a fine smile'	*fint* fine.INDF.MF	*smil* smile.INDF.SG	*det* 'it.N'
Neuter, inanimate, indefinite	*dette* this.N 'this fine smile'	*fine* fine.DEF	*smilet* smile.DEF.SG[N]	*det* 'it.N'
Neuter, animate, indefinite	*eit* a.N 'a fine child'	*fint* fine.INDF.N	*barn* child.INDF.SG	*det* 'it.N'
Neuter, animate, definite	*dette* this.N 'this fine child'	*fine* fine.DEF	*barnet* child.DEF.SG[N]	*det* 'it.N'

Table 2. Associated words showing gender in one variety of Norwegian Bokmål

	determiner	adjective	noun	pronoun, typically
Common, inanimate, indefinite	*en* a.C 'a fine car'	*fin* fine.INDF.C	*bil* car.INDF.SG	*den* 'it.C'
Common, inanimate, definite	*denne* this.C 'this fine car'	*fine* fine.DEF	*bilen* car.DEF.SG[C]	*den* 'it.C'
Common, animate, indefinite	*en* a.C 'a fine man'	*fin* fine.INDF.C	*mann* man.INDF.SG	*han* 'he'
Common, animate, definite	*denne* this.C 'this fine man'	*fine* fine.DEF	*mannen* man.DEF.SG[C]	*han* 'he'
Common, inanimate, indefinite	*en* a.C 'a fine file'	*fin* fine.INDF.C	*fil* file.INDF.SG	*den* 'it.C'
Common, inanimate, definite	*denne* this.C 'this fine file'	*fine* fine.DEF	*fila* file.DEF.SG[C]	*den* 'it.C'
Common, animate, indefinite	*en* a.C 'a fine woman'	*fin* fine.INDF.C	*dame* woman.INDF.SG	*hun* 'she'
Common, animate, definite	*denne* this.C 'this fine woman'	*fine* fine.DEF	*dama* woman.DEF.SG[C]	*hun* 'she'
Neuter, inanimate, indefinite	*et* a.N 'a fine smile'	*fint* fine.INDF.N	*smil* smile.INDF.SG	*det* 'it.N'
Neuter, inanimate, indefinite	*dette* this.N 'this fine smile'	*fine* fine.DEF	*smilet* smile.DEF.SG[N]	*det* 'it.N'
Neuter, animate, indefinite	*et* a.N 'a fine child'	*fint* fine.INDF.C	*barn* child.INDF.SG	*det* 'it.N'
Neuter, animate, definite	*dette* this.N 'this fine child'	*fine* fine.DEF	*barnet* child.DEF.SG[N]	*det* 'it.N'

For adjectives and determiners, Table 2 shows a two-gender contrast in Bokmål, between the 'common' and the neuter. The same holds for the Swedish and Danish written standards. The 'common' in Bokmål represents, historically, a continuation of the masculine and feminine genders – cf. the Nynorsk data in Table 1, which, historically speaking, represent an earlier stage. Nynorsk is here like Icelandic, Faroese and Elfdalian in having a traditional three-gender system. The incipient syncretism between the masculine and the feminine gender also shows in Nynorsk, though; cf. the glossing MF in Table 1.

These days, the feminine is receding (except in pronouns) in many varieties of Norwegian (see e.g., Lødrup 2011; Busterud et al. 2019; Haug in prep.) and even more so in Swedish (see e.g., Rabb 2007; Van Epps & Carling 2017). The status of the definite singular suffixes (*-en, -a, -et*) has been the subject of much debate. In some varieties, such as the Nynorsk standard in Table 1, these elements correlate well with the other gender markers, such as determiners and personal pronouns. However, in other varieties, such as the Bokmål standard in Table 2, the suffixes do not correlate so well with the determiner. On the traditional assumption that genders are classes of nouns reflected in the behaviour of associated words, the suffixes *-en, -a* and *-et* are not seen as gender exponents in this paper; after all, suffixes are not words. Still, the definite singular suffixes are indications of gender, and this paper follows Enger & Corbett (2012) in showing the indicated gender in square brackets. (For further discussion, see e.g., Dahl 2000b; Enger 2004a; Lødrup 2011; Enger & Corbett 2012; Svenonius 2017 and references there).

The evidence from personal pronouns correlates with that of the determiners in the Nynorsk data in Table 1. In the Bokmål data in Table 2, by contrast, personal pronouns tell a different story than determiners. The pronouns display a four-gender contrast, determiners a two-gender contrast. The neuter gender pronoun *det* and the common gender pronoun *den* (both meaning 'it') typically signal reference to a non-human. They also indicate that the noun in question typically selects *et* [INDF.NEUT] and *en* [INDF.C] (both meaning 'a') respectively. Historically speaking, *den* is an innovation (see e.g., Davidson 1990 and Enger 2004b).

The pronouns *han* 'he' and *hun* 'she' typically signal a human referent, and that the noun selects *en* [INDF.C]. The choice between *han* and *hun* typically depends on the sex of (human) referent. Higher animates such as dogs are in Bokmål often referred to with *den* 'it.C' (at least by those who do not own them). Essentially the same system is found in standard Danish and Swedish. To put it simply, there are four genders on pronouns, but not inside the NP. See Sections 3–5 below for complications, however.

Following Corbett (1991), Dahl (2000a), Wälchli & Di Garbo (2019), and others, we assume a 'semantic core' in the system. Importantly, this core does not relate exclusively to biological sex, but also to animacy (see Sections 3 and 4 for further discussion). This holds for gender systems in general, and for Scandinavian. Words for animates, human beings in particular, do not usually belong to the neuter gender in Norwegian (see Faarlund et al. 1997: 153f); *barn* 'child', included in the tables above, is a classic exception. In the Bokmål system in Table 2, they belong to the common gender, i.e., they take *en*, *fin* etc. (not *et*, *fint*, etc.). However, it does not follow that words for non-humans are restricted to the neuter. Compare the examples *bil* 'car', *fil* 'file' in Table 2 above.

2.2 Lexical vs. referential gender

As already noted, few nouns denoting humans are neuters, but some will take a neuter determiner. Examples include Norwegian Bokmål *postbud* 'mail carrier' (and the Nynorsk equivalent *postbod*), *vitne* 'witness'; cf. et *postbud*, *et vitne* with neuter determiner. However, if Norwegians refer to a mail carrier or a witness they can see, they will usually avoid the neuter pronoun *det*, which may seem offensive, as if one is degrading the referent. In other words, these nouns are hybrids, in Corbett's (2006) sense; they are not associated with an entirely consistent agreement pattern.

The *postbud* and *vitne* examples show that the semantics behind the gender system is clear in the pronouns, but less clear when it comes to the determiners. This is no surprise, typologically, given Corbett's (2006: 206ff) Agreement Hierarchy. Figure 1 shows a simplified version of the hierarchy (taken from Enger 2013, based on Corbett 2006), containing three 'pegs' for three different kinds of agreement controllers:

← Attributive — Predicative — Personal pronoun →

Figure 1. Corbett's Agreement Hierarchy, simplified version

For any controller that permits alternative agreements, the likelihood of referential (semantic) agreement will rise monotonically as we move towards the right along the Agreement Hierarchy (Corbett 2006: 207). So, if referential agreement is possible on the predicative, it will be possible on the personal pronoun too, but not necessarily the other way around. Gender agreement is subject to the Agreement Hierarchy, whether it shows on attributives, predicatives or pronouns.

It is difficult (at first sight) to explain why the Norwegian Bokmål noun *postbud* 'mail carrier' should take the neuter determiner *et*, since it refers to an animate. Yet it is not difficult to explain why we would use the pronoun *hun* 'she' if referring to an obviously female mail carrier, or *han* 'he' if referring to an obviously male one.

On closer inspection, there is a non-semantic reason why *postbud*, in isolation, takes the neuter determiner. The word is a compound (*post*+*bud*), and compounds usually have the same gender as the last 'member' has in isolation. Thus, the reason why *postbud* is a neuter is probably that *bud* is neuter. The noun *bud* is polysemous. It can denote a person, a messenger, but a more central meaning is probably 'message; bid; commandment', and apparently, core meanings tend to 'win' in lexical gender assignment (see e.g., Enger 2010 for some discussion of this issue). The neuter on *postbud* is lexical gender, the lexically specified gender of the noun in isolation. It is to be contrasted with referential gender, the gender we may choose on the pronoun, for example. There is normally no choice of determiner, which reflects lexical gender. Lexical gender is also known as syntactic or formal gender; referential gender is also known as semantic gender.

For many nouns, especially the inanimate ones, there is no obvious semantic reason for their belonging to one lexical gender or the other – *bil* 'car', *fil* 'file' and *smil* 'smile' in Tables 1 and 2 are examples. (See also Urek et al. 2022.) Thus, gender assignment for such words must be accounted for by morphological or phonological generalisations (or, alternatively, it may be arbitrary). Morphologically complex words, such as *postbud*, make up a large part of the nouns, and their gender is close to predictable, when it comes

to determiners and adjectives. Thus, while the Norwegian gender system is far from transparent, it is certainly not entirely arbitrary, either.

'Close to predictable' may be a strange wording, but then, predictability and arbitrariness are tricky notions – as is the entire issue of arbitrariness in gender. Trosterud (2001) argues explicitly that the Norwegian Nynorsk gender system (cf. Table 1) is not arbitrary, that lexical gender assignment is rule-governed. Drawing inspiration from Corbett (1991) and from Steinmetz (1986), Trosterud posits no fewer than 50 rules. Of those rules, 28 are semantic, 9 are phonological, 10 are morphological, and three have a more general character (one of them says the masculine is the default). By Trosterud's account, 94% of the nouns in the authoritative Nynorsk dictionary, *Nynorskordboka*, have rule-governed assignment. Trosterud's paper is an important achievement, and thus, after 2001, the arbitrariness thesis on Norwegian gender must be viewed with suspicion.

At the same time, the paper has some weak points which illustrate how difficult the notion of arbitrariness can be. While accepting Trosterud's main point, Enger (2002) argues that some rules lack independent evidence; they have no justification except to cover the data (and the alleged motivation behind some is implausible). Halse (2004) shows that some of Trosterud's rules (especially some of the semantic ones) have little practical value. Parkkonen (2011) shows that the assumption that the masculine is the default can be too simple. All these authors agree that the gender system of Norwegian is not arbitrary. Urek et al. (2022), who also support the claim that the Norwegian gender system is not entirely arbitrary, find psycholinguistic support for some of Trosterud's rules, but not for all, and, as so many before them, they point out that the rules are probabilistic.

While Kvinlaug (2011: 18) also agrees that gender assignment is basically not arbitrary, he points out that in the absence of a detailed presentation of the empirical data, it is hard to evaluate Trosterud's claim about exactly 94% of the nouns being rule-governed. Kvinlaug also criticises the claim about an exact percentage being rule-governed, since Trosterud explicitly refrains from taking a stand on the interaction between the rules. Kvinlaug's point is well taken: saying that sometimes formal rules win, sometimes semantic rules, making no claims about rule interaction, and then calling *both* possible outcomes 'rule-governed' does not seem entirely justified. Neither outcome is predicted, in the strict sense. On the other hand, as long as *some* plausible gender assignment rule bears on the outcome, it does not seem right to call the outcome arbitrary, either.

'Predictability vs. arbitrariness' is probably not the right way to frame the question of gender assignment (and may even be a case of what Langacker 1987 has dubbed 'the rule-list fallacy'). Van Epps et al. (2021: 266) say that in the literature,

> 'gender assignment is typically described in terms of rules, even though tendencies would be a more appropriate term. Gender assignment in general, as well as in our data, is highly variable and diverse, both synchronically as well as diachronically. Very few observed tendencies are exceptionless and would be seen as rules in the Neo-grammarian sense.'

Further arguments against the putative arbitrariness are easily found. As noted for *postbud* above, many compounds have the same gender as the 'last member' has in isolation. For example, *barneskole* 'children's school' has the same gender as *skole* 'school' (common), not *barn*, while *skolebarn* 'school child' will have the same gender

as *barn* ‘child’ (neuter), and not *skole*. For many derivations, the gender is not arbitrary: The noun *lydighet* ‘obedience’ has the same gender as other derivations in *-het*, such as *kjærlighet* ‘love’, *spydighet* ‘sarcasm’. Norwegian Bokmål *forelskelse* ‘infatuation’ has the same gender as other derivations in *-else*, compare *forsnakkelse* ‘slip of the tongue’, *besvergelse* ‘incantation, curse’, *anfektelse* ‘doubt, contestation’. The gender for most morphologically complex nouns is thus not arbitrary, and these obvious descriptive generalisations are uncontroversial. While gender on simplexes is less predictable, compounds and derivations together make up a considerable part of Norwegian nouns. Given that the gender of compounds and derivations is largely non-arbitrary, the arbitrariness claim does not seem appealing.

3 Splitting the system?

Many scholars have argued that the Scandinavian gender system should be ‘split in two’, as it were. (See e.g., Teleman 1969, 1987; Josefsson 2009, 2013, 2014.) On their view, there is not one Scandinavian gender system, but two. The idea is that the gender agreement found on pronouns is meaningful, while the gender agreement inside the NP is meaningless. The latter kind of agreement is labelled formal gender, the former is labelled semantic or referential gender. This clearly resembles the description in Section 2, but an important difference is that the dichotomy referential-lexical is now in practice equated with positions on the Agreement Hierarchy. ‘Formal gender’ is explained by Josefsson (2009: 40) in the following way, using the neuter as an example:

> ‘the neuter feature has a dual nature. First of all it is a morphosyntactic feature associated with nouns, in other words a “lexical gender feature”. As such the neuter gender does not carry any meaning; there is simply no element of meaning shared by all neuter nouns.’

Formal gender is seen as asemantic, and Josefsson (2013: 13) says that it ‘is simply not possible to predict the formal gender of a noun on the basis of its meaning’.

An even more radical view is advocated by Åfarli et al. (2022: 638): ‘gender assignment to non-sex nouns in Norwegian seems in general to be arbitrary and conventional.’ Åfarli et al. apparently take biological sex to be the sole, semantic basis of Norwegian gender. The implication would be that it is arbitrary that, for example, *spion* ‘spy’ and *skotte* ‘person from Scotland’ (nouns denoting human beings that are neutral as to the sex of the referent) belong to the common gender in Norwegian Bokmål, as do their cognates (*spion*, *skotte*) in Swedish. This claim is hard to defend, since very few animate nouns belong to the neuter (cf. Section 2 and references there, see also, e.g., Trosterud 2001; Faarlund et al. 1997). In his monumental catalogue of changes in lexical gender from Old Norse to Modern Norwegian, Beito (1954: 81f) records a group of animate nouns that have changed from the neuter to the masculine (which, to simplify, corresponds to the common gender, diachronically), but no group of animate nouns that have changed in the opposite direction. Dahl (2000b: 586) flatly rejects claims about the non-semantic character of the Swedish common-neuter distinction, calling them ‘in fact false’. Dahl points out that ‘animate nouns strongly tend to be uter [=common, HOE]’. Compare Section 2 above. In an extensive empirical study of six North Scandinavian varieties, also Van Epps et al. (2021) find solid support for the role of animacy.

Teleman and Josefsson both take a less radical stand than Åfarli et al., acknowledging that animate nouns usually belong to the common gender, and that, in Swedish, mass nouns often belong to the neuter. In support of the importance of arbitrariness, Josefsson (2013) nevertheless adduces the eight pairs of homonymous Swedish nouns with different genders in Table 3 below.

Table 3. Eight homonym pairs in Swedish

Common gender noun	Translation	Neuter gender noun	Translation
fax	'fax machine'	*fax*	'fax message'
bak	'butt'	*bak*	'baking'
visp	'whisk'	*visp*	'stuff being whipped'
lut	'lye'	*lut*	'angle of a hill'
kast	'caste'	*kast*	'throw'
pris	'snuff pouch'	*pris*	'prize'
e-mail	'e-mail (program)'	*e-mail*	'e-mail (message)'
as	'Norse god'	*as*	'carcass'

The argument is as follows: 'If we look carefully at the examples […] we find no particular meaning component that is shared by the common gender nouns in the left-hand column or the neuter nouns in the right-hand column' (Josefsson 2013: 13). Furthermore, '[i]n many cases, formal gender is simply arbitrary' (Josefsson 2013: 5), and gender is assumed to be practically useless, in a way that other categories usually are not: 'The whole purpose of formal gender morphophonology is […] to make visible […] other morphological categories' (Josefsson 2013: 11; see also Davidson 1990: 148). Finally, it is claimed that 'formal gender in Swedish does not have any meaning *per se*, but can be used to distinguish meanings conveyed by other morphological features' (Josefsson 2013: 57).

Josefsson's papers are admirably clear and explicit illustrations of a widespread view. We have already seen that Åfarli et al. (2022) push the idea of arbitrariness further than Josefsson does. Often, gender for nouns is described as not 'interpretable', in contrast to, say, number. We shall now consider counterarguments against the putative arbitrariness (some have already been presented in Section 2).

4 Lexical gender is non-arbitrary and meaningful, to some extent

4.1 Norwegian and Swedish homonyms

Homonyms of different genders are not restricted to Swedish; they are also found in Norwegian. Consider the examples in Table 4.

Table 4. Three homonym pairs in Norwegian

Common gender noun	Translation	Neuter gender noun	Translation
fyr	'bloke'	*fyr*	'lighthouse'
rev	'fox'	*rev*	'reef'
gap	'joker, fool'	*gap*	'gorge, mouth'

If gender ('formal' or 'lexical') really were 100% meaningless, there should be no discernible semantic pattern behind the distribution of gender on homonyms. Yet the three examples in Table 4 suggest a different story: The common is linked to words for animates, the neuter to words for inanimates. That is the 'semantic core' behind the Norwegian gender system, including pronouns. Compare Section 2.

However, the pattern in these three examples *could* be accidental. Josefsson's treatment of the Swedish gender system suggests as much. On closer inspection, however, animacy is relevant for the Swedish examples in Table 3 as well, once we accept a somewhat broader understanding of 'animacy' than the literal one. Such a broader understanding is worth clarifying, even though it is far from original. More than three decades ago, Comrie (1989: 197ff) made it clear that the broad label animacy covers more than the literal sense of the word. On this broader understanding, animacy relates to several factors, including individuation, agentivity, definiteness, abstractness and even empathy.

Figure 2 presents a simplified version of the animacy hierarchy (whether the hierarchy should be seen as a question of degrees or steps need not concern us now).

Words denoting	English examples
Humans	*woman, boy*
Animals	*cat, badger*
Non-animate tangible objects	*chair, bottle*
Abstract nouns and masses	*philosophising, ethanol*

Figure 2. A simple animacy hierarchy

Animacy and individuation are related (Comrie 1989: 199). Sasse (1993) even prefers to talk of a hierarchy of individuation rather than of animacy. Animate entities (boys and cats, for example) tend to occur in clearly delineated packages, as it were, unlike abstract entities (philosophising) and masses (ethanol). Many people do not think of yeast or moss as animate, presumably partly because they are not delineated, partly because animacy also relates to empathy. Versions of the animacy hierarchy have even been called an 'empathy hierarchy' and a 'me-first hierarchy', reflecting the anthropocentrism behind it.

Animacy is also related to agentivity. Animate entities such as women and badgers tend to be more agentive than non-animate entities such as bottles and philosophising. This does not rule out the possibility of non-animate nouns being agentive, but such cases

are less common. If a forensic pathologist says of a murder victim that the *bottle killed him*, we will interpret *the bottle* as an instrument applied by the killer. Instruments are closer than products to being agentive. Næss (2007) suggests the term 'force' for such cases. Animacy is also related to concreteness. Abstract entities cannot be animate, and animate entities must be concrete.

Armed with a broader understanding of animacy, we can return to Josefsson's pairs of nouns. The obvious case of animacy in the strictest sense being relevant is that of *as*, C 'Norse god' vs. *as*, N 'carcass'. Gods are near the top of the animacy hierarchy, carcasses near the bottom, so it is unsurprising that the former word belongs to the common gender, the latter to the neuter.

Two less obvious examples are *fax*, C 'fax machine' vs. *fax*, N 'fax message' and *e-mail*, C 'e-mail program' vs. *e-mail*, N 'e-mail message'. Clearly, there is no difference in animacy in the narrow sense of the word between machines/programmes on the one hand and messages on the other. None of them are alive. Yet a fax machine can 'produce', as it were a fax message, but not the other way around. In the same way, an e-mail program can 'produce' a message, but not the other way around. Machines and programmes are thus instruments and so closer to being agentive than messages are. The same difference holds for *visp*, C 'whisk' vs. *visp*, N 'stuff being whipped'; a whisk is instrumental in creating that which is whipped. These are cases of 'force', then.

As for *bak*, C 'butt' vs. *bak*, N '(act of) baking' there is, again, no difference in animacy in the narrow sense, but there is a difference in individuation and concreteness, in as much as the butt is more individuated and less abstract than is the act of baking. We have already noted that animates are more individuated and less abstract. The case for non-arbitrariness may seem weaker for the two instances of *pris*, but a snuff pouch (the masculine noun) is less abstract than a prize (the neuter noun). An intangible prize may be rare, but it is possible, an intangible snuff pouch belongs squarely to fiction. As for the two *kast*, a throw (the neuter noun) is more abstract than a caste (the common noun). As for the two *lut*, lye (the common noun denoting a strongly alkaline solution, typically sodium hydroxide or potassium hydroxide) is a mass, yet less abstract than an angle (the neuter noun).[2]

In short, out of Josefsson's original eight examples from Swedish, the majority, perhaps all, relate to animacy in the broadest sense. If 'formal gender' really were 100 % meaningless, there should be no discernible semantic pattern behind the distribution of genders on homonyms. What the discussion above of homonym pairs in Norwegian and Swedish suggests, however, is exactly the opposite. The eight Swedish noun pairs in Table 3 were meant to illustrate the arbitrariness doctrine. The fact that, on closer examination, they rather undermine it seems significant. Yet the possibility remains that this was due to chance. Fortunately, Bobrova (2013) has carried out a more comprehensive study of Norwegian homonym pairs.

4.2 Animacy and common gender

Bobrova (2013) went through every single homonym pair (such as common gender *fyr* 'bloke' vs. neuter gender *fyr* 'lighthouse' in Table 4) in *Bokmålsordboka*, an authoritative dictionary of Norwegian Bokmål. This amounted to roughly 430 pairs. Briefly, she found

[2] A less semantic argument against arbitrariness is that the noun meaning 'throw' is a conversion (compare the verb *kasta* 'throw'), and conversions are usually neuter in Swedish when they denote the action, common when they denote the instrument (Teleman et al. 1999: 60f, note 4). Thus, there is both semantic and formal motivation, so the example is not entirely arbitrary (and the difference between actions and instruments relates to animacy).

a statistically significant correlation between animacy/individuation on the one hand, and common gender on the other: 'The semantic analysis supports the suggested correlation between gender and the animacy/individuation hierarchy' (Bobrova 2013: iii, my translation).

Animacy and individuation are relevant to pronominal gender in Scandinavian, especially in a system like that of Bokmål (or Swedish), cf. Table 2. Part of the original motivation for 'splitting' the gender system in two, one part which is considered meaningful and one which is not, is that animacy and individuation were not seen as relevant for 'lexical' gender, but for 'referential' gender, cf. Section 3 above. Bobrova's findings cast serious doubt on this argument.

If there is one place where we would expect to find support for the arbitrariness hypothesis for Scandinavian gender, it is with homonym pairs like Norwegian *fyr*/*fyr* (or Swedish *as*/*as*). The reason is that gender on simplexes is less predictable, in Scandinavian as in German (on German, see Köpcke 1982.) There is usually morphological motivation for the gender of morphologically complex nouns (see Section 2 above). By contrast, there is usually not formal motivation behind the gender assignment for homonymous simplexes. Norwegian has a relatively transparent orthography (less than Finnish, but much more than English), so the homonym pairs in Table 4 sound the same, at least to many speakers. Phonological motivation thus seems irrelevant for the gender assignment of most Norwegian homonyms. Morphological motivation also seems less likely for Norwegian simplexes: in the usual case, gender predicts plural inflection, and not the other way round (cf. Enger 2004a).

To repeat, if there is one place we would expect arbitrariness in the Norwegian gender systems, this is it. On its own home turf, then, the arbitrariness hypothesis is beaten. The argument based on 'homonym pairs' simply does not hold.

Yet arguments like those offered by Josefsson and Teleman are widespread, and we have seen that Åfarli et al. (2022) go even further. One may wonder why. Perhaps this is a version of what Langacker (1987; 2008: 13) has called 'the exclusionary fallacy'. Given the clearly correct observation that not everything in Scandinavian gender is predictable or semantically motivated, scholars have gone to the other extreme, maybe because many linguists have frowned upon tendencies, at least until recently. Yet the positivist philosopher Hempel (1966) accepted tendencies even for the 'hard' sciences, so there is no obvious reason why linguists should have to reject them. We should look for 'regularities, not rules' (Dammel & Schallert 2019: 7).

Certainly, regularities do not cover everything in gender assignment (see also Fraurud 2000), but then, few observations about language are totally free from exceptions. Labelling a phenomenon arbitrary is simply unlikely to stimulate further research. The idea that gender assignment is (semantically) arbitrary should be examined just as critically as the idea that gender assignment is fully (semantically) regular. The fact that scholars do not understand the principles behind a certain phenomenon X is not sufficient argument for the conclusion that there are no principles behind X. The latter does not follow from the former, and the arbitrariness doctrine does not fare too well in this case. Bobrova found inspiration in Corbett's (1991) approach, which is close to the opposite.

Let us now return to Josefsson's (2013: 13) observation that it 'is simply not possible to predict the formal gender of a noun on the basis of its meaning'. If 'predict' means 'tell with full certainty', the remark is entirely correct. Yet in that case, we cannot, for example, 'predict' the argument structure of a word based on its meaning, either

(cf. Haugen 2014; Hudson 2010: 286f), and at least to my knowledge, nobody has concluded that argument structure is completely arbitrary.

However, if 'predict' means 'guess with a reasonable chance of being correct', the argument does not hold. There is ample psycholinguistic evidence indicating that speakers can learn tendencies (cf. Dąbrowska 2004), so linguists should not dismiss tendencies light-heartedly. Often, language learning/acquisition is about tendencies, not foolproof predictions. Van Epps et al. (2021) find many tendencies in gender assignment in North Scandinavian, and they find diachronic evidence in support.

5 Meaningful NP-internal agreement and meaningless pronominal agreement

5.1 The gender on the determiner can be meaningful, as can other NP-internal agreement

There is another side to the issue of 'meaning' for homonyms of different gender. Compare Examples (1)–(4) from Norwegian Bokmål:

(1)	*Jeg*	*ser*	*en*	*fyr.*
	I	see	a.C	bloke
	'I see a bloke.'			
(2)	*Jeg*	*ser*	*et*	*fyr.*
	I	see	a.N	lighthouse
	'I see a lighthouse.'			
(3)	*Båten*	*traff*	*en*	*rev.*
	boat.DEF	hit	a.C	fox
	'The boat hit a fox.'			
(4)	*Båten*	*traff*	*et*	*rev.*
	boat.DEF	hit	a.N	reef
	'The boat hit a reef.'			

An important clue to the difference in meaning between the bloke and the lighthouse, between the fox and the reef respectively, is in the different determiners – *en* vs. *et*. The gender value on the agreement target is meaningful, in these cases.

An obvious objection might be that the gender is really in the noun and that the determiners merely reflect the gender of the noun, so that the 'real' difference is not between *en* and *et*, but between *fyr* I and *fyr* II. While such an objection is understandable, it does entail a claim that what listeners can hear or readers can see is meaningless, while what is meaningful is what listeners cannot hear – or readers cannot see. This seems unappealing. My claim is not that the noun has no bearing on the meaning difference, nor is it that the context does not play a role. The point is only that listeners have no reason to dismiss an obvious and easily perceivable clue to the difference between *fyr* I and *fyr* II, the determiner. Syntagmatically, the determiner (*en/et*) is an index, in Peircean terms, and as such, it is a sign. A particular determiner does narrow down the set of possible nouns that may follow, and speakers notice (see e.g., Heim et al. 2005; Miceli et al. 2002).

The fact that the gender is in the noun does not exclude it also being in the determiner. Langacker (1991: 187) rejects the assumption that 'a marking cannot be meaningful if its occurrence is obligatory […] if an element is obligatory, there is certainly a sense in which its occurrence is uninformative, but that is very different from saying that it has no semantic content'. In other words, 'redundant' is not the same as 'meaningless' (see also 5.3).

Enger (2013) presents three other arguments for meaningful gender inside the Scandinavian NP; we shall look at two of them. In Norwegian Nynorsk, which has a three-gender system (see Table 1), the lexical gender of *dørvakt* 'bouncer' is, perhaps surprisingly, feminine. In colloquial style, one may use a pronoun (alternatively a determiner[3]) attributively before the noun, and the choice does not depend on the lexical gender of the noun, but on the sex of the bouncer in question. Compare the Nynorsk examples in (5)–(6):

(5)	*Har*	*du*	*sett*	*han*	*nye*	*dørvakta?*	*Han*	*er*	*stor.*
	Have	you	seen	he	new.DEF.SG	bouncer.DEF.SG[F]?	He	is	big
(6)	*Har*	*du*	*sett*	*ho*	*nye*	*dørvakta?*	*Ho*	*er*	*stor.*
	Have	you	seen	she	new.DEF.SG	bouncer.DEF.SG[F]?	She	is	big

The reason for choosing *han*/*ho* is the same, whether *han*/*ho* is inside or outside of the NP.

In at least one variety of Swedish, described in the Swedish Academy Grammar (Teleman et al. 1999: 227), such examples as (7)–(8) are acceptable:

(7)	*Alexander,*	*de-n*	*ny-e*	*chef-en,*	*han*	*är*	*trevlig.*
	Alexander	the-C	new-DEF.M	boss-DEF.SG[C]	he	is	nice.C.SG
	'Alexander, the new boss, he is nice.'						
(8)	*Alexandra,*	*de-n*	*ny-a*	*chef-en,*	*hon*	*är*	*trevlig.*
	Alexandra	the-C	new-DEF.F	boss-DEF.SG.[C]	she	is	nice.C.SG
	'Alexandra, the new boss, she is nice.'						

The choice of *han* in (7) depends on the same factor as the choice of *nye*, animacy and sex of the referent, so *nye* is an example of referential agreement inside the NP.[4]

The claim that even NP-internal agreement can be meaningful finds support outside of Scandinavian. Corbett (2006) has presented considerable evidence, and we shall just consider one more recent example here. Belyaev et al. (2015) show that agreement on the NP-internal adjective can be meaningful in Russian and Italian. Their Italian example is given in (9):

(9)	*le*	*bandiere*	*rossa*	*e*	*bianca*
	the.PL	flag.PL	red.SG	and	white.SG
	'the red flag and the white flag [2 flags total]'				

[3] It is a moot point whether determiners and pronouns constitute one class or two for Norwegian, and, for that matter, Scandinavian. Kristoffersen (2000) and Halmøy (2016) reject the two-class analysis for Norwegian. Hansen & Heltoft (2011: 183) reject the two-class analysis for Danish. (Examples such as *han nye dørvakta* in (5) are best analysed as one phrase in Norwegian, and not two; see Enger 2013: 285.)

[4] The fact that there is a particular masculine agreement inside the NP in (7) is clearly unusual in the Swedish agreement system overall. Nevertheless, the possibility is there.

The phrase refers to two flags in total, with the attributes 'red' and 'white' each holding of a different flag (i.e., not two Polish or Austrian flags). According to Belyaev et al. (2015: 29), there 'is no direct number agreement between each adjective and the noun'; note that the adjectives are in the singular, the noun in the plural. And yet the 'number marking on the adjectives makes a very clear semantic contribution to the interpretation of the phrase'.

To sum up, NP-internal agreement can be meaningful (as argued also by e.g., Corbett 2006 and Landau 2016), also in Scandinavian.

5.2 Pronominal agreement is not always meaningful

As noted above, part of the justification for 'splitting' the Scandinavian gender system is that lexical gender, found inside the NP/DP, is not semantic. We have seen that this does not hold. In a discussion of Italian gender and meaning, Percus (2011: 173) hits the nail on its head: 'what seems to be the right approach for some cases just doesn't seem to work for other cases'.

The other part of the justification for a split in Scandinavian is that pronominal gender is semantic. However, this does not always hold, either. As already mentioned, the Swedish gender system is in principle like that of Table 2, in that there are four genders in pronouns, but usually two elsewhere. Yet in their Swedish grammar, Holmes & Hinchliffe (2013: 4) note that '[n]ouns ending in *-a* [in the indefinite singular, HOE] which denote animals are often treated as feminine irrespective of their true gender [biological sex, HOE]: *råttan – hon* the rat – she, *åsnan – hon* the donkey – she'. (See also Teleman et al. 1999: 277.) So, if we wish to explain why *råttan* 'the rat' is referred to by the pronoun *hon* 'she', *musen* 'the mouse' by the pronoun *han* 'he', the account will involve the observation that *råtta* 'rat' ends in an unstressed *-a* in the indefinite singular and that *mus* does not. Ending in an unstressed -a can hardly be called a semantic property. Thus, it is not the case that pronominal agreement is always semantically motivated in Swedish.

In Norwegian, the noun *barn* 'child' takes a neuter determiner. This may relate to its semantics, but it is nevertheless a lexical fact; the near-synonym *unge* 'child' is a masculine in Nynorsk, a common gender noun in Bokmål. One can refer to *barnet* 'the child' with the neuter pronoun *det*, and Faarlund et al. (1997: 327-28) see this as a violation of the usual tendency for pronouns to agree according to natural sex. This indicates that pronominal agreement is not always semantically motivated in Norwegian Bokmål, either. Example (10) stems from a Bokmål novel from 2019 *(Festningsverket*, by David Lie):

(10)	*Guttebarnet,*	*det*	*slipper*	*å*	*drite*	*seg*	*ut.*
	boy-child.DEF.SG[N]	it.N	is.spared	to	shit	itself	out

'The boy-child is spared from making a fool of itself (lit. shitting itself out).'

Since to refer to *guttebarnet* with *han* 'he' would count as referential gender agreement here, the use of the neuter pronoun *det* here reflects lexical gender.

Compare also the following Bokmål example in (11), a philosophical joke on solipsism:

(11) *Vi er mange som tror vi er den eneste i verden.*
we are many who believe we are the.SG only in world.DEF.SG[C]
'We are many who think we are the only one in the world.'

Here, the second occurrence of the plural pronoun *vi* does not carry any 'real plural meaning'. It is a case where the use of the plural is 'meaningless' even on a pronoun.[5] In short, pronominal agreement is not always meaningful in Scandinavian; it can be fairly meaningless, in the same way as NP-internal agreement (if less often).

5.3 Comparing gender with other categories

In parts of the literature, gender comes out as strange, which is unfortunate. It is also seen as having no reasonable purpose, except to support other categories – that idea is also unappealing. Nevertheless, when gender comes out as so strange, this may be partly because gender sometimes has been treated in 'splendid isolation' (a tendency criticised also by Wälchli & Di Garbo 2019). Corbett (1991: 1) says that '[g]ender is the most puzzling of the grammatical categories'. This often-quoted remark may be correct, but it has sometimes been misread. Gender is not the *only* grammatical category that is puzzling: they all are, to some extent. Spencer (2002: 280) says that when it comes to being puzzling, gender 'competes with verbal aspect'. Kürschner & Nübling (2011: 357) say that 'declension seems no less puzzling'. Gender can at least in principle help in disambiguation (even if it rarely does so in practice, cf. Feist 2020). By contrast, declensions and conjugations are often held to be entirely pointless. (Carstairs-McCarthy 2010 argues for a different view, however.)

Let us compare gender values with values for an apparently 'reasonable' morphosyntactic category, verbal tense. Tense is usually held to be a 'meaningful' category, so some might object: 'There are puzzling things *within* tense, but it is hardly puzzling that languages regularly distinguish present from past. There are also puzzling things *within* gender. But worse than that, gender is puzzling *externally* – why have it at all? Particularly when plenty of languages manage without.'

Many languages manage without tense as well (or, for that matter, without most other grammatical categories). Dahl & Villupelai (2013) say that it 'is only from a Eurocentric point of view that the marking of the distinction between present and past appears to be a necessary part of grammar. Languages may or may not distinguish […] grammatically, and there is no clear majority for either alternative.' We should not mistake the familiar for the inevitable. There is an advantage in taking a 'Martian perspective', i.e., to try to see language from outside (Carstairs-McCarthy 2010). While the distinction between present and past is common in European languages, an inflectional distinction between present and future is not, for example. The difference between this distinction and that between present and past would presumably not be obvious to a Martian.

Gender may not have an obvious *raison d'être*, but grammar is generally anthropocentric. The semantic core of gender is invariably related to animacy and to biological

[5] In his grammar of Nynorsk, Venås (1990: 71) commented on what at the time was new feminism in society and language, where *-mann* 'man' in compounds was no longer seen as uncontroversially sex-neutral. Given a compound such as *fylkesmann* 'county governor', Venås comments that if the *fylkesmann* is a woman, it seems 'unreasonable' (*urimeleg*) to use *han* 'he'. However, Venås adds as an afterthought that in such cases, where 'natural [≈referential] gender' and 'grammatical [≈lexical] gender' compete, usage has varied. This means that not all pronominal agreement has been referential in Nynorsk, either.

sex (cf. Section 2), and these properties are clearly important to human beings – in much the same way as the distinction versus the present and the past. Animacy is pervasive in grammar, and easily overlooked.

The Norwegian tense system is broadly comparable to the English one. There are two tenses in the strictest sense, viz. the present and the past. Neither correlates perfectly with 'logical time', as the following examples illustrate. Sentences (12)–(15) exemplify this for the present tense (all examples in Bokmål):

(12) *Akkurat* *nå* *går* *jeg* *på* *trikken.*
right now go I on tram.DEF.SG[C]
'Right now I'm entering the tram.' 'Logical time': right now

(13) *I går* *var* *vi* *i* *skogen.* *Plutselig* *kommer* *det* *en* *ulv.*
yesterday were we in forest.DEF.SG.C suddenly comes it a.C wolf
'Yesterday, we were in the woods. Suddenly, a wolf emerges.' 'Logical time': past

(14) *Tenk* *deg* *at* *du* *dør.*
imagine yourself that you.SG die
'Imagine that you die.' 'Logical time': hypothetical

(15) *I morgen* *går* *jeg* *på* *kino.*
tomorrow go I on cinema.INDF.SG
'Tomorrow, I'll be going to the cinema.' 'Logical time': future

The present tense, then, covers all conceivable 'logical times'. So what about the past tense?

(16) *Dette* *var* *god* *kaffe!*
this.N was good.INDF.SG.C coffee
'This coffee tastes good!' (right now, while I am drinking it)
'Logical time': right now

(17) *I går* *dukket* *det* *opp* *et* *nytt* *problem.*
yesterday popped it.N up a.N new.N problem
'Yesterday, a new problem turned up.' 'Logical time': past

(18) *Om* *jeg* *var* *en* *rik* *mann,* […]
if I were a.C rich.INDF.SG.C man […]
'If I were a rich man, […]' 'Logical time': hypothetical

(19) *Da* *gikk* *vi,* *da!*
then went we, then!
'It's time we went!' (as I have been telling you several times before this)
'Logical time': immediate future

The past tense *also* appears to cover all conceivable logical times, even though, admittedly, its use for 'right now' and 'immediate future' is much more restricted than what holds for the present. Recall now the arbitrariness claim for gender from Section 3. The argument was that there is 'no particular meaning component that is shared by the common gender nouns […] or the neuter nouns […]'. In the same way, there is no particular meaning component shared by all and only the present tense forms or all and only the past tense forms.

We should treat gender more like tense in seeing both its semantic basis and the quirks. Even tense is not always fundamentally meaningful. Smith (2022: 63) makes the point well:

> 'some functional or grammatical properties (such as gender and case) have no necessary extralinguistic correlate, whereas others (such as number and tense) appear to be correlated with independent conceptual notions, such as quantity and time. […] However, the correlation of, say, number with quantity and tense and aspect with time is at best indirect.'

The past tense is semantically motivated, but it is not dictated by semantics. The same holds for the common gender in Swedish, for example, even if the common gender is less motivated by semantics than is the past. Also in Norwegian Bokmål, it is easy to see a semantic reason for most uses of, e.g., past tense, and not so easy for many uses of, e.g., the common gender. Yet a quantitative difference is not a qualitative one. My argument is trivial: grammatical meaning does not always make complete sense. Still, this trivial message seems necessary. The difference between gender and tense is a matter of degree: we are not dealing with a difference in kind.[6]

At this stage, another possible objection could be as follows: 'Gender is not a grammatical category, the way tense is. The gender value for a particular noun is lexically fixed; the tense value for a particular verb will vary. Meaning, according to a venerable structuralist doctrine, lies in oppositions. Since gender does not vary for a noun, it is meaningless.'

On closer inspection, this objection does not hold. While the gender on a particular noun in language L need not vary, it varies *between* nouns. Otherwise, we would not say that language L has a gender system at all. Here is an analogy from forestry: Say that some trees in a wood are marked for being felled, e.g., with a big X. Others are unmarked (the default). The marking on each tree does not vary, yet it seems clear that the big X has a meaning.

Furthermore, an anonymous reviewer points out that in many languages of the world, 'gender assignment is highly flexible and contextual'. It can contribute to reference construal in many African languages, for example (see Di Garbo & Agbetsoamedo 2018). Finally, the argument that gender does not vary excludes 'hybrid nouns' (Corbett 1991), and thus rests on controversial assumptions. The gender of many nouns does vary also in Scandinavian, if we consider more than the determiner.[7]

---

[6] For further discussion of this point, see e.g., Smith (2022), O'Neill (2013) and Herce (2020, 2023: 28 *et passim*).

[7] In Norwegian Bokmål of a different kind than the one in Table 2, one may even find, if very rarely, semantically variable gender on the determiner (Enger 2015). We may find *ei lærer* 'a.F teacher' when reference is made to a woman, and *en lærer* 'a.M teacher' otherwise.

A comparison with number may help. Number on nouns is not as puzzling as gender, but number can be lexically fixed on nouns, just like gender. In Norwegian, several nouns occur practically only in the singular, and we may leave it open whether a plural can be formed at all, in natural speech. Examples include *dansing* 'dancing', *mjøl* 'flour', *oppdragsforskning* 'commissioned research'. Conversely, another, smaller set of nouns occur practically only in the plural; these include e.g., *opptøyer* 'riots', *innvoller* 'guts'. Thus, neither set of nouns will usually inflect for number. Clearly, there is a semantic motivation behind the absence of inflection. As for *mjøl* 'flour', mass nouns often do not pluralise, and as for *dansing* 'dancing', abstract nouns often do not pluralise. As for *opptøyer* 'riots' and *innvoller* 'guts', the English translational equivalents are also plural nouns. In short, lack of number inflection seems to be motivated by semantics. Yet this is hardly semantically predictable. For example, English *money* is a singular-only noun, just like German *Geld*, which means roughly the same thing. For many Norwegians, the translational equivalent noun *penger* is a plural-only noun, like the near-synonyms *kontanter* and *grunker*. Lack of number inflection is thus motivated by semantics, but it is not predictable. Besides plural-only nouns, there are plural-dominant nouns, so there are various relations a noun can have with number (see Corbett 2019 for extensive typological discussion). If meaning really lies in oppositions only, one is forced to claim that number is meaningful only for those nouns for which it is inflectional, and not for the rest. We may leave it to adherents of the structuralist doctrine to work out whether this implies that there is no semantic motivation behind cases of nouns that do not inflect for number.

The upshot is that we need not insist on either predictability or arbitrariness (either rules or lists), and that gender can be meaningful. There is an interesting convergence here with work in different research traditions. In a study in formal semantics, Percus (2011: 168) argues that 'gender features on Italian nouns are associated with an interpretation, but there are ways in which the grammar manages to hide this fact'. Percus (2011: 167) defends 'a view on which interpretable features can sometimes go uninterpreted. Gender features are among the elements that can do so'. Furthermore, in the framework of a 'dual model of language', Nordström (2024: 78) argues that both 'semantic and grammatical [person-number-gender, HOE] agreement affixes represent semantically interpretable features'. As for the alleged 'uselessness' of gender (Section 3), languages can vary in many ways, and most grammatical features in any particular language L are probably dispensable. That does not support singling out gender. Rather, it indicates that 'redundant', from the linguist's birds-eye perspective, cannot be equated with 'meaningless', from the perspective of language L or its users (cf. Langacker 1991: 187 and 379). Here is an illustration: Any linguist writing a paper will probably state the same main idea at least three times during the paper. It follows that two out of these three occurrences are redundant, strictly speaking, but it does not render them meaningless.

Some redundancy is probably necessary and useful in any human language, cf. Langacker (2008: 188): 'Redundancy is not to be disparaged, for […] every language makes extensive use of it. By providing the listener with extra clues, it helps ensure that a partially degraded message can still be understood'. Linguists tend to think of gender as an unnecessary complication, because speakers have to think of which gender to use, but listeners may sometimes be thankful for the disambiguating effect of gender. This is not to say that disambiguation is a primary function of gender, but it may be a convenient by-product (Feist 2020).

While linguists traditionally have aimed for redundancy-free descriptions, 'it remains perfectly possible for speakers to have systematic redundant knowledge'

(Maiden 2016: 136). Some redundancy may even be useful to speakers, who are known to be less than perfect. Ackerman & Malouf (2015: 310) say that '[a]ll natural languages show a certain degree of what Baerman et al. […] call "gratuitous" morphological complexity and Wurzel […] describes as "ballast" in the linguistic system.' On such a basis, it is not clear that gender is so special. There has been a tendency in the literature to treat gender in splendid isolation and as redundant and strange, yet gender is hardly more redundant or bizarre than, say, inflection classes or the German *Fugenelement* found in some compounds such as *Schafskopf* 'head of mutton', to mention but two examples.

Also Dolberg (2019: 25) presents arguments that 'asserting gender to be non-functional or meaningless is simply untenable'. Dolberg concedes that 'its effects are mostly moderate and its functions can generally be fulfilled by other means as well', but this could also be said about many other grammatical categories. Verbal tense is, after all, also dispensable (cf. above).

6 Conclusion

We are used to agreement targets contributing meaning when they are in the pronominal domain, and less so when they are attributive. Yet it is not the case that gender agreement on determiners always is meaningless – as cases like *en fyr* / *et fyr* show. Here, the agreement target contributes even inside the NP. This seems problematic for the repeated claim that agreement in general and inside the NP/DP in Scandinavian in particular is without meaning. The idea of agreement as mere feature copying is in trouble when the 'target' is making an independent contribution in terms of meaning. My argument is thus in line with, for instance, Corbett (2006), Haug & Nikitina (2016) and Kibrik (2019).

The argument in this paper supports Corbett's (2006) reluctance towards drawing a sharp line around the NP for agreement purposes. Equating the meaningful/meaningless distinction with some particular position on the Agreement Hierarchy remains less promising. Agreement is 'a multifaceted phenomenon' (Thorvaldsdóttir 2019), and since gender is defined by agreement, gender is multi-faceted too.

Even if we should not exclude some arbitrariness for lexical gender in Scandinavian, this arbitrariness has sometimes been overstated, and an emphasis on arbitrariness can be positively unhelpful, heuristically. Gender may be puzzling, but it is not quite as different from other categories as we may have thought.

Acknowledgments

This paper has been long in the making, and several colleagues have helped at various stages. Thanks to Marit Westergaard and Terje Lohndal for kindly inviting me to their project *MultiGender* at the Centre for Advanced Studies at the Norwegian Academy of Science and Letters, 2019–2020, with its excellent working conditions, and to Marit, Terje and many MultiGender participants for their helpful response to different presentations. I am also grateful to Grev Corbett for generous and constructive comments on an early manuscript, and to Ragnhild Eik, Dag Haug and Michele Loporcaro for helping me in various ways. Finally, sincere thanks to two incisive and generous FJL referees and the Editor, Olli Silvennoinen.

Abbreviations

C	common
DEF	definite
F	feminine
INDF	indefinite
M	masculine
MF	masculine/feminine
N	neuter
PL	plural
SG	singular

References

Ackerman, Farrell & Malouf, Robert. 2015. Implicative relations in word-based morphological systems. In Hippisley, Andrew & Stump, Greg T. (eds.), *Cambridge handbook of morphology*, 297–328. Cambridge: Cambridge University Press. https://doi.org/10.1017/981139814720.012

Åfarli, Tor Anders & Nygård, Mari & Riksem, Brita R. 2022. Gender 'translation' and distributed gender: Evidence from the Norwegian DP and Language Mixing. *Studia Linguistica* 76. 626–669. https://doi.org/10.1111/stul.12190

Barlow, Michael. 1999. Agreement as a discourse phenomenon. *Folia Linguistica* XXXIII. 187–211. https://doi.org/10.1515/flin.1999.33.1-2.187/html

Beito, Olav T. 1954. *Genusskifte i nynorsk* [Gender change in New Norwegian]. Oslo: Jacob Dybwad.

Belyaev, Oleg & Dalrymple, Mary & Lowe, John. 2015. Number mismatches in coordination: An LFG analysis. In Butt, Miriam & King, Tracy Holloway (eds.), *Proceedings of the LFG 15 Conference*, 26–46. Stanford: CSLI Publications.

Bobrova, Maria. 2013. *Genus ved homofone substantiver* [Gender in homophonous nouns]. MA thesis, University of Oslo. https://www.duo.uio.no/handle/10852/39118

Busterud, Guro & Lohndahl, Terje & Rodina, Yulia & Westergaard, Marit. 2019. The loss of feminine gender in Norwegian. *Journal of Comparative Germanic Linguistics* 22. 141–167. https://doi.org/10.1007/s10828-019-09108-7

Carstairs-McCarthy, Andrew. 2010. The evolution of morphology. Oxford: Oxford University Press.

Carstens, Vicki. 2000. Concord in minimalist theory. *Linguistic Inquiry* 31. 319–355. https://doi.org/10.1162/002438900554370

Comrie, Bernard. 1989. *Language universals and linguistic typology.* 2nd edition. Oxford: Blackwell.

Corbett, Greville G. 1991. *Gender*. Cambridge: Cambridge University Press.

Corbett, Greville G. 2006. *Agreement*. Cambridge: Cambridge University Press.

Corbett, Greville G. 2019. Pluralia tantum nouns and the theory of features. *Morphology* 29. 51–108. https://doi.org/10.1007/s11525-018-9336-0

Dąbrowska, Ewa. 2004. *Language, Mind and Brain*. Edinburgh: Edinburgh University Press.

Dahl, Östen. 2000a. Animacy and the notion of semantic gender. In Unterbeck, Barbara & Rissanen, Matti (eds.), *Gender in grammar and cognition*, 99–115. Berlin: Mouton de Gruyter.

Dahl, Östen. 2000b. Elementary gender distinctions. In Unterbeck, Barbara & Rissanen, Matti (eds.), *Gender in grammar and cognition*, 577–595. Berlin: Mouton de Gruyter.

Dahl, Östen &Velupillai, Viveka. 2013. The past tense. In Dryer, Matthew & Haspelmath, Martin (eds.), *The World atlas of language structures online* (v.2020.3). https://doi.org/10.5281/zenodo.7385533 (Also available at http://wals.info/chapter/66)

Dammel, Antje & Schallert, Oliver. 2019. Introduction: On the benefits of analysing morphological variation by linking theory and empirical evidence. In Dammel, Antje & Schallert, Oliver (eds.), *Morphological variation: Theoretical and empirical perspectives* (Studies in Language Companion Series 207), 1–27. Amsterdam: John Benjamins. https://doi.org/10.1075/slcs.207.01.sch

Davidson, Herbert. 1990. *Han hon den: Genusutvecklingen i svenskan under nysvensk tid*. [He she it: Gender development in Swedish during the New Swedish period]. Lund: Lund University Press.

Di Garbo, Francesca & Agbetsoamedo, Yvonne. 2018. Non-canonical gender in African languages: A typological survey of interactions between gender and number, and between gender and evaluative morphology. In Fedden, Sebastian & Audring, Jenny & Corbett, Greville G. (eds.), *Non-canonical gender systems*, 176–210. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780198795438.003.0008

Dolberg, Florian. 2019. *Agreement in language contact: Gender development in the Anglo-Saxon Chronicle* (Studies in Language Companion Series 208). Amsterdam: John Benjamins.

Enger, Hans-Olav. 2002. Stundom er ein sigar berre ein sigar: Problem i studiet av leksikalsk genus [Sometimes, a cigar is just a cigar: Problems in the study of lexical gender]. *Maal og Minne* 2002. 135–151.

Enger, Hans-Olav. 2004a. On the relationship between gender and declension: A diachronic perspective from Norwegian. *Studies in Language* 28. 51–82. https://doi.org/10.1075/sl.28.1.03eng

Enger, Hans-Olav. 2004b. Tre endringer i det skandinaviske genussystemet i lys av grammatikaliseringsteori [Three changes in the Scandinavian gender system in light of grammaticalisation theory]. *Arkiv för nordisk filologi* 119. 125–147.

Enger, Hans-Olav. 2010. Partial and competing motivations for gender. In Dammel, Antje & Kürschner, Sebastian & Nübling, Damaris (eds.), *Kontrastive Germanistische Linguistik = Germanistische Linguistik* [GL] 206–209. 673–693. Hildesheim: Olms.

Enger, Hans-Olav. 2013. Scandinavian pancake sentences revisited. *Nordic Journal of Linguistics* 36. 275–301. https://doi.org/10.1017/S0332586513000280

Enger, Hans-Olav. 2015. When friends and teachers become hybrids (even more than they were). In Fleischer, Jürg & Rieken, Elisabeth & Widmer, Paul (eds.), *Agreement from a diachronic perspective* (Trends in Linguistics Studies and Monographs), 215–233. Berlin: De Gruyter Mouton. https://doi.org/10.1515/9783110399967-011

Enger, Hans-Olav & Corbett, Greville G. 2012. Definiteness, gender, and hybrids: Evidence from Norwegian dialects. *Journal of Germanic Linguistics* 24. 287–324. https://doi.org/10.10717/S1470542712000098

Faarlund, Jan Terje & Lie, Svein & Vannebo, Kjell Ivar. 1997. *Norsk referansegrammatikk* [Norwegian Reference Grammar]. Oslo: Universitetsforlaget.

Feist, Timothy. 2020. Nominal classification: Does it play a role in referent disambiguation? *Studies in Language* 44. 191–230. https://doi.org/10.1075/sl.19026.fei

Fraurud, Kari. 2000. Proper names and gender in Swedish. In Unterbeck, Barbara & Rissanen, Matti (eds.), *Gender in grammar and cognition*, 167–221. Berlin: Mouton de Gruyter.

Haig, Geoffrey & Forker, Diana. 2018. Agreement in grammar and discourse: A research overview. *Linguistics* 56. 715–734. https://doi.org/10.1515/ling-2018-0014

Halmøy, Madeleine. 2016. *The Norwegian nominal system: A neo-Saussurean perspective* (Trends in Linguistics. Studies and Monographs 294). Berlin: de Gruyter.

Halse, Gro Egset. 2004. *Genustilordning i nynorsk: Ei datamaskinell etterprøving* [Gender assignment in New Norwegian: A computational assessment]. MA thesis, University of Bergen. https://bora.uib.no/bora-xmlui/handle/1956/2823

Hansen, Erik & Heltoft, Lars. 2011. *Grammatik over det Danske Sprog* [Grammar of the Danish Language]. Odense: Syddansk Universitetsforlag.

Haug, Dag Trygve Truslew & Nikitina, Tatiana. 2016. Feature sharing in agreement. *Natural Language & Linguistic Theory* 34. 865–910. https://doi.org/10.1007/s11049-015-9321-9

Haug, Kristin Nordbø. In prep. *Gender loss at an early stage*. Manuscript.

Haugen, Tor Arne. 2014. Adjectival predicators and approaches to complement realisation. *Lingua* 140. 83–99. https://doi.org/10.1016/j.lingua.2013.12.007

Heim, Stefan & Alter, Kai & Frederici, Angela. 2005. A dual-route account for access to grammatical gender. *Anatomy and Embryology* 210. 473–483. https://doi.org/10.1007/s00429-005-0032-6

Hempel, Carl Gustav. 1966. *Philosophy of natural science*. Englewood Cliffs NJ: Prentice-Hall.

Herce, Borja. 2020. On morphemes and morphomes: Exploring the distinction. *Word Structure* 13. 45–68. https://doi.org/10.3366/word.2020.0159

Herce, Borja. 2023. *The typological diversity of morphomes*. Oxford: Oxford University Press.

Holmes, Phil & Hinchliffe, Ian. 2013. *Swedish: A Comprehensive Grammar*. 3rd edn. London: Routledge.

Hudson, Richard A. 2010. *An introduction to Word Grammar*. Cambridge: Cambridge University Press.

Josefsson, Gunlög. 2009. Peas and pancakes: On apparent disagreement and (null) light verbs in Swedish. *Nordic Journal of Linguistics* 32. 35–72. https://doi.org/10.1017/S0332586509002030

Josefsson, Gunlög. 2013. Gender in Scandinavian: On the gender systems in Mainland Scandinavian, with focus on Swedish. Manuscript. LingBuzz. http://ling.auf.net/lingbuzz/001966

Josefsson, Gunlög. 2014. Scandinavian gender and pancake sentences: A reply to Hans-Olav Enger. *Nordic Journal of Linguistics* 37. 431–449. https://doi.org/10.1017/S0332586514000286

Kibrik, Andrej A. 2019. Rethinking agreement: Cognition-to-form mapping. *Cognitive Linguistics* 30. 37–83. https://doi.org/10.1515/cog-2017-0035

Köpcke, Klaus-Michael. 1982. Untersuchungen zum Genussystem der deutschen Gegenwartssprache [Studies in the gender system of contemporary German] (*Linguistische Arbeiten* 122). Berlin: de Gruyter.

Kristoffersen, Kristian Emil. 2000. Ordklassane pronomen og determinativ – kor gode er argumenta for å skilje dei i norsk? [The word-classes pronoun and determinative – how good are the arguments for keeping them apart in Norwegian?] *Maal og Minne* 2000, 181–194.

Kürschner, Sebastian & Nübling, Damaris. 2011. The interaction of gender and declension in Germanic languages. *Folia Linguistica* 45. 355–388. https://doi.org/10.1515/flin.2011.014

Kvinlaug, Anders. 2011. *Genustilordning i Kristiansandsdialekten* [Gender assignment in the Kristiansand dialect]. MA thesis, University of Oslo. https://www.duo.uio.no/bitstream/handle/10852/26731/Kvinlaug_Master.pdf?sequence=2&isAllowed=y

Landau, Ivan. 2016. DP-internal semantic agreement: A configurational analysis. *Natural Language & Linguistic Theory* 34. 975–1020. https://doi.org/10.1007/s11049-015-9319-3

Langacker, Ronald W. 1987. *Foundations of Cognitive Grammar, volume I: Theoretical foundations*. Stanford: Stanford University Press.

Langacker, Ronald W. 1991. *Foundations of Cognitive Grammar, volume II: Descriptive application*. Stanford: Stanford University Press.

Langacker, Ronald W. 2008. *Cognitive Grammar: A basic introduction*. Oxford: Oxford University Press.

Lehmann, Christian. 1988. On the function of agreement. In Barlow, Michael & Ferguson, Charles (eds.), *Agreement in natural language*, 55–65. Stanford: CSLI publications.

Lødrup, Helge. 2011. Hvor mange genus er det i Oslo-dialekten? [How many genders are there in the Oslo dialect?] *Maal og Minne* 2011. 120–137. https://ojs.novus.no/index.php/MOM/article/view/330

Maiden, Martin. 2016. The Romanian alternating gender in diachrony and synchrony. *Folia Linguistica Historica* 37. 111–144. https://doi.org/10.1515/flih-2016-0004

Miceli, Gabriele & Turriziani, Patrizia & Caltagirone, Carlo & Capasso, Rita & Tomaiulo, Francesco & Caramazza, Alfonso. 2002. The neural correlates of grammatical gender. *Journal of Cognitive Neuroscience* 14. 618–628. https://doi.org/10.1162/08989290260045855

Nordström, Jackie. 2024. Semantic agreement and the dual model of language. *Zeitschrift für Sprachwissenschaft* 43. 65–92. https://doi.org/10.1515/zfs-2024-2009

Næss, Åshild. 2007. *Prototypical transitivity* (Typological Studies in Language 72). Amsterdam: John Benjamins.

O'Neill, Paul. 2013. The morphome and morphosyntactic/semantic features. In Cruschina, Silvio & Maiden, Martin & Smith, John Charles (eds.), *The boundaries of pure morphology*, 221–246. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199678860.003.0012

Palmer, F.R. 1984. *Grammar. 2nd edn*. Harmondsworth: Penguin.

Papazian, Eric. 1978. Han og ho: Eit uromantisk oversyn over formene av desse pronomena og bruken av dei i norsk. [He and she: A non-romantic survey of the forms of these pronouns and their use in Norwegian]. In Hoff, Ingeborg (ed.), *På leit etter ord* [In search of words], 235–281. Oslo: Universitetsforlaget.

Parkkonen, Lotta. 2011. *Genus ved nyord i norsk* [Gender in new words in Norwegian]. MA thesis, University of Oslo. https://www.duo.uio.no/handle/10852/26735

Percus, Orin. 2011. Gender features and interpretation: A case study. *Morphology* 21. 167–196. https://doi.org/10.1007/s11525-010-9157-2

Rabb, Viveca. 2007. *Genuskongruens på reträtt: Variation i nominalfrasen i Kvevlaxdialekten* [Gender agreement on retreat: Variation in the NP in the Kvevlax dialect]. Dissertation. Åbo [Turku]: Åbo Akademis förlag.

Sasse, Hans-J. 1993. Syntactic categories and subcategories. In Jacobs, Joachim & von Stechow, Arnim & Sternefeld, Wolfgang & Vennemann, Theo (eds.), *Syntax: An international handbook of contemporary research*, 646–686. Berlin: de Gruyter.

Smith, John Charles. 2022. The boundaries of inflexion and periphrasis. In Ledgeway, Adam & Smith, John Charles & Vincent, Nigel (eds.), *Periphrasis and inflection in diachrony*, 61–91. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780198870807.003.0003

Spencer, Andrew. 2002. Gender as an inflectional category. *Journal of Linguistics* 38. 279–312. https://doi.org/10.1017/S0022226702001421

Steinmetz, Donald. 1986. Two principles and some rules for gender in German: Inanimate nouns. *Word* 37. 189–217.

Svenonius, Peter. 2017. Declension class and the Norwegian definite suffix. In Gribanova, Vera & Shih, Stephanie S. (eds.), *The morphosyntax-phonology connection*, 325–361. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780190210304.003.0012

Teleman, Ulf. 1969. On gender in a generative grammar of Swedish. *Studia Linguistica* 23. 27–67.

Teleman, Ulf. 1987. Hur många genus finns det i svenskan? [How many genders are there in Swedish?]. In Teleman, Ulf (ed.), *Grammatik på villovägar*, 106–115. Stockholm: Esselte.

Teleman, Ulf & Hellberg, Staffan & Andersson, Erik. 1999. *Svenska Akademiens Grammatik 2: Ord* [The Swedish Academy Grammar 2: Words]. Stockholm: Norstedts.

Thorvaldsdóttir, Thorbjorg. 2019. Agreement with conjoined singular noun phrases in Icelandic. *Glossa* 4(53), 1–33. https://doi.org/10.5334/gjgl.696

Trosterud, Trond. 2001. Genustilordning i norsk er regelstyrt [Gender assignment in Norwegian is rule-governed]. *Norsk Lingvistisk Tidsskrift* 19. 29–58.

Urek, Olga & Lohndal, Terje & Westergaard, Marit. 2022. En splyv eller et splyv? Tilordning av grammatisk genus til pseudosubstantiv i norsk [A.C splyv or a.N splyv? Assignment of grammatical gender on pseudo-nouns in Norwegian]. *Norsk Lingvistisk Tidsskrift* 40. 7–27. https://ojs.novus.no/index.php/NLT/article/view/2070

Van Epps, Briana & Carling, Gerd. 2017. From three genders to two: The sociolinguistics of gender shift in the Jämtlandic dialect of Sweden. *Acta Linguistica Hafniensia* 49. 53–84. https://doi.org/10.1080/03740463.2017.1286811

Van Epps, Briana & Carling, Gerd & Sapir, Yair. 2021. Gender assignment in six North Scandinavian Languages. *Journal of Germanic Linguistics* 33. 264–315. https://doi/org/10.1017/S1470542720000173

Venås, Kjell. 1990. *Norsk grammatikk: Nynorsk* [Norwegian grammar: Nynorsk]. Oslo: Det norske samlaget.

Wälchli, Bernard & Di Garbo, Francesca. 2019. The dynamics of gender complexity. In Di Garbo, Francesca & Olsson, Bruno & Wälchli, Bernard (eds.), *Grammatical gender and linguistic complexity, vol. II,* 201–364. Berlin: Language Science Press. https://doi.org/10.5281/zenodo.3462784

Contact information:

Hans-Olav Enger
University of Oslo
Department of Linguistics and Scandinavian Studies
h.o.enger@iln.uio.no

‘To do or not to do’: Semi-lexical affixes in (post)classical Greek

Victoria Beatrix Fendel
University of Oxford

Abstract

The support verbs ποιέομαι *poieomai* ‘to do’ and τίθημι *tit*^{h}*ēmi* ‘to put’ exist in bound and unbound forms from classical into medieval times, ποιέομαι *poieomai* ‘to do’ as ποιέομαι *poieomai* and -ποιέομαι *-poieomai*, τίθημι *tit*^{h}*ēmi* ‘to put’ as τίθημι *tit*^{h}*ēmi* and -θετέω -*t*^{h}*eteō*. They differ from auxiliaries, in that they are semi-lexical as they contribute to the event structure of the verb phrase. The article draws on the literary corpus of the *Thesaurus Linguae Graecae* and the documentary corpus of the *Duke Database of Documentary Papyri* to answer three research questions: (i) Are support verbs the unbound alternative of bound affixes? (ii) How do semi-lexical support verbs become semi-grammatical affixes? (iii) Why did -ποιέομαι *-poieomai* ‘to do’ and -θετέω - *t*^{h}*eteō* ‘to put’ not become productive? It finds that the bound and unbound forms of the support verb differ in the semantics of the lexical unit and its pragmatic embedding. Through the univerbation of leader words and their subsequent reanalysis, a new word-formation pattern emerges, which never becomes productive but is available as a creative option especially in technical registers (akin to English *readable* alongside established *legible*) from the Ptolemaic/Roman periods onwards. This new word-formation pattern involves a semi-lexical affix, just like the unbound support verb.

Keywords: support-verb construction, boundedness, univerbation, productivity, technical register

1 Introduction

The verbs ποιέω *poieō* (active)/ποιέομαι *poieomai* (middle) ‘to do’ and τίθημι *tit*^{h}*ēmi* (active)/τίθεμαι *tit*^{h}*emai* (middle) ‘to put’[1] can be lexical, semi-lexical, or grammatical in (post)classical Greek. When the verbs function as lexical verbs, they fill the predicate slot of the sentence by themselves, e.g., *Paul made me a cake*. When they function as semi-lexical support verbs, they fill the predicate slot of a sentence in combination with

[1] The active / middle distinction, e.g., active ποιέω *poieō* vs. middle ποιέομαι *poieomai* is theoretically a distinction between a three-argument (causative) and a two-argument frame (active). However, verb lability is attested already in classical times such that we find instances of ‘to do’ with active morphology in two-argument frames (or even one-argument frames) (Lavidas 2009). In post-classical times, the distinction between active and middle morphology was apparently no longer determined semantico-syntactically but socio-pragmatically, i.e., a middle form would be used in higher-register contexts (Vives Cuesta & Acero 2022). This is similarly to what has been proposed for verbal complementation patterns (Bentein 2017).

VERTAISARVIOITU
KOLLEGIALT GRANSKAD
PEER-REVIEWED
www.tsv.fi/tunnus

Finnish Journal of Linguistics 37 (2024), 33–72
DOI: 10.61197/fjl.142559

a nominal predicative component, e.g., *Paul made a suggestion*.[2] When they function as grammatical auxiliaries, they fill the predicate slot of a sentence in combination with a verbal component, e.g., *Paul made me laugh*.

The present article argues that ποιέω *poieō* 'to do' and τίθημι *tit^{h}ēmi* 'to put' when used as semi-lexical support verbs can become semi-lexical affixes, i.e., we witness a process from an unbound to a bound form. The battle in the literature is blazing as to whether support verbs become affixes due to grammaticalization. This article suggests an alternative pathway. However, rather than the transition from unbound to bound being a linear process, the bound and unbound forms co-exist over centuries with varying degrees of extensibility of the patterns (Barðdal 2008: 20; Baayen 2009) moderated by indexicality (Bentein 2019). The following three sub-sections introduce the reader briefly to (post)classical Greek, support-verb constructions, and the blazing battle over the grammaticalization of support verbs before presenting the research questions and the structure of the remaining article.

1.1 (Post)classical Greek

(Post)classical Greek is a morphology-rich stage of the language with a fully fletched compounding system lexically speaking and a fully fletched system of derivational morphology grammatically speaking. Against this background, the development of lexical and/or grammatical affixes is conceivable (as mechanisms are in place) but also raises the question of how the new item fits into the system. Redundancy in language does not exist from a variationist perspective.

The Greek language as a whole offers a diachronic depth of at least 3,000 years of written records, from the Homeric epics (ca. 9th / 8th c. BC) to the modern day. The present article is concerned with those about first 2,000 years of its history which are by now corpus varieties, i.e., varieties whose native speakers are the texts (Fleischman 2000: 34). The available diachronic depth allows us to observe developments over long periods of time. The fact that the stages of interest are corpus varieties challenges approaches to multi-word expressions developed for modern languages, specifically in Natural Language Processing. These draw on grammaticality statements and the reverse-engineering of syntactic configurations from attestations.

(Post)classical Greek word order is not syntactically but information-structurally driven. Thus, unlike in languages such as modern English or French in which the support verb and the predicative noun have to appear in a specific order due to syntactic pressures, (post)classical Greek allows for any order as long as there is an information-structural reason or the elements in question are functionally linked. This functional link surfaces in languages such as modern English in the form of constraints on discontinuity and modi-

[2] Standard grammar of (post)classical Greek label 'predicative' adjectives and nouns in combination with the copular verb 'to be' (εἶναι *einai* or especially in the perfect / aorist / pluperfect γίγνομαι *gignomai*). These adjectives or nouns complete the predication as the copula alone is not a full predicate. Similarly, the support verb by itself is not a full predicate but the predication needs to be completed by a predicative element. Formally, the predicative adjectives and nouns with copular verbs appear in the same case as the subject component of the copula; conversely, predicative nouns with support verbs appear in the verb's object slot (see however also Tronci 2016; Tronci 2017a). In both constructions, multiple elements form the complex predicate functionally and constitute a verbal multi-word expression formally. Gross (1998) would accept the verb 'to be' as a support verb and Jiménez López (2021) suggests that the copular verb acts as the lexical passive in support-verb constructions.

fiability, e.g., *Paul took heart* but not **Paul heart took* or **Paul took great heart* (see further Fendel 2024).

For reasons of accessibility, one-word translations of Greek terms are provided in the running text. These should not be taken as more than guidance. As regards support-verb constructions, standard dictionaries are a minefield. Three cases are discussed in detail in Fendel (2024) with suggestions for the improvement of dictionary entries. In short, dictionaries do not consider support-verb constructions except when they form non-compositional lexical units, thus overall inconsistently. Some members of a support-verb-construction family around a predicative noun may be non-compositional, whereas others may not be.[3] Furthermore, standard dictionaries, such as Liddell-Scott-Jones (LSJ), are reductionist in that they view support-verb constructions when listed as the equivalent to a simplex verb derived from the same root as the predicative noun in the support-verb construction (e.g. LSJ s.v. χάρις *kʰaris* III.1.a (‘favour’) χ. δοῦναι *kʰ[arin] dounai* = χαρίζεσθαι *kʰarizesthai* ‘to do a favour’). They thus often omit patterns and associated meanings. Finally, if listed, support-verb constructions are labelled ‘prose phrases’ indicating a stylistic nuance. However, they are not limited to prose texts nor are all support-verb construction variants of simplex verbs, neither semantically nor pragmatically. Research in modern languages has amply shown this (Wittenberg & Levy 2017; Wittenberg & Snedeker 2014; Wittenberg & Trotzke 2021).

1.2 Support-verb constructions

Support-verb constructions[4] (SVCs) consist of a verbal and a nominal component that together fill the predicate slot, as *made the suggestion* in *I made the suggestion that he join*. The nominal component, the predicative noun, must be eventive. The article takes a broad approach in that it accepts any noun that can be reconceptualised as eventive, whether formally related to a verb (by derivational morphology) or not, including items with a primarily concrete meaning that need reconceptualising, e.g., *picture* in *to take a picture* (Radimský 2011), and items whose meaning undergoes metaphorical extension, e.g., *heart* in *to take heart* (Sheinfux et al. 2019). This entails several exclusions,

---

[3] An anonymous reviewer pointed out BGU.3.941 (Herakleopolis, AD 376–377 (EBG), receipt), 13–14 ὁμολογῶ ἐντεῦθεν μηδένα λόγον ἔχειν πρὸς σὲ μηδὲ ἐπελθεῖν σοι *ʰomologō enteutʰen mēdena logon ekʰein pros se mēde epeltʰein soi* ‘I thus agree not to have any charge against you nor proceed against you’ as an example of a support-verb construction in an idiomatic phrase in the papyri. Support-verb constructions appear in many idiomatic phrases in the papyri and otherwise, e.g., ἐξουσίαν ἔχω *exousian ekʰō* ‘to have power (over)’ in wills (Fendel 2023a; Fendel 2025). An example with ποιέομαι *poieomai* ‘to do’ is expressions of origin, such as P.Grenf.2.78 (Kysis, AD 307–308, petition), l. 15 τὴν ὁρμὴν ποιούμενος *tēn ʰormēn poioumenos* ‘coming from’ (Zilliacus 1956).

[4] Structures, such as *to make a remark*, have been dealt with in three major research traditions. Each has its own terminology. Primarily in studies on German, we find the term *Funktionsverbgefüge* ‘function-verb construction’ (von Polenz 1963; Kamber 2008; Storrer 2009; De Knop & Hermann 2020). Originating in Jespersen’s (1954) *Modern English Grammar on Historical Principles*, we find the term light-verb construction. This was adopted in language-contact studies (e.g., Bakker 2003: 132; Myers-Scotton 2002: 134–139) and Natural Language Processing frameworks (e.g., PARSEME). Originating in the Lexicon-Grammar framework of the *Laboratoire d’Automatique Documentaire et Linguistique* (e.g., Gross 1998), we find the term support-verb construction (*construction à verbe support*). The latter term is adopted in line with Jiménez López’ (2016) seminal article on the structures in classical Greek and because the support-verb-construction tradition includes structures in which the support verb contributes to the event structure in the form of voice, *aktionsart*, polarity, and register (e.g., Vivès 1983; Gross 1989). Thus, the verb is neither purely a ‘function’ word nor fully semantically ‘light’.

i.e., items referring to (i) human beings (professions, kinship terms, etc.), e.g., τέκνον *teknon* 'child', (ii) concrete locations (place names, references to places), e.g., ἀγορά *agora* 'market square', along with (iii) syntactic nominalisation, including nominalised adjectives, e.g., (τὸ) ἀσφαλές (to) *asp^{h}ales* 'safe', and (iv) nouns that cannot through reconceptualization or metaphorical extension adopt an eventive meaning.

The verbal component, the support verb, plays a supporting role. The support verb is morphosyntactically marked for tense, aspect, and mood (1). Through lexical substitution, it can also indicate *aktionsart* (2), polarity (3), and voice (4) (cf. Collins 2018):

(1)	*He*	***made***	ART	*suggestion*	*that*	*she*	*join.*
	he.S	**make.PST.3SG**	*the*	suggestion.PN	CONJ	she.S	join.PRS.3SG
(2)	*He*	***has***	*a*	*suggestion.*			
	he.S	**have.PRS.3SG**	ART	suggestion.PN			
(3)	*He*	***lacks***	*patience.*				
	he.S	**lack.PRS.3SG**	patience.PN				
(4)	*He*	***got***	*word*	*yesterday.*			
	he.S	**get.PST.1SG**	word.PN	yesterday			

Support verbs differ from auxiliary verbs, in that (i) they contribute to the event structure of the support-verb construction (Butt 2010), (ii) they are never phonetically reduced (Loporcaro 2022), (iii) they are not fully productive but selected based on the collocational field of the predicative noun (Kamber 2008; Bonami 2015). Auxiliaries are grammatical function words, support verbs are not.

The support verb in support-verb constructions profiles lexical aspect (Vendler 1967), recipient passive (Fendel 2024), negation of intensity as opposed to contrast (Fendel 2023b), and allows for semantic (external, i.e. with the support-verb construction) and syntactic (internal, i.e. with the predicative noun or the support verb) agreement with the support-verb construction (Fendel 2023a; Janse 2023). This sets support verbs apart from auxiliaries, which indicate grammatical aspect and patient passive and do not allow for internal agreement.

Boye (2023) offers a distinction between lexical and grammatical items that is based on discursively primary and secondary status. Lexical items are discursively primary; grammatical items are discursively secondary. This is tested through the permissibility of being (i) focussed, (ii) addressed in subsequent discourse, (iii) modified, and (iv) of standing alone in an utterance. (5) to (7) test English *to make a suggestion*:

(5) *It is a **suggestion** that I made, **not a plan**.*

(6) *I made a suggestion. The others liked **my suggestion** and implemented it.*

(7) a. *I made a **quick** suggestion.*
b. *I **quickly** made a suggestion.*

Focalisation (contrastive focus) is shown in (5). This is possible for the predicative noun and sometimes the support verb. Lexical anaphora is shown in (6); morphosyntactic anaphora by means of pronominalisation is equally applicable to the predicative noun (*it* instead of *my suggestion*). The support verb cannot be anaphorically resumed, neither lexically nor morphosyntactically. Finally, (7) illustrates external vs internal modification in support-verb constructions. An attributive phrase qualifies the predicative noun; an adverb applies to the support-verb construction rather than just the support verb. The support verb *to make* cannot stand on its own in any of these utterances.

(8) to (10) apply the observations to Greek examples from the classical literary Attic sample:

(8) CG Plato, *Republic*, 337e[5]

ἄλλου	δ’	ἀποκρινομένου	**λαμβάνῃ**
allou	*d’*	*apokrinomenou*	*lambanē*
other.GEN.SG.M	PRT	answer.PRS.PTCP.MID.GEN.SG.M	take.PRS.SBJV.ACT.3SG

λόγον	**καὶ**	ἐλέγχῃ.
logon	*kai*	*elegk^{h}ē*
word.ACC.SG.M	and	refute.PRS.SBJV.ACT.3SG

‘(such that) when someone else gives an answer he listens and rejects (it)’

In (8), two support verbs are contrasted, compare (5).

(9) CG Plato, *Gorgias*, 484d

τῶν	**λόγων**	**οἷς**
tōn	*logōn*	*^{h}ois*
the.GEN.PL.M	word.GEN.PL.M	which.DAT.PL.M

δεῖ	**χρώμενον**	ὁμιλεῖν
dei	*k^{h}rōmenon*	*^{h}omilein*
be.necessary.PRS.IND.ACT.3SG	use.PRS.PTCP.MID.ACC.SG.M	mitigate.PRS.INF.ACT

‘of the words using which it is necessary to mitigate’

In (9), the predicative noun is resumed by a relative pronoun, compare (6).

(10) a. CG Antiphon, *Speech*, 4.2.1[6]

Ὅτι	μὲν	**τοὺς**	**βραχεῖς**	**λόγους**	ἐποιήσαντο
^{h}oti	*men*	*brak^{h}eis*	*tous*	*logous*	*epoiēsanto*
that	PRT	brief.ACC.PL.M	the.ACC.PL.M	word.ACC.PL.M	do.AOR.IND.MID.3SG

‘that they kept their remarks brief’

[5] ἐλεγχω *elegk^{h}ō* ‘to reject’ is technically a verb of realization, ‘qui [a] le comportement syntaxique des V_{supp}, mais qui (…) [est] sémantiquement pleins’ (Mel’čuk 2004: 208).

[6] The adjective is in predicative position.

b. CG Thucydides, *Histories*, 5.18.11

λόγοις	**δικαίοις**	**χρωμένοις**
logois	*dikaios*	*k^{h}rōmenois*
word.DAT.PL.M	just.DAT.PL.M	use.PRS.PTCP.MID.DAT.PL.M

'speaking justly'

In (10), an attributive adjective modifies the predicative noun, compare (7). Thus, by Boye's definition, support verbs are semi-lexical and semi-grammatical (Butt & Geuder 2001; Grimshaw & Mester 1988).

Some verbs, such as ἔχω *ek^{h}ō* 'to have', can function as auxiliaries, (11) and support verbs,[7] (12):

(11) CG Euripides, *Trojan Women*, 315–321

καταστένουσ'	ἔχεις
katastenous'	*ek^{h}ō*
lament.PRS.PTCP.ACT.NOM.SG.F	have.PRS.IND.ACT.2SG

'you are continually lamenting' (Bentein (2016: sec. 4.3.4), 'exceptional example')

(12) CG Antiphon, *Speech*, 2.3.6

οὐδεμίαν	ἐλπίδα	εἶχε
oudemian	*elpida*	*eik^{h}e*
no.ACC.SG.F	hope.ACC.SG.F	have.IMPF.IND.ACT.3SG

'he had no hope'

In (12), ἔχω *ek^{h}ō* 'to have' forms a support-verb construction with the lexical noun ἐλπίδα *elpida* 'hope'; in (11), ἔχω *ek^{h}ō* 'to have' forms an imperfective periphrastic in combination with a present participle. While support-verb constructions can be discontinuous (Savary et al. 2018), periphrastic verbal forms tend towards adjacency (Bentein 2016: sec. 2.3.2; Ledgeway & Vincent 2022: 51–54). While support verbs reverse-select their predicative nouns (Bonami 2015), auxiliary verbs do not. While support-verb constructions contain verbal and nominal components, periphrastics as in (11) consist of two verbal components (see also Tronci & Logozzo 2022). Thus, while the same verb can function as a support verb and auxiliary, its properties differ depending on its function.[8]

Support-verb constructions form an internally heterogenous group with some more tending towards a lexical unit and others towards an analytic syntagm (Heine 2020; Croft 2022). Lexicographically, Mel'čuk (2023: 119) subsumes most support-verb constructions under collocations, which have a semantic pivot (Mel'čuk 2023: 35) in the predicative noun and can contain a quasi-unilexeme – "a degenerate lexeme […] [which] appears only in a particular collocation (or in a handful of collocations) and has at least one non-degenerate lexeme in its vocable, that is, it co-exists in the language with normal lexemes which have the same signifier and the same syntactics and from which it differs only by its strictly context-bound signified" (Mel'čuk 2023: 46), e.g., *pay* in *to pay*

[7] This is not uncommon in Greek and across languages (e.g., Vincent & Wheeler 2022; Concu 2022).

[8] We can either assume a polyfunctional item or a situation of homonymy.

attention. To collocations, operations such as passivisation and pronominalisation can be applied unlike to idioms, which form lexical units (Mel’čuk 2023: 74–75).[9] The items of interest in this article, i.e., ποιέομαι *poieomai* ‘to do’ and τίθημι *tit^{h}ēmi* ‘to put’ appear in analytic support-verb constructions.

ποιέομαι *poieomai* ‘to do’ appears in support-verb constructions in which (i) the profiling of the subject component introduced by the support verb coincides with the subject component implied by the predicative noun, (ii) the support verb does not add voice, polarity, *aktionsart*, or aspect information, and (iii) the semantic structure is fully compositional.[10] An example is ἐξέτασιν ποιέομαι *exetasin poieomai* ‘to make an inspection’ in (13):

(13) CG Xenophon, *Anabasis*, 1.7.1

Κῦρος	**ἐξέτασιν**	**ποιεῖται**
Kuros	*exetasin*	*poieitai*
Cyrus.NOM.SG.M	inspection.ACC.SG.F	make.PRS.IND.MID.3SG
τῶν	Ἑλλήνων	
tōn	*^{h}ellēnōn*	
the.GEN.PL.M	Greek.GEN.PL.M	

‘Cyrus made an inspection of the Greeks’

The subject of ἐξέτασιν ποιέομαι *exetasin poieomai* ‘to make an inspection’ is Cyrus. ἐξέτασις *exetasis* ‘inspection’ is a deverbal action noun with -σι- *-si-* being ‘the most productive action noun suffix; it could be added to virtually any verbal root’ (van Emde Boas et al. 2019: 262–269). The implied subject component is an Agent. ποιέομαι *poieomai* ‘to do’ profiles its subject as an Agent. In (13), it only adds verbal morphology to the construction. The semantically speaking object component τῶν Ἑλλήνων *tōn ^{h}ellēnōn* ‘of the Greeks’ appears as an objective genitive dependent on the predicative noun. The structure is analytic (Ledgeway & Vincent 2022: 51) and the predicative noun is the semantic head, Mel’čuk’s semantic pivot. The subsequent anaphora μετὰ δὲ τὴν ἐξέτασιν *meta de tēn exetasin* ‘and after the inspection’ shows this (Xenophon, *Anabasis* 1.7.2).

1.3 The battleground

Cross-linguistically/typologically, ‘to do’ as an auxiliary and support verb appears in typologically unrelated languages (Hoffmann 2023). In the Afro-Asiatic family, Kilani (2023) hypothesises a diachronic relationship between *jr(j)* ‘to do’ and the sentence-initial particle *jw* in Classical Egyptian (ca. 2000–1300 BC) (Loprieno 1995: 5–8). Later

[9] Mel’čuk’s (2023) theory on the distinction between collocations and idioms reflects his categorical approach to compositionality, e.g., *to spill the beans* is an idiom for him (pp. 74–76) although more transparent than *to kick the bucket*. Mel’čuk (2023: 53) dismisses the notion of transparency as a “psychological property of idioms”. Sheinfux et al. (2019: 66) take the opposite approach (cf. figuration). Here, only Mel’čuk’s collocations are of interest.

[10] Frameworks developed for verbal multi-word expressions in non-corpus languages, such as PARSEME, have developed test batteries including those relating to the semantics of the noun inside vs. outside of a support-verb construction and those relating to the deletion of the support verb (Test 10 [N-SEM] and Test 12 [V-REDUC] in particular). See: https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.0/?page=060_Specific_tests_-_categorize_VMWEs/020_Light-verb_constructions (last accessed 07 May 2024). Note that PARSEME is at heart a deterministic Natural Language Processing framework, whereas this article adopts a variationist approach.

stages of the language (Coptic, from ca. AD 100) show overt (e.g., Bohairic) or covert (e.g., Sahidic) 'to do' especially when integrating Greek loan verbs into the Egyptian morpho-syntactic frame (Reintges 2001; Quack 2017; Funk 2017; Egedi 2017; Grossman & Richter 2017; Zakrzewska 2017). In the Indo-European family, English has a 'to do' periphrastic to indicate negation and illocutionary force (e.g., Ellegård 1953; Schwarz 2004) but uses 'to do' also in support-verb constructions, e.g., *to do a favour*. Thus, English 'to do' can function as an auxiliary, a support verb, and a full lexical verb. In typologically isolated Ainu, Dal Corso (2022) hypothesises that the analytic *kii* 'to do' construction gradually replaces the preverbal phrasal clitic functioning as a negative. In Udi, a language of the northeast Caucasus, Harris suggests that *-b-* 'to do, make' becomes a classifier productively applied to transitive verbs (Harris 2008: 225).

In the Natural-Language-Processing-inspired PARSEME terminology, 'to do' support-verb constructions fall under the label LVC.full (light-verb construction full), in Jiménez López' collocation-focussed framework (Jiménez López & Baños 2022) under SVC-base (support-verb construction base), and in the function-verb-construction framework under *Nominalisierungsverbgefüge* 'nominal-verb construction' (Schutzeichel 2014: 10). Thus, 'to do' support-verb constructions fall into the most basic, most prototypical category in each framework. 'To do' can often replace the prototypical support verb (Langer 2004; Brown et al. 2012: 237) to form a comprehensible but unidiomatic construction, e.g., *to make a walk*.

We include -θετέω *-t^{h}eteō* 'to put' as a foil because (i) a diachronic relationship between ποιέομαι *poieomai* 'to do' and τίθημι *tit^{h}ēmi* 'to put' in support-verb constructions exists (De Pasquale 2023: 263; Cock 1981: 24; Schutzeichel 2014: 84, 154, and 162) due to lexical substitution (Tribulato 2015: 278 n. 32),[11] (ii) both verbs exist in bound and unbound forms in classical Greek already (Schutzeichel 2014: 136–138), and (iii) the bound forms appear in various fundamentally different combinations, e.g., + adjective (φανεροποιέω *p^{h}aneropoieō* 'to make clear') [causative], + non-eventive noun (τεκνοποιέω *teknopoieō* 'to procreate') [causative / discursively primary], + eventive noun (λόγοποιέομαι *logopoieomai* 'to remark', νομοθετέω *nomot^{h}eteō* 'to legislate'). These have often been grouped together (Tribulato 2015: 57 ποιος *poios* is not a separate lexeme; Asraf 2021), but should not be.

The literature is divided between those assuming a diachronic link between support and auxiliary verbs (e.g., Anderson 2006; Slade 2013; Ittzés 2022) and those who hypothesise one lexeme appearing in various frames (e.g., Butt 2010; Butt & Lahiri 2013). Butt and Lahiri (2013) argue against a diachronic link because support-verb constructions are complex predicates, i.e. the verbal and nominal components contribute to the event structure and the structure is monoclausal (Butt 1995; Butt 2010). It gets complicated when, as with 'to do' and 'to put' in a two-argument frame, the otherwise clear contribution of the support verb as regards profiling the subject component coincides with the subject argument implied by the predicative noun. This is the case in (1) *he made the suggestion that she join* (Agent); in (3) *he lacks patience*, the subject component implied by the predicative noun is a Volitional Undergoer (Næss 2007) but the support verb profiles the subject component as a Frustrative. The issue of a diachronic relation-

[11] Lexical substitution is diachronically common in Greek, cf. AG ἔρδω *erdō* – CG ποιέομαι *poieomai* – MG κάνω *kanō* 'to do' (Meissner 2016: 28; Ittzés forthcoming on the protolanguage), but also e.g., CG λαμβάνω *lambanō* – MG παίρνω *pairnō* 'to take', AG λαγχάνω *lagk^{h}anō* – CG τυγχάνω *tugk^{h}anō* 'to receive', AG τίθημι *tit^{h}ēmi* – CG δίδωμι *didōmi* 'to let' (causative).

ship between support and auxiliary verbs is here approached by considering the relationship between λογοποιέομαι *logopoieomai* and λόγον / λόγους ποιέομαι *logon* / *logous poieomai* ‘to make (a) remark(s)’ and νομοθετέω *nomot*^{h}*eteō* and νόμον / νόμους τίθημι *nomon* / *nomous tit*^{h}*ēmi* ‘to establish laws / to legislate’.

1.4 Research questions and overview

The article addresses three research questions: (i) Are support verbs the unbound alternative of bound affixes? (ii) How do semi-lexical support verbs become semi-grammatical affixes? (iii) Why did -ποιέομαι *-poieomai* ‘to do’ and -θετέω - *t*^{h}*eteō* ‘to put’ not become productive? The way the terms semi-lexical and semi-grammatical are used throughout is determined by which property of the support verb or affix has slight prevalence. The support verb and the affix are semi-lexical and semi-grammatical in that they contribute to the event structure (lexically) and act akin to the derivational (denominal) and inflexional morphology (grammatically). The article is based on (post)classical Greek corpora.

After this introductory section, Section 2 introduces the five data samples that the discussion draws on, Section 3 addresses research question one, Section 4 research question two, and Section 5 research question three. Section 6 summarises the findings and offers conclusions.

2 Datasets

The article is based on (i) the *Thesaurus Linguae Graecae*, which compiles literary texts from the archaic to the early modern periods, https://stephanus.tlg.uci.edu (subscription-only), and (ii) the *Duke Database of Documentary Papyri*, which compiles documentary texts of the postclassical and early medieval periods, https://papyri.info (open-access). From these datasets are drawn.

Literary texts were written with an artistic intent and in Classics and related disciplines are taken to include oratory, historiography, prose, and multiple verse-genres. Literary texts also include technical writing, such as commentaries, medical writings, lexicographic writings, or later homilies. In most cases, literary texts have come down to us through a tradition of copying and re-copying of manuscripts. Conversely, documentary texts are texts that were written for specific purposes in daily life in personal and professional contexts. They include texts such as letters, receipts, contracts, and wills. They can be divided into higher-register (H) and lower-register (L) texts based on their link or lack thereof with governmental affairs of any kind (Palme 2009: 361–363). They have come down to us in the form of papyri and potsherds that have survived in the sands of Egypt. The chance of preservation affects both corpora. Both corpora reflect diatopic, diastratic, and diachronic diversity of sources.

An internally homogenous (regarding register, dialect, timeframe, non-poetic literary genre) sub-sample of (i) is the *ECF Leverhulme* corpus, which is implemented in Sketch Engine, an online corpus-analysis tool (Fendel & Ireland 2023). The *ECF Leverhulme* corpus consist of half a million words of literary classical Attic oratory, historiography, and prose, Table 1.

Table 1. *ECF Leverhulme* corpus

Historiography (**203,186 words**):	Thucydides, *Histories*, vol. 1–5 (**98,945**); Xenophon, *Anabasis*, vol. 1–4 (**32,034**), *Memorabilia*, vol. 1–4 (**36,465**), *Hellenica*, vol. 1–4 (**35,742**);
Oratory (**143,937 words**):	Antiphon, *Speeches* 1–6 (**18,605**); Isocrates, *Speeches* 1–6 and 13 (**37,311**); Isaeus, *Speeches* 1–8 (**25,018**), Lysias, *Speeches* 1, 3, 7, 12, 14, 19, 22, 30, 31, 32 (**24,130**); Demosthenes, *Speeches* 1, 2, 3, 4, 6, 9, 18 (**38,873**);
Prose (**145,497 words**):	Plato, *Gorgias* (**27,790**), *Phaedrus* (**17,271**), *Republic*, vol. 1–3 (**28,688**); Aristotle, *Rhetoric* (**44,312**), *Politics*, vol. 1–3 (**27,436**)

20% of this corpus are fully annotated for support-verb constructions and the remainder is annotated for a select number of support-verb-construction families.[12]

Five datasets were annotated and analysed for Sections 3 to 5[13], as shown in Table 2:

Table 2. Datasets

Dataset 1	*ECF Leverhulme* Sketch Engine corpus, concordance[14] operating on lemmata for (i) λόγος *logos* + ποιέω *poieō* and (ii) νόμος *nomos* + τίθημι *tit^{h}ēmi* within 5 words of each other (manual correction for support-verb constructions and voice of the verb)
Dataset 2	*ECF Leverhulme* Sketch Engine corpus, concordance operating on lemmata for (i) λογοποιέομαι *logopoieomai* and (ii) νομοθετέω *nomot^{h}eteō* (manual correction for part-of-speech = verb and manual assessment of lexical anaphora)
Dataset 3	Duke *Database of Documentary Papyri* search for (i) οποι *opoi* and (ii) οθετ *ot^{h}et*, manual correction for appearance in verbs consisting of -ποιέω *-poieō*/-θετέω *-t^{h}eteō* and a nominal or adjectival component and subsequent classification on whether these verbs contain an eventive nominal component[15]
Dataset 4	*Duke Database of Documentary Papyri* search for (i) λόγον *logon* + lemma ποιέω *poieō*, (ii) λόγους *logous* + lemma ποιέω *poieō*, (iii) νόμον *nomon* + lemma τίθημι *tit^{h}ēmi*, and (iv) νόμους *nomous* + lemma τίθημι *tit^{h}ēmi* within 5 words of each other (manual correction for support-verb constructions and voice of the verb)[16]
Dataset 5	*Thesaurus Linguae Graecae* Text search – simple – lemma (i) ποιέομαι *poieomai* and (ii) θετεω *t^{h}ēteō* – substring match (the full list of lemmata returned is manually corrected for combination with an eventive nominal component)

[12] Dataset: DOI 10.5287/ora-g2op5v0em

[13] Datasets: DOI 10.5287/ora-0zzadxvj5

[14] Concordances are vertical tables showing the context of the selected item / lemma.

[15] Papyri.info: a simple string search. αποι *apoi* and ηποι *ēpoi* only returned one relevant hit, P.Babatha 22 καθαραποιοῦντός *katharapoiountos* 'cleansing' (Maoza, 130 AD, sale), which is a combination of adjectival and verbal components. Note the a-stem of the adjective, see Section 5.

[16] Papyri.info – Search: [accusative case of the predicative noun] NEAR LEX [lemma of the support verb] within 5 words of each other. (i) the singular and plural of the predicative noun in the accusative need to be searched separately, (ii) the search returns all the documents that contain the relevant string, but these must be searched manually for relevant structures, (iii) manual correction for the voice of the verb is needed, and (iv) false positives abound due to incorrect lemmatisation.

The following timeframes are adopted: Archaic Greek (AG) pre 5th c. BC; Classical Greek (CG) 5th / 4th c. BC; Ptolemaic Greek (PG) 3rd–1st c. BC; Roman Greek (RG) 1st–3rd c. AD; Early Byzantine Greek (EBG) 4th–7th c. AD; Medieval Greek (MG) post 8th c. AD.[17] If items are e.g., 4th–3rd c. BC, they are counted in PG; if items are e.g., 7th–8th c. AD, they are counted in EBG.

3 Variability, ambiguity, discontinuity: the unbound form

The first research question (*Are support verbs just the unbound alternative of bound affixes?*) arises from the observation that the bound and unbound forms co-exist, e.g., Plato, *Republic* 456b12 ἐνομοθετοῦμεν *enomot*^{h}*etoumen* next to ἐτίθεμεν τὸν νόμον *etit*^{h}*emen ton nomon* ‘to legislate’ and BGU.1.4 (Arsinoites, AD 177–178, petition), 9–15 ἐλο[γ]οποιούμην *elo[g]opoioumēn* next to λόγον … πεποίηται *logon … pepoiētai* ‘to make a remark’.[18] The unbound form differs in its semantic and pragmatic characteristics. The differences are illustrated with λόγον / λόγους ποιέομαι *logon* / *logous poieomai* ‘to make a remark’ below but also apply to νόμον / νόμους τίθημι *nomon* / *nomous tit*^{h}*ēmi* ‘to legislate’. The section draws on Datasets 1 and 4.

The unbound support verb is part of a verbal multi-word expression that has an internal syntax, unlike the bound affix which is part of a multimorphemic word (Asraf 2021). The extent to which the internal syntax of the support-verb construction is accessible depends on its (overt and covert) analyticity and compositionality (Ledgeway & Vincent 2022: 51). Support-verb constructions display degrees of variability, discontinuity, and ambiguity of components (Savary et al. 2018: 88–90; Tutin 2016). Variability and discontinuity are specific to the unbound form.

Tables 3 and 4 show variability, as measured by the permissibility of adding determiner phrases (DP), attributive phrases (ATT), and pluralisation to the predicative noun, and discontinuity, as measured by the number and type of items intervening between the support verb and the predicative noun (the support-verb-construction field, see columns 5 and 6) (Fendel 2024)[19], along with the preferred order of support verb and predicative noun (NV, VN) (see further Section 4). In column 6, only >1 tokens are indicated in brackets.

[17] MG is included, but our main interest ends with EBG.

[18] The same is true of typologically related Latin for which the phenomenon has been studied in more depth as in Latin and Romance, for a long time (Baños 2012; Baños 2013; Marchello-Nizia 1996).

[19] Instances which show (i) co-ordination of predicative nouns and/or support verbs, (ii) morphological passivisation of the support verb with the predicative noun becoming the subject, and (iii) pronominalisation (including zero anaphora) or relativisation of the predicative noun are not counted when assessing the support-verb-construction field. Only instances of the canonical form are assessed (cf. PARSEME https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.0/?page=010_Definitions_and_scope/030_Syntactic_variants_of_VMWEs (accessed 04 Jan 2024)).

Table 3. Dataset 1. *ECF Leverhulme* Sketch Engine corpus[20]

	Total occurrences	DP	ATT	Mean field size	Field type	NV	VN
λόγος *logos* & ποιέομαι *poieomai* 'to make a remark'	13	11 85%	3 22%	0.31	PRN, PRT, DP (2)	10 77%	3 23%
λόγοι *logoi* & ποιέομαι *poieomai* 'to make remarks'	56	40 71%	7 13%	0.58	PRT (3), PP (2), DP (22), VP (2), ATT, ADV	29 53%	27[21] 47%
νόμος *nomos* & τίθημι *tit^{h}ēmi* 'to legislate'	9	6 67%	–	0.44	DP (3), PRT	5 56%	4 44%
νόμοι *nomoi* & τίθημι *tit^{h}ēmi* 'to legislate'	22	12 55%	4 18%	0.64	ATT (2), DP (7), ADV, PRT (2), indO	15 68%	7 32%

Table 4 is divided by the singular and plural forms of the predicative noun (λόγος *logos* vs λόγοι *logoi*). Each such created half is further subdivided by the period of time that the relevant instances data from, i.e. Ptolemaic Greek (PG) 3rd–1st c. BC; Roman Greek (RG) 1st–3rd c. AD; Early Byzantine Greek (EBG) 4th–7th c. AD.

[20] One instance of λόγοι *logoi* 'remarks' & ποιέομαι *poieomai* 'to do' involves co-ordination of predicative nouns; 41 instances of νόμος / νόμοι *nomos* / *nomoi* 'law(s)' & τίθημι *tit^{h}ēmi* 'to put' involve morphological passivisation (of these 9 are singular), 4 involve pronominalisation (of these 1 is singular).

[21] Aristotle, *Rhetoric* 1355a show a structure with co-ordinated predicative nouns, ποιεῖσθαι τὰς πίστεις καὶ τοὺς λόγους *poieisthai tas pisteis kai tous logous* 'to prove and argue/speak'.

Table 4. Dataset 4. *Duke Database of Documentary Papyri*

	Total occurrences	DP	ATT	Mean field size	Field type	NV	VN	High register	Low register
λόγος *logos* & ποιέομαι *poieomai* ‘to make a remark’	48	30	8	0.69		35	13	39	9
PG	28	21 75% NEG (15)	5 18%	0.61	PRN, DP (3), ATT (3), PP (5)	23 82%	5 18%	27 94%	1 4%
RG	10	3 30% NEG (1)	2 20%	0.4	PRN (2), NEG (2)	8 80%	2 20%	6 60%	4 40%
EBG	10	6 60% NEG (0)	1 10%	1.2	PP (2), DP (5), ADV, PRN	4 40%	6 60%	6 60%	4 40%
λόγοι *logoi* & ποιέομαι *poieomai* ‘to make remarks’	13	7	1	0.62		10	3	13	–
PG	5	–	1 20%	1.2	REL clause, PP	5 100%	–	5 100%	–
RG	1	–	–	0	–	1 100%	–	1 100%	–
EBG	7	7 100%	–	0.29	DP (2)	4 57%	3 43%	7 100%	–

The adding of determiner phrases, attributive phrases, and pluralisation to the predicative noun constitutes a modification of the internal morpho-syntax of the support-verb construction. Such operations are possible on analytic support-verb constructions but may be constrained even with those. E.g., *to hold in X esteem* accepts attributive phrases (*X*) but they must indicate a degree, e.g., *high*, *low*, cf. *British National Corpus*. Determiner phrases, attributive phrases, and/or pluralisation that render the predicative noun referential or break the co-referentiality between the subject of the event referred to by the

predicative noun and the subject of the support verb, break up the support-verb construction and force a verb-object reading. E.g., *Paul and Mary made a suggestion* contains a support-verb constructions, but *Tim offered their suggestion to the committee* a verb-object structure. For compositional support-verb constructions, the change in meaning attached to the adding of determiner phrases, attributive phrases, and pluralisation to the predicative noun is no greater than the change expected based on the change in form (e.g., pluralisation correlates with particularisation of meaning).

Tables 3 and 4 evidence diachronic developments, from the classical to the early Byzantine periods, for λόγος *logos* 'remark' and ποιέομαι *poieomai* 'to do'; observations for νόμος *nomos* 'law' and τίθημι *tit^{h}ēmi* 'to put' are absent from the *Duke Database of Documentary Papyri* sample:

> (i) Determiner phrases while reasonably established in the classical sample (85% of instances in the singular, 71% of instances in the plural) become fewer over time.[22] In Ptolemaic times, the availability of determiner phrases seems to have been exploited in the singular for negation (71% of all instances), thus indicating contrastive focus (Fendel 2023b) (e.g., *I made no suggestion but a comment*).[23] Negative determiner phrases do not combine with attributive phrases; 8 of the post-classical passages with a determiner phrase that is not negative show an attributive phrase.[24]
>
> (ii) The incidence of attributive phrases decreases markedly by the early Byzantine period. Attributive phrases are always limited to those that do not render the predicative noun referential and thus force a verb-object reading, e.g., λόγος *logos* as 'message' rather than 'remark' (cf. Rusten 2020). They provide internal modification as opposed to external modification by means of adverbs, e.g., *I gave a **good** speech* vs *I gave the speech **well*** where the former refers to the content of the speech and the latter to its delivery (Didakowski & Radtke 2020).
>
> (iii) The ratio of singular vs plural is reverse in the classical as opposed to the postclassical sample (19% singular vs 81% plural in classical times; 79% singular vs 21% plural in postclassical times). In the classical sample, the plural prevails and is the more flexible option; in the post-classical sample, the singular prevails and is the more flexible option. In the post-classical sample, the singular appears gradually also in the lower registers, whereas the plural remains confined to the higher registers throughout. Pluralisation as a flexibly available modification (indicating particularisation) seems to have become unavailable (Giry-Schneider 1991: 105 and 120).

---

[22] 100% of instances showing determiner phrases in early Byzantine times in the plural may reflect a fixed expression. The support-verb-construction field is maximally small (a mean of 0.29 items) and only determiner phrases can appear in it. The only remnant of flexibility pertains to the word order.

[23] E.g., P.Cair.Zen. 1.59018, 6–8 (Palestine, 258 BC, letter) **μηθένα λόγον πεποιῆσθαι** τῶι ἐπιστο [λίωι μου], αὐτοῖς δὲ *mēthena logon pepoiēst^{h}ai tō episto[liō mou], autois de* '(that they) did not make any reference to my letter, but … on them'.

[24] Unlike in classical times, e.g., Demosthenes, Speech, 18.34 ἂν ἐγὼ **λόγον** οὐδέν' **ἐποιούμην** ἕτερον *an egō logon ouden' epoioumēn heteron* 'I would not have made another remark'.

As the bound form constitutes part of a morphological word, these modulations of meaning by modifications in form are impossible (Asraf 2022).

The unbound support verb displays flexibility vis-à-vis the bound affix not only as regards the semantics of the lexical unit, as shown, but also as regards its discursive embedding. The unbound form creates a discontinuous structure with the predicative noun. The extent of discontinuity is shown in Tables 3 and 4 in the form of the size and type of the support-verb-construction field. Information-structurally closely related pieces of information can be sandwiched iconically (Lakoff & Johnson 1980: 130), as the prepositional phrases referring to recipients in (14) and (15):

(14) CG Isaeus, *On the Estate of Menecles*, 15

λόγους	οὖν	πρὸς	ἡμᾶς	**ἐποιεῖτο**
logous	*oun*	*pros*	*hēmas*	*epoieito*
word.ACC.PL.M	PRT	towards	we.ACC.PL	do.IMPF.IND.MID.3SG

καὶ **ἔφη** (…)
kai ephē (…)
and say.IMPF.IND.ACT.3SG
‘thus, he spoke to us and said (…)’

(15) PG P.Erasm.1.1 (Oxyrhynchos, 148–147 BC, petition), 22–26

τὸν	**προσήκοντα**
ton	*prosēkonta*
the.ACC.SG.M	be.suitable.PRS.PTCP.ACT.ACC.SG.M

λόγον	πρὸς	αὐτοὺς	**ποιήσασθαι**
logon	*pros*	*autous*	*poiēsasthai*
word.ACC.SG.M	towards	they.ACC.SG.M	do.AOR.INF.ACT

‘to make a suitable remark to them’

The more the support-verb construction tends towards a word and away from a syntagm, the smaller and more constrained the support-verb-construction field becomes as regards permissible parts-of-speech. This is especially obvious in the plural in the documentary sample.[25] By the early Byzantine period a mean of only 0.29 items intervene between the support verb and the predicative noun and the only permissible part-of-speech is a determiner phrase.

The unbound form makes syntactic anaphora possible, in the form of pronominalisation (including null anaphora) and relativisation. In the classical sample, ποιέομαι *poieomai* ‘to do’ is replaced by χράομαι *khraomai* ‘to use’ when more freedom regarding attributive phrases and pronominalisation is required or when larger support-verb-construction fields are needed for discursive reasons.[26] While Squeri (forthcoming) finds the latter to be an option preferred in technical contexts, no meaning or context differences can be observed in the literary classical Attic sample. In the postclassical sample, no relevant instances appear. Conversely, for νόμον/νόμους τίθημι *nomon* /

[25] Non-restrictive relative clauses have their own illocutionary force and are irrelevant for this, cf. SB.6.9225, 12 (unknown, 300–201 BC, law) (see, e.g., Koev 2022).

[26] Relevant examples appear in classical Greek, e.g., in Plato, *Gorgias* 484d; Plato, *Gorgias* 451d; Thucydides, *Histories* 4.17.2.

nomous tit^{h}ēmi ‘to legislate’ the syntagm seems preserved exactly because it allows for pronominalisation and morphological passivisation, Table 3.[27]

Note in passing that the word order is not fixed and the only combination that displays a strong preference for Noun-Verb is λόγος *logos* ‘remark’ and ποιέομαι *poieomai* ‘to do’ in the literary data. In the documentary data, there seems to be shift in preference from the Ptolemaic to the Early Byzantine periods. Word-order preferences are discussed in Section 4.

4 Cliticization, incorporation, or univerbation: from unbound to bound

The second research question (*How do semi-lexical support verbs become semi-grammatical suffixes?*) arises from the observation in Section 3 that support verbs are semi-lexical and from the debate over their relationship to auxiliary verbs outlined in Section 1. The section draws on Datasets 1 and 2.

When a diachronic link between support verbs and auxiliary verbs (e.g., Anderson 2006; Slade 2013; Ittzés 2022) is posited, a lexical item is assumed to become a grammatical item by means of grammaticalization, i.e. the conventionalisation of discursively secondary status according to Boye (2023) (cf. Section 1).[28] Grammaticalization thus defined can co-occur with ‘phonological reduction, bondedness and semantic bleaching’ (Boye 2023: 280). An extreme case is the French future tense, which has arisen from the combination of *habere* ‘to have’ and an infinitive (Hopper & Traugott 2003: 52–53; Adams 2003: 822–823), possibly through an intermediate stage of **cliticization** (Crystal 2008: 80). While such a development is regularly attested for auxiliary verbs, e.g., (ἐ)θέλω *(e)t^{h}elō* ‘to want’ vs ‘will’ (Markopoulos 2009: 85), it is not for support verbs, which maintain positional freedom and phonological shape.

Unlike auxiliaries, support verbs contribute to the event structure of the support-verb construction (Butt 2010) and contain a lexical noun[29] instead of a non-finite verb (participle, infinitive), which must be fit into a verbal frame. These observations have given rise to positing **noun incorporation** for cases such as λογοποιέω *logopoieō*. The noun “is stripped of its nominal suffixes such as case and number endings” and “cannot be qualified by determiners or adjectives” (Asraf 2021: 37–38; see also Asraf 2022; Pompei 2006; Pompei 2014; Pompei & Grandi 2012). Asraf (2021: 40) explains formations such as λογοποιέω *logopoieō* and νομοθετέω *nomot^{h}eteō* as noun incorporation with the verb ending in -έω *-eō* as a secondary denominal formation (yet Tribulato 2015: 57).[30] In his sample, the “incorporated noun can fulfil various semantic roles” including location, goal, instrument, similative, time, and patient (Asraf 2021: 54). Predicative nouns fulfil none of these adjunct roles and never the role of Patient. This is because in noun-

[27] Lexical anaphora relies on the compositionality of the structure and thus seems to cut across bound and unbound support verbs, e.g., P.Koeln.7.317 (Hermopolites, AD 501–600, letter), 26–28, 37–38, and 46–47 with λόγος *logos* ‘remark’. See further Section 4.

[28] Boye (2023: 282–289) considers the distinction between lexical and grammatical categorical, unlike the continuum often assumed (e.g., Lehmann 1988: 217).

[29] Work on syntactic nominalisations (cf. English *to give someone a beating*) is a desideratum, yet see for initial thoughts Fendel (submitted).

[30] Similarly, Lehmann (2020: 211) on German *staubsaugen* ‘to hoover’.

incorporation contexts, the verb is discursively primary, whereas in support-verb constructions the noun is (see also Fendel 2023a). This semantic make-up explains why noun-incorporated items contain nouns referring, e.g., to people, τεκνοποιέω *teknopoieō* ‘to bear a child’, which cannot be reconceptualised as eventive and are thus excluded in support-verb constructions (cf. Section 1). The semantic make-up of support-verb constructions is reflected in **anaphora patterns**, as anaphoric references point back to the semantic head.[31]

The syntagm λόγον/λόγους ποιέομαι *logon/logous poieomai* ‘to make (a) remark(s)’ first appears in Aesop’s fables (6th c. BC) (Fable 120, version 1, line 10 οὐδένα λόγον τῶν χρημάτων ποιοῦνται *oudena logon tōn kʰrēmatōn poiountai* ‘they made no mention of these matters’), the word λογοποιέομαι *logopoieomai* ‘to remark’ in Thucydides’ *Histories* (5th c. BC) (6.38, there active). The syntagm νόμον/νόμους τίθημι *nomon/nomous titʰēmi* ‘to legislate’ first appears in a fragment by Heraclitus (6th / 5th c. BC) (fragment 1, νόμους θεῖναι *nomous tʰeinai* ‘to establish legislation’)[32], the word νομοθετέω *nomotʰeteō* ‘to legislate’ in Herodotus’ *Histories* (5th c. BC) (2.42)[33]. Both syntagms and words are absent from Homer’s epics (9th / 8th c. BC).[34]

-ποιέομαι *-poieomai* ‘to do’ and -θετέω *-theteō* ‘to put’ differ. ποιέομαι *poieomai* ‘to do’ is a lexeme in Archaic and Classical Greek such that the univerbation hypothesis holds for a period for which we have extant evidence. -θετέω *-theteō* seems to align with denominal formations as suggested for a range of items by Asraf (2021). However, νομοθετέω *nomotʰeteō* ‘to legislate’ does not fit the semantic profile, i.e., its nominal component is in fact the semantic head, and the unbound form is a support verb. Thus, if Schutzeichel’s (2014: 136–138) univerbation hypothesis is correct, univerbation happened before the Archaic period.

Table 5 illustrates (i) positional freedom as evidenced by the order of predicative noun and support verb and (ii) phonological shape as evidenced by the support-verb-construction field in the classical literary Attic sample. Only >1 tokens are indicated in brackets.

[31] Not all of Asraf’s examples qualify as support-verb constructions. Rather, esp. λογοποιέω *logopoieō* ‘to remark’ and νομοθετέω *nomotʰeteō* ‘to legislate’ do not align with the rest of the sample (cf. Asraf 2021: 56–57).

[32] The instance in the *Apophthegmata* by the Septem Sapientes (division 2, apophthegm 31) is difficult to date.

[33] The instance in the *Apophthegmata* by the Septem Sapientes (division 10, apophthegm 18) is difficult to date.

[34] The lack of attestations of λόγον/λόγους ποιέομαι *logon/logous poieomai* ‘to make (a) remark(s)’ would be explicable by the lexical substitution of ἔρδω *erdō* ‘to do’ or ἔπος *epos* ‘word’, but no relevant examples appear.

Table 5. Dataset 1. *ECF Leverhulme* Sketch Engine corpus[35]

	λόγος *logos* & ποιέομαι *poieomai* 'to make a remark'	λόγοι *logoi* & ποιέομαι *poieomai* 'to make remarks'	νόμος *nomos* & τίθημι *tit^{h}ēmi* 'to legislate'	νόμοι *nomoi* & τίθημι *tit^{h}ēmi* 'to legislate'
Mean field size	0.31	0.58	0.44	0.64
Field type	PRN, PRT, DP (2)	PRT (3), PP (2), DP (22), VP (2), ATT, ADV	DP (3), PRT	ATT (2), DP (7), ADV, PRT (2), indO
NV word order	10 77%	29 53%	5 56%	15 68%
VN word order	3 23%	26 47%	4 44%	7 32%
Total occurrences	13	56	9	22

Table 5 shows that the support verbs in question retain positional freedom, as opposed to structures such as προσέχω τὸν νοῦν *prosek^{h}ō ton noun* 'to pay attention' (Fendel 2024). The only word-order preference appears with singular λόγος *logos* 'remark' and ποιέομαι *poieomai* 'to do'. For νόμος *nomos* 'law' and τίθημι *tit^{h}ēmi* 'to put', most instances display morphological passivisation or pronominalisation. Neither are considered when assessing word-order patterns (see Section 3), but both attest to positional freedom, in that components can be modified and moved independently. Table 5 also shows that the support verb seems to retain its phonological shape with support-verb-construction fields being small, but no combination showing a field that has a mean of 0.

Anaphora patterns help to determine the semantic head of a support-verb construction. (16) to (19) show anaphora patterns, both lexical and morphosyntactic.[36] Lexical anaphora can take the form of collocation, i.e., the appearance of "lexical items that often occur in the same lexical environment" (Halliday & Hasan 1976: 284–286), and reiteration, i.e., "the repetition of a lexical item [...] the use of a general word to refer back to a lexical item [...] and a number of things in between – the use of a synonym,

[35] One instance of λόγοι *logoi* 'remarks' & ποιέομαι *poieomai* 'to do' involves co-ordination of predicative nouns; 41 instances of νόμος / νόμοι *nomos / nomoi* 'law(s)' & τίθημι *tit^{h}ēmi* 'to put' involve morphological passivisation (of these 9 are singular), 4 involve pronominalisation (of these 1 is singular).

[36] The types of anaphora differ discursively. Lexical anaphora appears when "less prominent referents can only be picked up by full descriptive terms", morphosyntactic anaphora when "[s]peakers choose pronouns or zero forms for prominent discourse referents" (von Heusinger & Schumacher 2019: 123 and 125).

near-synonym, or superordinate” (Halliday & Hasan 1976: 278), (17) and (18). Morpho-syntactic anaphora can be by means of pronominalisation (using relative, personal, and demonstrative pronouns) (Manolessou 2001) or zero anaphora (Luraghi 2003a), (16) and (19).[37]

(16) Isocrates, *Speech*, 12.249

πεποίησαι	**πολλοὺς**	**λόγους,**	**τοὺς μὲν**	δικαίους
pepoiēsai	*pollous*	*logous*	*tous men*	*dikaious*
do.PRF.IND.MID.2SG	many.ACC.PL.M	word.ACC.PL.M	some.ACC.PL.M	right.ACC.PL.M

καὶ	σεμνοὺς,	**τοὺς δ’**	ἀσελγεῖς
kai	*semnous*	*tous d’*	*aselgeis*
and	just.ACC.PL.M	others.ACC.PL.M	brutal.ACC.PL.M

καὶ	λίαν	φιλαπεχθήμονας·
kai	*lian*	*philapekhthēmonas*
and	too	quarrelsome.ACC.PL.M

‘you made many remarks, some right and just, others brutal and excessively quarrelsome’

(17) Isocrates, *Speech*, 12.215

θρασέως	μὲν	οὐδὲ	πρὸς	ἓν
thraseōs	*men*	*oude*	*pros*	*ʰen*
arrogant.ADV	PRT	NEG	towards	one.ACC.SG.N

ἀντεῖπε	**τῶν**	**εἰρημένων**,	οὐδ’
anteipe	*tōn*	*eirēmenōn*	*oud’*
reply.AOR.IND.ACT.3SG	the.GEN.PL.N	say.PRF.PTCP.PASS.GEN.PL.N	NEG

αὖ	παντάπασιν	ἀπεσιώπησεν,	ἀλλ’
au	*pantapasin*	*apesiōpēsen*	*all’*
in.turn	entirely	be.silent.AOR.IND.ACT.3SG	but

ἔλεγεν,	ὅτι.	«Σὺ	μὲν
elegen	*ʰoti*	*su*	*men*
say.IMPF.IND.ACT.3SG	that	you.NOM.SG.M	PRT

πεποίησαι	**τοὺς**	**λόγους»** […]
popoiēsai	*tous*	*logous* […]
do.PRF.IND.MID.2SG	the.ACC.PL.M	word.ACC.PL.M

‘he did not reply arrogantly to anything that was said nor did he remain silent but he said: ‘You have made remarks […]’

[37] In the postclassical period, the system of demonstratives (Manolessou 2001) and the permissibility of null objects changes (Luraghi 2003a: 180; Luraghi 2004: 245; Lavidas 2015).

(18) Aristotle, *Politics*, 1274b

τοὺς	νόμους	ἔθηκεν,	ἴδιον
tous	*nomous*	*et^{h}ēken*	*idion*
the.ACC.PL.M	law.ACC.PL.M	establish.AOR.IND.ACT.3SG	unique.NOM.SG.N

δ’	**ἐν**	**τοῖς**	**νόμοις**	οὐδέν
d’	*en*	*tois*	*nomois*	*ouden*
PRT	in	the.DAT.PL.M	law.DAT.PL.M	nothing.NOM.SG.N

ἐστιν	ὅ τι	καὶ	μνείας
estin	*o ti*	*kai*	*mneias*
be.PRS.IND.ACT.3SG	which.NOM.SG.N	also	mention.GEN.SG.F

ἄξιον,	πλὴν […]
axion	*plēn* […]
worthy.NOM.SG.N	except

‘He (sc. Drakon) legislated, but in the laws there is nothing unique which is worthy of mention except […]’

(19) Isocrates, *Speech*, 12.152

με	διεξιέναι	**τοὺς**	**νόμους**
me	*diexienai*	*tous*	*nomous*
I.ACC.SG.M	go.THROUGH.PRS.INF.ACT	the.ACC.PL.M	law.ACC.PL.M

οὓς	**Λυκοῦργος**	**μὲν**	**ἔθηκε,**
hous	*Lukourgos*	*men*	*et^{h}ēke*
which.ACC.PL.M	lycurgus.NOM.SG.M	PRT	establish.AOR.IND.ACT.3SG

Σπαρτιᾶται	δ’	**αὐτοῖς**	**χρώμενοι**
Spartiatai	*d’*	*autois*	*k^{h}rōmenoi*
spartan.NOM.PL.M	PRT	they.DAT.PL.M	use.PRS.PTCP.MID.NOM.PL.M

τυγχάνουσιν.
tugk^{h}anousin
happen.to.PRS.IND.ACT.3PL

‘(that) I am going through the laws which Lycurgus established, and which the Spartans (still) happen to use’

Anaphora patterns show that the semantic head of the support-verb construction is the predicative noun. This points towards a different process to connect λογοποιέομαι *logopoieomai* ‘to remark’ with λόγον / λόγους ποιέομαι *logon* / *logous poieomai* ‘to make (a) remark(s)’ and νομοθετέω *nomot^{h}eteō* ‘to legislate’ with νόμον / νόμους τίθημι *nomon / nomous tit^{h}ēmi* ‘to legislate’, namely univerbation.[38]

Univerbation is “the diachronic unification of two or more erstwhile autonomous words in a single one, regardless of whether the resulting word is monomorphemic or internally complex” (Giomi 2023: 48). The formerly autonomous words appear in

[38] Anaphora patterns for the bound form marginally matter (cf. Asraf 2022). They appear for λογοποιέομαι *logopoieomai* ‘to remark’ (one instance) and νομοθετέω *nomot^{h}eteō* ‘to legislate’ (5 instances) in Dataset 2.

“syntagmatically adjacent” position frequently beforehand (Lehmann 2020: 206). Univerbation relies upon a synchronically valid syntactic structure rather than a word-formation pattern and downgrades a syntactic boundary to a morphological one (Berg 2020; Lehmann 2020: 228–229, 238, and 245–246). As support-verb constructions form phrasemes (Mel’čuk 2023: 74–78), their univerbation falls under phrasal univerbation which can follow onto lexicalisation and grammaticalisation (Lehmann 2020: 214–223).

λόγον ποιέομαι *logon poieomai* ‘to make a remark’ in particular exhibits symptoms typical of univerbation in the literary classical Attic data (cf. Lehmann 2020: 232–238). The small support-verb-construction field only allows for function words to appear and reflects “enforcement of continuity” of the structure. The strong preference for noun-verb (77% of all instances) points towards the “fixation of word order”. The limited availability of attributive phrases (see Section 3) reflects the “reduction of syntactic structure”. In λογοποιέομαι *logopoieomai* ‘to remark’, “the internal structure is suppressed, [and the] external structure is added as needed” (Lehmann 2020: 235), i.e., verbal inflexional endings. Compositionality is retained, yet the “loss of compositionality is neither a necessary nor a sufficient condition for univerbation” (Lehmann 2020: 236). The new lexeme is prosodically adapted: the stress appears on the penultimate syllable if the ultimate is long and on the ante-penultimate when the ultimate is short. It is also segmentally adapted, i.e., the inflexional endings on the former predicative noun disappear.[39]

Univerbation produces new lexemes rather than new grammatical forms (Giannakis 2023: 202). It is a spontaneous process at the level of *parole* (Lehmann 2020: 248). Over time, univerbation may “lead to the emergence of new rules of word formation” (Giomi 2023: 50) due to reanalysis of the structure of the univerbate (Asraf 2021: 40; Giomi 2023: 53; Tribulato 2015: 40; Giannakis 2023: 202), e.g., the Latin ablative *mente* (from *mens* ‘mind’) into the Romance adverb suffix *-ment* (Booij 2014: 173). Items such as *claramente* ‘in a clear manner’ could “function as models for new deadjectival adverbs” (Booij 2014: 173). This has been cast in the notion of leader words, i.e. “words which might have served as models for word formation processes due to their morphological transparency” (Burdy 2019: 43 with references).

Extensive work relates to the Latin support verb *facere* ‘to do’ and its relationship with the suffixes *-fico(r)* and *-facio*, e.g., Brucale and Mocciaro (2016) on *-facio* being the more modern type without vowel reduction; Galdi (2018; 2019) on *facere* as a support verb in late Latin, Marini (2005; 2014; 2018) on *-fico* vs *-ficor*. Rosén (2020: 266–267) speaks of the ‘mechanization of the conjugating element (*-ficare* vel sim.)’. Tronci (2017b: 298) considers *-facio* the native formation as opposed to *-isso*/*-izo*/*-idio* which appear primarily with Greek loans. By the time of Old French, “*-(i)fier* could never be identified with *faire* ‘to make’ and therefore qualifies as a suffix” (Rainer & Buridant 2015: 1980). Apparently, erstwhile univerbates with the support verb *facere* ‘to do’ (leader words) were reanalysed over time into stem + suffix and thus a word-formation pattern emerged. A similar analysis is proposed by Asraf (2021: 42) for his Greek noun incorporates.

[39] Phonologically, this involves loss/assimilation of a nasal before a plosive. Semantically, the singular is unmarked vs the particularising plural.

As regards word-formation, we may initially think of compounding. The Greek compounding system is largely right-headed but "tolerates a number of left-headed types" (Tribulato 2015: 46), the productivity of which may be limited (Tribulato 2015: 103). Greek compounds contain a first component "that does not correspond to a full 'word' but to a stem" and a second component "which may consist of either a stem or an independently attested word" (Tribulato 2015: 18). Formations such as λογοποιέομαι *logopoieomai* 'to remark' would fit the bill in that they are left-headed (λόγο- *logo*-), the first component is a stem rather than a full word, and the second component is an independently attested word. Section 5 shows that the development of -ποιέομαι *-poieomai* 'to do' has gone beyond compounding and seems to align rather with the development of *-fico* 'to do' and *-(i)fier* in Latin / Romance. In Latin / Romance, the bound and unbound forms of 'to do' co-exist (e.g., Giry-Schneider 1987) and we observe splitting, i.e., "a grammatical descendent is gradually differentiated from its lexical source, with which it co-exists" (Boye 2023: 285). Such an analysis also aligns with Butt and Lahiri's (2013) objection to the grammaticalization of support verbs.

5 Productivity and paradigmaticity in diachrony: the bound form

The third research question (*Why did -ποιέομαι* -poieomai *'to do' and -θετέω* -t^{h}eteō *'to put' not become productive?*) arises from the observation in Section 4 that λογοποιέομαι *logopoieomai* 'to remark' and νομοθετέω *nomot^{h}eteō* 'to legislate' are products of spontaneous univerbation but that through reanalysis a word-formation pattern may have developed. The section draws on Datasets 3 and 5.

5.1 A new word-formation pattern

The bound form of the support verb qualifies as an affix. Inflexional affixes determine the function of an item in the sentence and are applicable to any item that is not invariable (such as particles). Derivational affixes enact transitions between different parts-of-speech and apply to content words only (e.g., not to determiner phrases). Lexical affixes change the semantics of an item (e.g., the prefixed negatives alpha privative and δυσ- *dus*- (Joshi 2020), the suffixed diminutives in -ιον *-ion* (Hopper & Traugott 2003: 5)) and apply to small groups of items depending on their meaning. Thus defined, -ποιέομαι *poieomai* 'to do' acts akin to a derivational suffix in transforming an event referred to by a noun into a verb phrase **and** a lexical affix in changing the event structure (e.g., Wittenberg & Levy 2017). Section 3 showed that unbound ποιέομαι *poieomai* 'to do' is semi-lexical and semi-grammatical and subject to reverse selection by the predicative noun, such that its combinatorial freedom is limited. What about the bound form?

Table 6 shows the lemmata in the *Thesaurus Linguae Graecae* which contain the substring ποιέομαι *poieomai*, in which the first half is nominal and can be re-conceptualised as eventive, and ποιέομαι *poieomai* 'to do' is not causative.

Table 6. Dataset 5. *Thesaurus Linguae Graecae*[40]

Lemma	AG	CG	PG	RG	EBG	MG	Total	Dictionary meaning
διασκηνοποιέομαι *diaskēnopoieomai*						#1	1	Trapp: inszeniert werden ‘to be put on stage’
δυναμοποιέομαι *dunamopoieomai*					#2	11	13	– (‘to be able’)
ἐμετοποιέομαι (ἔμετος) *emetopoiomai (emetos)*		3			#2		5	Trapp: zum Erbrechen gereizt werden ‘to be made to vomit’
μισθοποιέομαι *mist^{h}opoieomai*						#1	1	LSJ: to derive rent
ὁρκοποιέομαι *^{h}orkopoieomai*						#1	1	Trapp: schwören ‘to swear’
πραγματοποιέομαι *pragmatopoieomai*						8	8	Trapp: verwirklichen ‘to realise’
σπονδοποιέομαι *spondopoieomai*			3	#6		4	13	LSJ: to pour a libation
συνειδοποιέομαι *suneidopoieomai*					2	#2	4	LSJ: to be specified together with
συνθηκοποιέομαι *sunt^{h}ēkopoieomai*				#1	#4	3	8	LSJ: = συνθήκας ποιέομαι *sunthēkas poieomai* ‘to make an agreement’
χαραποιέομαι *k^{h}arapoieomai*					#1	#1	2	Lampe: to rejoice
χρεωποιέομαι *k^{h}reōpoieomai*					#1	6	7	Lampe: to need the assistance of, need
λογοποιέω *logopoieō*		31	9	70	56	260	426	‘to compose / to remark’
λογοποιέομαι *logopoieomai*			#1	17	7	40	65	‘to remark’
Number of tokens [count λογοποιέω *logopoieō*]	0	34	12	77	68	298	n/a	
Number of types	0	2	2	3	7	11	n/a	

[40] Abbreviations used in Table 6: V = Verse, # = hapax, idiolectal. Idiolectal (#) means that the item is used repeatedly in the same author’s writings.

Table 6 allows for three observations: Firstly, -ποιέομαι *-poieomai* ‘to do’ seems suffixed to an o-stem noun with one exception, which is χαραποιέομαι *kharapoieomai* ‘to rejoice’, whereas in univerbation contexts we would expect that any stem type can appear. Secondly, the earliest attestations of -ποιέομαι *-poieomai* ‘to do’ formations appear in technical registers and the formation is limited to prose rather than verse contexts. Thirdly, there is a striking numerical preponderance of one lemma (λογοποιέομαι *logopoieomai* ‘to remark’) from the classical period onwards and no relevant lemmata appear before the classical period. Table 7 shows the results for the same search parameters in the *Duke Database of Documentary* Papyri, to which the above observations also apply:

Table 7. Dataset 3. *Duke Database of Documentary Papyri*[41]

	Total of tokens	Total of types	Mean tokens per type	Register	Lemmata
PG	4	1	4	all H	λογοποιέομαι *logopoieomai* ‘to remark’ (4)
RG	35	6	5.8	H = 33 L = 2 (λογοποιέομαι *logopoieomai* ‘to remark’)	λογοποιέομαι *logopoieomai* ‘to remark’ (20) μαρτυροποιέω/ομαι *marturopoieō/omai* ‘to testify’ (11) μετροποιέω *metropoieō* ‘to measure’ (1) στιχοποιέω *stik^{h}opoieō* ‘to compose (verses)’ (1) βλαβοποιέω *blabopoieō* ‘to harm’ (1) κακιοποιέω *kakiopoieō* ‘to wrong’ (1)
RG/EBG	1	1	1	all H	λογοποιέομαι *logopoieomai* ‘to remark’ (1)
EBG	7	5	1.4	all H	λογοποιέω/ομαι *logopoieomai* ‘to remark’ (3) μαρτυροποιέω/ομαι *marturopoieō/omai* ‘to testify’ (1) μετροποιέω *metropoieō* ‘to measure’ (1) συμφεροποιέομαι *sump^{h}eropoieomai* ‘to benefit’ (1) χρεωποιέομαι *k^{h}reōpoieomai* ‘to need’ (1)

[41] αποι *apoi* and ηποι *ēpoi* did not return relevant hits.

Regarding observation one, the apparent limitation to o-stem nouns in combination with -ποιέομαι *-poieomai* ‘to do’, two aspects are of interest: Firstly, there is one exception to this rule, χαραποιέομαι *kʰarapoieomai* ‘to rejoice’, which appears only twice in the literary data, in the 6th / 7th c. AD *Commentarius in Ecclasiaten* by Gregorius (6.16, 28–34) and in the astrological *Zodiologicum*, which is of uncertain date (e cod. Mus. Hist. Mosq. 186, fol. 144, 29–33). These instances may attest to two author’s creativity rather than a new pattern. Secondly, there are several formations for which the underlying support-verb construction contains an a-stem noun but the -ποιέομαι *-poieomai* formation does not. διασκήνη *diaskēnē* ‘stage’, σπονδή *spondē* / σπονδαί *spondai* ‘libation / truce’, συνθήκη *suntʰēkē* ‘agreement’, and χρεία *kʰreia* ‘need’ are a-stem nouns that appear in the support-verb construction but seem transformed into o-stem nouns in the -ποιέομαι *-poieomai* ‘to do’ formations. δύναμις *dunamis* ‘power’ is an s-stem seemingly transformed into an o-stem. Relevant instances also appear in the documentary data. Instead of a-stem βλάβη *blabē* ‘harm’, s-stem βλάβος *blabos* ‘harm’ seems to appear, instead of a-stem κακία *kakia* ‘badness’, the comparative adjective κακίων / κάκιον *kakiwn / kakion* ‘worse’, instead of a-stem συμφορά *sumpʰora* ‘chance’, the participle συμφέρον *sumpʰeron* ‘suitable’, instead of a-stem χρεία *kʰreia* ‘obligation’, the s-stem χρέος *kʰreos* ‘obligation’. Instead of positing a random stem change for the nominal component of the -ποιέομαι *-poieomai* ‘to do’ formations, one might posit an emerging word-formation pattern. The o-vowel would be part of the word-formation pattern and the nominal component would be a bare stem (cf. Tribulato 2015: 18).[42] From this, two questions emerge: why the o-vowel and what is the function of the new pattern?

5.2 Leader words

Observation two, the numerical preponderance of λογοποιέομαι *logopoieomai* ‘to remark’ from the classical period onwards without relevant lemmata before the classical period, links to the notion of leader words introduced in Section 4. Table 8 shows the lemmata in the *Thesaurus Linguae Graecae* which contain the substring -θετέω *-tʰeteō* ‘to put’, in which the first half is nominal and can be reconceptualised as eventive, and -θετέω *-tʰeteō* ‘to put’ is not causative.

Table 8. Dataset 5. *Thesaurus Linguae Graecae*[43]

Lemma	AG	CG	PG	RG	EBG	MG	Total	Dictionary meanings
διαγωνοθετέω *diagōnotʰeteō*[44]			2			#2	4	LSJ: to set at variance
δικοθετέω *dikotʰeteō*						#1 (V)	1	Trapp: Recht sprechen ‘to judge’
δογματοθετέω *dogmatotʰeteō*						#1	1	Trapp: ein Dogma darlegen ‘to lay out a dogma’

[42] This is inexplicable under a noun-incorporation approach except if assuming noun incorporation as the first step and subsequent reanalysis (Giannakis 2023: 202).

[43] Abbreviations used in Table 8: V = Verse, # = hapax, idiolectal.

[44] διαγωνία *diagōnia* ‘struggle’ is a cranberry word.

θημονοθετέω *tʰēmonotʰeteō*						#1	1	LSJ: to put in a heap
καιροθετέω *kairotʰeteō*						#1	1	Trapp: die rechte Zeit festsetzen 'to determine the right time'
λογοθετέω *logotʰeteō*			#1		74	126 (V)	201	LSJ: to call to account, to keep accounts
ὀνοματοθετέω *onomatotʰeteō*					27	41	68	LSJ: to give a name
ὁροθετέω *ʰorotʰeteō*				3	13	100 (V)	116	LSJ: to fix boundaries
συμψηφοθετέω *sumpsēpʰotʰeteō*				#1			1	LSJ: = συμψηφίζω *sumpsēphizō*, to reckon together, to count up
σφαιροθετέω *sphairotʰeteō*						#1	1	Trapp: die himmlische Sphäre festlegen 'to determine the heavenly sphere'
φοροθετέω *pʰorotʰeteō*				2			2	Lampe: to make a charge for
ὡροθετέω *ʰōrotʰeteō*					2	21	23	LSJ: to take note of a thing in casting a nativity
νομοθετέω *nomotʰeteō*	#1	361	141 (V)	573 (V)	2550 (V)	3285 (V)	6,911	LSJ: to legislate
ἀντινομοθετέω *antinomotʰeteō*				6	44	132 (V?)	182	
διανομοθετέω *dianomotʰeteō*		#2		#10	2	3	17	
ἐπινομοθετέω *epinomotʰeteō*		#1			#3		4	
κατανομοθετέω *katanomotʰeteō*		#1					1	
παρανομοθετέω *paranomotʰeteō*						#1	1	

προνομοθετέω *pronomot^{h}eteō*				#1	2	7	10	
προσεπινομοθετέω *prosepinomot^{h}eteō*					#1		1	
προσνομοθετέω *prosnomot^{h}eteō*				10	6	3	19	
συννομοθετέω *sunnomot^{h}eteō*		3			#1	3	7	
νουθετέω *nout^{h}eteō*[45]	#4	1 2 8 (V)	8 3 (V)	3 8 3 (V?)	1009	1 4 4 9 (V)	3,053	LSJ: to put in mind
ἀντινουθετέω *antinout^{h}eteō*				#1			1	
ἀπονουθετέω *aponout^{h}eteō*						#1	1	
ἐπινουθετέω *epinout^{h}eteō*						#1	1	
κατανουθετέω *katanout^{h}eteō*					#1	6	7	
προνουθετέω *pronout^{h}eteō*						3	3	
ὑπερνουθετέω *^{h}upernout^{h}eteō*						#1	1	
ὑπονουθετέω *^{h}uponout^{h}eteō*				#1			1	
Number of tokens [prefixed options are counted under root]	5	489	227	962	3675	5029	n/a	
Number of types [prefixed options are counted under root]	2	2	4	5	6	12	n/a	

Table 8 reflects the same seeming preference for o-stem nouns with -θετέω *-t^{h}eteō* ‘to put’ as observed for -ποιέομαι *-poieomai* ‘to do’ (esp. διαγωνοθετέω *diagōnot^{h}eteō* ‘to set at variance’ but διαγωνία *diagōnia* ‘struggle’ and δικοθετέω *dikot^{h}eteō* ‘to judge’

[45] νουθετ-εύω *nout^{h}et-euō* in the 14th c. AD.

but δίκη *dike* ‘judgement’). It also reflects a numerical preponderance of two lexemes, νουθετέω *nout*^{h}*eteō* ‘to put in mind’ and νομοθετέω *nomot*^{h}*eteō* ‘to legislate’, similarly to λογοποιέομαι *logopoieomai* ‘to remark’. However, those two lexemes appear already in archaic Greek.

λογοποιέομαι *logopoieomai* ‘to remark’ appears in the passive from the 2/3 c. AD onwards in the *Thesaurus Linguae Graecae*. Passivisation applies externally to the word rather than taking the internal syntagmatic structure into account. Only future and aorist passive formations are counted as in all other tenses, the middle and passive share the same set of endings. νομοθετέω *nomot*^{h}*eteō* ‘to legislate’ and νουθετέω *nout*^{h}*eteō* ‘to put in mind’ exist in the passive from classical times onwards reflecting their earlier lexicalisation and becoming a word (Mel’čuk 2023: 72; Langer 2004: 177–178). From classical times, both lexemes can be modified by lexical prefixes (preverbs) (cf. Luraghi 2003b) further pointing towards their status as a word rather than a syntagm.

All three leader words, i.e., νουθετέω *nout*^{h}*eteō* ‘to put in mind’, νομοθετέω *nomot*^{h}*eteō* ‘to legislate’, and λογοποιέομαι *logopoieomai* ‘to remark’, contain an o-stem noun before -θετέω -*t*^{h}*eteō* ‘to put’ or -ποιέομαι -*poieomai* ‘to do’.[46] They are morphologically transparent and semantically compositional, at least until passive formations and/or preverbs start appearing. Thus, the o-vowel in the hypothesised word-formation pattern may result from re-segmentation of the univerbates of support-verb constructions and subsequent reanalysis of the former support verb as a suffix.

5.3 Productivity

If we posit -ποιέομαι -*poieomai* ‘to do’, and possibly -θετέω -*t*^{h}*eteō* ‘to put’, as a word-formation pattern, the question of productivity arises. Productivity means that an item “is repeatedly used in language to produce further instances of the same type” such as the inflexional past-tense suffix -*ed* in English (Crystal 2008: 390).

As -ποιέομαι -*poieomai* ‘to do’ and -θετέω -*t*^{h}*eteō* ‘to put’ are not purely derivational but have a lexical element to them, productivity is not expected to be high. With less productive items, we distinguish between available patterns and patterns that language users generally accept. E.g., German *öffbar* ‘can be opened’ results from the availability of the derivational pattern for verbs in -*nen* such as *öffnen* ‘to open’ (Finkbeiner 2008: 401) and English *readable* relies on the pattern of deverbal adjective formations in -*able*. Neither is universally accepted. German prefers the formation with -*en*-, e.g., *einordnen* ‘to put in order’ with *einordbar/einordenbar*, English prefers *legible*.[47] Patterns that are theoretically available but not generally accepted would qualify as creativity rather than productivity (Goldberg 2019; Hoffmann 2018). These patterns appear in the corpus data as *hapaces* or idiolectal attestations (cf. Baayen 2009). Patterns that are generally accepted appear with relatively high token-type ratios. However, the text type may impose limitations in a corpus language such as (post-)classical Greek (Hoffmann 2005: chap. 8). Productivity can increase and decrease diachronically (Barðdal et al. 2024; Hartmann 2018).

Table 9 shows the lemmata in the *Duke Database of Documentary Papyri* which contain the substring οθετ *ot*^{h}*et*, in which the first half is nominal and can be reconceptualised as eventive, and -θετέω -*t*^{h}*eteō* ‘to put’ is not causative. Searches for ηθετ *ēt*^{h}*et* and αθετ *at*^{h}*et* did not return relevant hits.

---

[46] νόος *noos* regularly contracts to νοῦς *nous*.

[47] The German and English suffixes add modality.

Table 9. Dataset 3. *Duke Database of Documentary Papyri*

	Total of tokens	Total of types	Mean tokens per type	Register	Lemmata
PG	1	1	1	all H	ὁριοθετέομαι *^{h}oriot^{h}eteō* ‘to set boundaries / divide’
RG	15	3	5	all H	νομοθετέω/ομαι *nomot^{h}eteō/omai* ‘to legislate’ (6) λογοθετέω/ομαι *logot^{h}eteō/omai* ‘to call to account’ (4) ἀγωνοθετέω *agōnot^{h}eteō* ‘to exhibit games’ (5)
EBG	5	2	2.5	all H	λογοθετέω/ομαι *logot^{h}eteō/omai* ‘to call to account’ (3) νομοθετέω/ομαι *nomot^{h}eteō/omai* ‘to legislate’ (2)

Compared to Table 7 for -ποιέομαι *-poieomai* ‘to do’ formations, the same o-vowel before the suffix and the smaller range of types is noticeable. Table 7 showed 1 type for the Ptolemaic period, 6 types for the Roman period, and 5 types for the early Byzantine period. Table 10 synthesises counts of types that are not *hapaces*, idiolectal, or leader words (and their compounds) for -ποιέομαι *-poieomai* ‘to do’ and -θετέω *-t^{h}eteō* ‘to put’ formations based on Tables 6 to 9. *Hapaces* and idiolectal items are excluded as they reflect creativity rather than productivity; leader words are excluded as they are the origin of reanalysis and do not constitute new types.

Table 10. Counts without *hapaces*, idiolectal items, and (compounds of) leader words

Number of types	**AG**	**CG**	**PG**	**RG**	**EBG**	**MG**
-ποιέομαι *-poieomai* ‘to do’ literary	–	1[48]	1	–	1	5
-ποιέομαι *-poieomai* ‘to do’ documentary	–	–	–	1[49]	–	–
-θετέω *-t^{h}eteō* ‘to put’ literary	–	–	1	2	4	4
-θετέω *-t^{h}eteō* ‘to put’ documentary	–	–	–	2	1	–

Productivity of the hypothesised word-formation patterns seems relatively low throughout although with a slight increase over time in the literary sources. Productivity of the hypothesised word-formation patterns is more limited in the documentary than the literary data.

[48] ἐμετοποιέομαι *emetopoieomai* ‘to be made to vomit’: Hippocrates Med., *De affectionibus*; Diocles Med., *Fragmenta* (2x).

[49] μαρτυροποιέω/ομαι *marturopoieō/omai* ‘to testify’.

5.4 Technical contexts and idiolects

This brings us to observation two, i.e., that the earliest attestations of -ποιέομαι *-poieomai* 'to do' formations appear in technical registers, primarily medical writing in literary contexts and witness statements in documentary contexts, and the formation seems limited to prose contexts. This applies to -θετέω *-tʰeteō* 'to put' formations less strictly, in that the leader words appear in verse contexts from classical times and some non-leader formations in later periods. Both formation patterns were apparently exploited for creative purposes. This is productivity in the sense of Barðdal's extensibility, Baayen's potential productivity, and Sampson's E(nlarging)-creativity.[50] It points towards the function of this new word-formation pattern. Table 11 shows the number and context of *hapaces* and idiolectal formations extracted from Tables 6 to 9.

Table 11. *hapaces* (without compounds of leader words) and idiolectal items[51]

Types	PG	RG	EBG	MG
-ποιέομαι *-poieomai* 'to do'	–	2	5	5
Register (literary)	n/a	Apollonius, *Lexicon Homericum*; Athenaus, *Deipno-sophistae*	Pseudo-Dionysios Areopagita, *De divinis nominibus* (2x); Paulus Med., *Epitomae medicae libri septem* (2x); Hesychius, *Lexicon* (4x); Gregorius, *Commentarius in Ecclesiasten*; Ephraem, *Sermo compunctorius*	Photius, *Epistulae et Amphilochia*; Achmet Astrol., *Achmetis Oneirocriticon*; Michael Psellus, *Epistulae*; Manuel Bryennius Mus / Math / Astron, *Harmonica* (2x); Astrologica, *Zodiologium*
-ποιέομαι *-poieomai* 'to do'	–	4	4	–
Register (documentary)	n/a	list (2x); petition; taxes	contract (2x); order; protocol	n/a

[50] Barðdal's (2008: 24–25) 'extensibility of a word formation pattern' refers to the 'degree, [to which] the word formation patterns of a given language are available when new words come into existence' and is 'a function of a construction's type frequency'. Baayen's (2009: 902) potential productivity 'estimates the growth rate of the vocabulary of the morphological category' based on the hapax-token ratio. Sampson's (2016: 19) E(nlarging) creativity is 'creativity that enlarges or expands our system(s)' by breaking the existing system rules intentionally or unintentionally (see also Bergs 2019).

[51] No relevant hits for AG and CG.

-θετέω *-theteō* ‘to put’	1	1	–	6
Register (literary)	Antio-chus Astrol., Fragmenta (1x).	Vettius Valens, *Anthologiarum libri xi*	n/a	Constantinus VII Porphyrogenitus (2x); Constantinus Manasses Poeta, *Breviarium Chronicum* (1x); Michael Syncellus Gramm, *Laudatio sancti Mocii* (1x); Scholia in Theocritum; Cosmas Vestitor, *Vita Joannis Chrysostomi* (1x); Scholia in Aratum
-θετέω *-theteō* ‘to put’	1	–	–	–
Register (documentary)	petition	n/a	n/a	n/a

-ποιέομαι *-poieomai* ‘to do’ formations in the literary data appear primarily in technical writing – medical, astrological, and commentary / lexicon projects – and in the documentary data in technical documents – contracts, protocols, and tax documents. In the literary and documentary sources, -ποιέομαι *-poieomai* ‘to do’ formations appear to be exploited as a creative means only from the Roman period onwards. This aligns with the observation that the leader word had by this point become a word and reanalysis must have happened before. A noticeable example is συνθηκοποιέομαι *sunthēkopoieomai* ‘to make an agreement’, all eight instances of which come from lexica and scholia, apparently a formation created for technical writing.

-θετέω *-theteō* ‘to put’ formations in the literary data appear from the Ptolemaic period onwards, initially primarily in technical writing – astrological and anthological writings. Yet, they do not seem to be exploited for creative purposes before the medieval period. The documentary data is very limited. A noticeable instance is Hesychius’ *Lexicon* (5th / 6th c. AD) (640) νομοθετεῖ : νομοποιεῖ *nomothetei : nomopoiei* ‘to legislate’, where the (newer) more transparent formation explains the apparently no longer fully transparent formation. Both are causative.

The above shows, akin to Baayen (2009: 908), that the suffixes in question are most common in technical writing in the sense of a creative exploitation of a theoretically available pattern, and that -ποιέομαι *-poieomai* ‘to do’ formations seem to emerge productively in technical contexts (cf. Van Camp 2005; Squeri forthcoming; Schutzeichel 2014: 136–138). Yet, Greek has extensive derivational morphology to transform verbs into nouns and vice versa (van Emde Boas et al. 2019: 262–269), such that a desire to

integrate nominal technical terms in the predicate slot cannot explain the emergence of a new word-formation pattern.

Rather, neither the bound nor the unbound forms of ποιέομαι *poieomai* 'to do' and τίθημι *tit^{h}ēmi* / -θετέω *-t^{h}eteō* 'to put' are fully grammatical or fully lexical. ποιέομαι *poieomai* 'to do' in particular is a hybrid element, in that it converts nouns into verbs but also interferes with the event structure of the resulting verb phrase (in profiling the subject component), and seems specialised for technical writing.[52] The resulting verb phrase can take direct objects (cf. Fendel 2023a) unlike an antipassive (cf. Marini 2010; Creissels 2016; Asraf 2021), which could be considered a categorial periphrastic (Haspelmath 2000).[53]

ποιέομαι *poieomai* 'to do' seems to exist in a bound and an unbound form from the classical period onwards (after lexical renewal struck down ἔρδω *erdō* 'to do') and until lexical renewal strikes it down (cf. modern Greek κάνω *kanō* 'to do' (Anastassiadis-Symeonidis, Fotopoulou & Kyriacopoulou 2019)). Possibly due to its shorter lifespan, it does not reach the degree of productivity of Latin *-facio*/*-fico* 'to do' (cf. Section 4). -θετέω *-t^{h}eteō* 'to put' seems to be the older counterpart that saw a brief revival in the early medieval period (cf. Schutzeichel 2014: 136–138 on early specialization).[54]

6 Summary, conclusion, outlook

The support verbs ποιέομαι *poieomai* 'to do' and τίθημι *tit^{h}ēmi* 'to put' exist in bound and unbound forms from classical into medieval times, ποιέομαι *poieomai* 'to do' as ποιέομαι *poieomai* and -ποιέομαι *-poieomai*, τίθημι *tit^{h}ēmi* 'to put' as τίθημι *tit^{h}ēmi* and -θετέω *-t^{h}eteō*. They differ from auxiliaries, in that they are semi-lexical as they contribute to the event structure of the verb phrase. The article used five data samples drawn from the literary corpus of the *Thesaurus Linguae Graecae* and the documentary corpus of the *Duke Database of Documentary Papyri* (cf. Section 2) to answer three research questions: (i) Are support verbs the unbound alternative of bound affixes? (ii) How do semi-lexical support verbs become semi-grammatical affixes? (iii) Why did -ποιέομαι *-poieomai* 'to do' and -θετέω - *t^{h}eteō* 'to put' not become productive?

Section 3 found by considering the variability and discontinuity of support-verb constructions with ποιέομαι *poieomai* 'to do' and τίθημι *tit^{h}ēmi* 'to put' (Datasets 1 and 4) along with their anaphora patterns that the bound and unbound forms differ in the semantics of the lexical unit and its discursive embedding. When applying Boye's (2023) criteria for distinguishing between lexical and grammatical items, i.e., the permissibility of being (i) focussed, (ii) addressed in subsequent discourse, (iii) modified, and (iv) of standing alone in an utterance, support verbs appear to be semi-lexical. Section 4

[52] It seems to index a technical context (cf. Bentein 2019 on indexing in Greek documentary data).

[53] Categorial periphrastics have "a sufficiently high degree of grammaticalization to be described as part of the verbal paradigm", yet they do not equal forms in the paradigm, but are add-ons, such as the French *aller*-future (Haspelmath 2000: 664).

[54] An anonymous reviewer remarked that "in Modern Greek the verb ποιέω [*poieō*] has some very common bound forms which do not seem to appear in earlier Greek, such as χρησιμοποιώ [*k^{h}rēsimopoiō*] 'to use'". The remark is interesting because ποιέω *poieō* 'to do' has yielded as the main doing verb to κάνω *kanō* 'to do' in Modern Greek. If there are indeed bound forms that have no pre-modern ancestor, this would lend support to the hypothesis of a word-formation pattern. However, one should caution that in the example provided, χρήσιμος *k^{h}rēsimos* 'useful' is an adjective. Further investigation of syntactic nominalisations in the predicative-noun slot of the support-verb construction and the slot before the affix in the bound form would be needed to draw further conclusions.

rejected the grammaticalization and cliticization hypotheses for support verbs turned affixes based on their positional freedom and morphological shape (Datasets 1 and 2) and the noun-incorporation hypothesis based on anaphora phenomena. It showed that formations such as λογοποιέομαι *logopoieomai* ‘to remark’ result from univerbation. The co-existence of homonymous support and auxiliary verbs is explained by splitting. Section 5 argued for the emergence of a new word-formation pattern due to reanalysis of leader words (λογοποιέομαι *logopoieomai* ‘to remark’, νουθετέω *nout^{h}eteō* ‘to put in mind’, and νομοθετέω *nomot^{h}eteō* ‘to legislate’). Re-segmentation in the process explains the o-vowel in -ποιέομαι *-poieomai* ‘to do’ and -θετέω *-t^{h}eteō* ‘to put’ formations. The new suffixes are semi-lexical like the unbound support verb, largely limited to prose contexts, and more common in literary than documentary texts. Creative formations, i.e. *hapaces* and idiolectal items, attest to their availability for creative purposes, especially in technical registers, from the Ptolemaic/Roman period onwards, while productivity remains low throughout.

In Greek, ποιέομαι *poieomai* ‘to do’ is an outlier. Diachronically, the support verb (bound and unbound) replaces from the classical period earlier ἔρδω *erdō* ‘to do’ but yields to κάνω *kanō* ‘to do’ by the modern period. Synchronically, it combines with so large a range of nouns that reverse selection due to lexical collocation seems lower than with other support verbs. Perhaps therefore, ποιέομαι *poieomai* ‘to do’ can be replaced by verbs of realization, including πράττω *prattō* ‘to achieve’ and ἐργάζομαι *ergazomai* ‘to work on/at’ (De Pasquale 2023: 263; Baños & Jiménez López forthcoming). ποιέομαι *poieomai* ‘to do’ is the only support verb with form-identical bound and unbound forms. Conversely, ‘to do’ exists as a support verb across typologically unrelated languages, e.g., Persian (Saeedi 2017) and Coptic (Reintges 2001; Grossman 2023), and its bound form constitutes a semi-lexical affix (see also Croft 2022: 397–431).

Acknowledgements

The article was written in the context of the project *Giving gifts and doing favours: Unlocking Greek support-verb constructions* (University of Oxford, 2020–2024), which was funded by the *Leverhulme Trust* (grant n. ECF-2020-181) with the author as the PI.

Abbreviations

AG	Archaic Greek (pre 5th c. BC)
AOR	Greek aorist tense
ATT	attributive phrase
CG	Classical Greek (5th / 4th c. BC)
CONJ	conjunction
DP	determiner phrase
EGB	Early Byzantine Greek (4th–7th c. AD)
H	high register
IMPF	Greek imperfect tense
indO	indirect object
L	low register
Lampe	Lampe, Geoffrey, *A Patristic Greek Lexicon* 1961 (accessible via http://stephanus.tlg.uci.edu/lsj/)
LSJ	Liddell, Henry, and Robert Scott and Henry Jones, *Greek-English Lexicon* 1996 (accessible via http://stephanus.tlg.uci.edu/lsj/)

LVC	light-verb construction (PARSEME)
MG	Medieval Greek (post 8th c. AD)
MID	middle voice
NV	noun-verb order
PG	Ptolemaic Greek (3rd–1st c. BC)
PN	predicative noun
PP	prepositional phrase
PR	present tense
PRN	pronoun
PRT	particle
RG	Roman Greek (1st–3rd c. AD)
SV	support verb
SVC	support-verb construction
Trapp	Trapp, Erich, and Wolfram Hörandner, *Lexikon zur byzantinischen Gräzität: besonders des 9.–12. Jahrhunderts* 2001 (accessible via https://stephanus.tlg.uci.edu/lbg/#context=lsj&eid=17288)
VN	verb-noun order
VP	verb phrase

References

Adams, James. 2003. *Bilingualism and the Latin language*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511482960

Anastassiadis-Symeonidis, Anna & Fotopoulou, Angeliki & Kyriacopoulou, Tita. 2019. Multiword expressions in Modern Greek: Synthetic review on their nature. *Bulletin of Scientific Terminology and Neologisms 2019, Special Issue: MWEs in Greek and other languages: From theory to implementation.* https://hal.science/hal-02378693

Anderson, Gregory. 2006. *Auxiliary verb constructions.* Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199280315.001.0001

Asraf, Nadav. 2021. The mechanism of noun incorporation in Ancient Greek. *Glotta* 97. 36–72. https://doi.org/10.13109/glot.2021.97.1.36

Asraf, Nadav. 2022. The Syntax-Morphology Interface in Ancient Greek. *Mnemosyne* 76(5). 1–30. https://doi.org/10.1163/1568525x-bja10141

Baayen, R. Harald. 2009. Corpus linguistics in morphology: Morphological productivity. In Kytö, Merja & Lüdeling, Anke (eds.), *Corpus linguistics: An international handbook*, 899–919. Berlin; Boston: Mouton De Gruyter. https://doi.org/10.1515/9783110213881.2.899

Bakker, Peter. 2003. Mixed languages as autonomous systems. In Bakker, Peter & Matras, Yaron (eds.), *The mixed language debate: Theoretical and empirical advances*, 113–156. Berlin: Mouton de Gruyter. https://doi.org/10.1515/9783110197242.107

Baños, José Miguel. 2012. Verbos soporte e incorporación sintáctica en latín: El ejemplo de ludos facere. *Revista de Estudios Latinos (RELat)* 12. 37–57. https://doi.org/10.23808/rel.v12i0.87788

Baños, José Miguel. 2013. Sobre las maneras de "hacer la guerra" en latín: Bellum gero, belligero, bello. In Beltrán Cebollado, José Antonio & Encuentra Ortega, Alfredo & Fontana Elboj, Gonzalo & Magallón García, Ana Isabel & Marina Sáez, Rosa María & Iso Echegoyen, José Javier (eds.), *Otium cum dignitate: Estudios en homenaje al profesor José Javier Iso Echegoyen*, 27–40. Zaragoza: Universidad de Zaragoza.

Baños, José Miguel & Jiménez López, María Dolores. forthcoming. Translation as a mechanism for the creation of collocations (I): The alternation ἐργάζομαι / ποιέω in the Bible.

Barðdal, Jóhanna. 2008. *Productivity: Evidence from case and argument structure in Icelandic.* (Constructional Approaches to Language 8). Amsterdam: John Benjamins. https://doi.org/10.1075/cal.8

Barðdal, Jóhanna & Enghels, Renata & Feltgen, Quentin & Van Hulle, Sven & Lauwers, Peter. 2024. Productivity in diachrony. In Ledgeway, Adam & Breitbarth, Anne & Kiss, Katalin & Salmons, Joseph & Simonenko, Alexandra (eds.), *Wiley Blackwell Companion to diachronic linguistics*. Chichester: Wiley Blackwell. https://doi.org/10.13140/RG.2.2.21677.56804

Bentein, Klaas. 2016. *Verbal periphrasis in ancient Greek: have- and be- constructions*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198747093.001.0001
Bentein, Klaas. 2017. Finite vs. non-finite complementation in Post-classical and Early Byzantine Greek: Towards a pragmatic restructuring of the complementation system? *Journal of Greek Linguistics* 17(1). 3–36. https://doi.org/10.1163/15699846-01701002
Bentein, Klaas. 2019. Dimensions of social meaning in Post-classical Greek: Towards an integrated approach. *Journal of Greek Linguistics* 19(2). 119–167. https://doi.org/10.1163/15699846-01902006
Berg, Kristian. 2020. Changes in the productivity of word-formation patterns: Some methodological remarks. *Linguistics* 58(4). 1117–1150. https://doi.org/10.1515/ling-2020-0148
Bergs, Alexander. 2019. What, If Anything, Is Linguistic Creativity? *Gestalt Theory* 41. 173–183. https://doi.org/10.2478/gth-2019-0017
Bonami, Olivier. 2015. Periphrasis as collocation. *Morphology* 25. 63–110. https://doi.org/10.1007/s11525-015-9254-3
Booij, Geert. 2014. The structure of words. In Taylor, John (ed.), *The Oxford handbook of the word*, 157–174. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199641604.013.002
Boye, Kasper. 2023. Grammaticalization as Conventionalization of Discursively Secondary Status: Deconstructing the Lexical–Grammatical Continuum. *Transactions of the Philological Society* 121(2). 270–292. https://doi.org/10.1111/1467-968X.12265
Brown, Dunstan & Chumakina, Marina & Corbett, Greville & Popova, Gergana & Spencer, Andrew. 2012. Defining “periphrasis”: Key notions. *Morphology* 22. 233–275. https://doi.org/10.1007/s11525-012-9201-5
Brucale, Luisa & Mocciaro, Egle. 2016. Composizione verbale in latino: Il caso dei verbi in -facio, -fico. In Poccetti, Paolo (ed.), *Latinitatis rationes: Descriptive and historical accounts for the Latin language*, 279–297. Berlin; Boston: Mouton De Gruyter. https://doi.org/10.1515/9783110431896-020
Burdy, Philipp. 2019. On the importance of leader words in word formation: The popular transmission of the Latin abstract-forming suffix -io in French. *Word Structure* 12(1). 42–59.
Butt, Miriam. 1995. *The structure of complex predicates in Urdu*. Stanford: CSLI Publications.
Butt, Miriam. 2010. The Light Verb Jungle: Still Hacking Away. In Amberger, Mengistu & Baker, Brett & Harvey, Mark (eds.), *Complex Predicates: Cross-Linguistic Perspectives on Event Structure*, 48–78. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511712234.004
Butt, Miriam & Geuder, Wilhelm. 2001. On the (semi)lexical status of light verbs. In van Riemsdijk, Henk & Corver, Norbert (eds.), *Semi-lexical categories: The function of content words and the content of function words*, 323–370. Berlin; Boston: Mouton De Gruyter. https://doi.org/10.1515/9783110874006.323
Butt, Miriam & Lahiri, Aditi. 2013. Diachronic pertinacity of light verbs. *Lingua* 135. 7–29. https://doi.org/10.1016/j.lingua.2012.11.006
Cock, A. 1981. ΠΟΙΕΙΣΘΑΙ: Ποιein. Sur les critères déterminant le choix entre l’actif Ποιein et le moyen ΠΟΙΕΙΣΘΑΙ. *Mnemosyne* 34(1/2). 1–62. https://www.jstor.org/stable/4431012
Collins, Chris. 2018. *NEG NEG. *Glossa: A journal of general linguistics* 3(1). https://doi.org/10.5334/gjgl.611
Concu, Valentina. 2022. Werden and periphrases with present participles and infinitives: A diachronic corpus analysis. *Journal of Germanic Linguistics* 34(1). 1–34. https://doi.org/10.1017/S1470542721000064
Creissels, Denis. 2016. Univerbation of light verb compounds and the obligatory coding principle. In Nash, Léa & Samvelian, Pollet (eds.), *Approaches to complex predicates*, 46–69. Leiden; Boston: Brill. https://doi.org/10.1163/9789004307094_004
Croft, William. 2022. *Morphosyntax: Constructions of the world’s languages*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781316145289
Crystal, David. 2008. *A dictionary of linguistics and phonetics*. Malden, MA; Oxford: Blackwell. 10.1002/9781444302776
Dal Corso, Elia. 2022. The development of analytic negatives in Sakhalin Ainu. *Japanese/Korean Linguistics* 28. 211–228.
De Knop, Sabine & Hermann, Manon (eds.). 2020. *Funktionsverbgefüge im Fokus: Theoretische, didaktische und kontrastive Perspektiven.* Berlin; Boston: Mouton De Gruyter. https://doi.org/10.1515/9783110697353
De Pasquale, Noemi. 2023. Making a move towards Ancient Greek light verb constructions. In Pompei, Anna & Mereu, Lunella & Piunno, Valentina (eds.), *Light verb constructions as complex verbs*, 257–274. Berlin; Boston: Mouton De Gruyter. https://doi.org/10.1515/9783110747997-010
Didakowski, Jörg & Radtke, Nadja. 2020. Verwendung der deutschen Stützverbgefüge mit Adjektiven und ihre Ermittlung mithilfe des DWDS-Wortprofils. In De Knop, Sabine & Hermann, Manon (eds.), *Funktionsverbgefüge im Fokus*, 101–136. Berlin; Boston: Mouton De Gruyter. https://doi.org/10.1515/9783110697353-005

Egedi, Barbara. 2017. Remarks on Loan Verb Integration into Coptic. In Grossman, Eitan & Dils, Peter & Richter, Tonio & Schenkel, Wolfgang (eds.), *Greek influence on Egyptian-Coptic: Contact-induced change in an ancient African language. DDGLC Working Papers 1*, 195–206. Hamburg: Widmaier.

Ellegård, Alvar. 1953. *The Auxiliary do, the establishment and regulation of its use in English*. Stockholm: Almqvist och Wiksell.

Emde Boas, Evert van & Rijksbaron, Albert & Huitink, Luuk & de Bakker, Mathieu. 2019. *Cambridge grammar of classical Greek*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781139027052

Fendel, Victoria. 2023a. Support-verb constructions with objects: Greek-Coptic interference in the documentary papyri? In Meyer, Robin & Bianchoni, Michele (eds.), *Contact-induced change in morphosyntax*, 382–403. Chichester: Wiley-Blackwell. https://doi.org/10.1111/1467-968X.12279

Fendel, Victoria. 2023b. 'I haven't got a clue!': Assessing negation in classical Greek Support-Verb Constructions. *Journal of Greek Linguistics* 23(2). 139–163. https://doi.org/10.1163/15699846-02302004

Fendel, Victoria. 2024. Celebrating diversity: The origins and pathways of three support-verb constructions. *Lexis 24*.

Fendel, Victoria. 2025. Taking stock of Greek support-verb constructions: Synchronic and diachronic variability in the documentary papyri. In de la Villa, Jesus (ed.), *Advances in Ancient Greek Linguistics. Proceedings of the 10th International Colloquium of Ancient Greek Linguistics (ICAGL)*. Berlin; Boston: Mouton De Gruyter.

Fendel, Victoria. submitted. When the lines get blurred: Support-verb constructions in the documentary papyri. *Pylon*.

Fendel, Victoria & Ireland, Matthew. 2023. Discourse cohesion in Xenophon's On Horsemanship through Sketch Engine. *Digital Humanities Quarterly* 17(3). https://www.digitalhumanities.org/dhq/vol/17/3/000683/000683.html

Finkbeiner, Rita. 2008. Zur Produktivität idiomatischer Konstruktionsmuster. Interpretierbarkeit und Produzierbarkeit idiomatischer Sätze im Test. *Linguistische Berichte* 216. 391–430.

Fleischman, Suzanne. 2000. Methodologies and ideologies in historical linguistics: On working with older languages. In Herring, Susan & Reenen, Pieter & Schøsler, Lene (eds.), *Textual parameters in older languages*, 33–58. Amsterdam: John Benjamins. https://doi.org/10.1075/cilt.195.03fle

Funk, Wolf-Peter. 2017. Differential loan across the Coptic literary dialects. In Grossman, Eitan & Dils, Peter & Richter, Tonio & Schenkel, Wolfgang (eds.), *Greek influence on Egyptian-Coptic: Contact-induced change in an ancient African language. DDGLC Working Papers 1*, 369–397. Hamburg: Widmaier.

Galdi, Giovanbattista. 2018. On the use of facio as support verb in late and Merovingian Latin. *Journal of Latin Linguistics* 17(2). 231–257. https://doi.org/10.1515/joll-2018-0011

Galdi, Giovanbattista. 2019. Verbi a supporto nel latino tardo: Il caso di facio. *Acta Antiqua Academiae Scientiarum Hungaricae* 59. 145–160. https://doi.org/10.1556/068.2019.59.1-4.15

Giannakis, Georgios. 2023. At the Crossroads of Linguistics and Philology: The Tmesis-to-Univerbation Process in Ancient Greek. In Giannakis, Georgios & Filos, Panagiotis & Crespo, Emilio & de la Villa, Jesús (eds.), *Classical Philology and Linguistics: Old Themes and New Perspectives*, 175–211. Berlin; Boston: Mouton De Gruyter. https://doi.org/10.1515/9783111272887-008

Giomi, Riccardo. 2023. *A functional discourse grammar theory of grammaticalization*. Leiden: Brill. https://doi.org/10.1163/9789004520578

Giry-Schneider, Jacqueline 1987. *Les prédicats nominaux en français: Les phrases simples à verbe support*. Geneva: Droz.

Giry-Schneider, Jacqueline 1991. Relation entre le sens des noms et leur structure prédicative. *Revue québécoise de linguistique* 20(1). 99–124.

Goldberg, Adele. 2019. *Explain me this. Creativity, Competition, and the Partial Productivity of Constructions*. Princeton, NJ: Princeton University Press. https://doi.org/10.1515/9780691183954

Grimshaw, Jane & Mester, Armin. 1988. Light verbs and θ-marking. *Linguistic Inquiry* 19(2). 205–232.

Gross, Maurice. 1998. La fonction sémantique des verbes supports. *Travaux de Linguistique: Revue Internationale de Linguistique Française* 37(1). 25–46.

Gross, Gaston. 1989. *Les constructions converses du français*. Geneva: Droz.

Grossman, Eitan. 2023. Transitive verbs as lexical affixes in Coptic. Presented at the Affixes symposium, Turku.

Grossman, Eitan & Richter, Tonio. 2017. Dialectal variation and language change: The case of Greek loan-verb integration strategies in Coptic. In Grossman, Eitan & Dils, Peter & Richter, Tonio & Schenkel, Wolfgang (eds.), *Greek influence on Egyptian-Coptic: Contact-induced change in an ancient African language. DDGLC Working Papers 1*, 207–236. Hamburg: Widmaier.

Halliday, Michael & Hasan, Ruqaiya. 1976. *Cohesion in English*. London: Longman. https://doi.org/10.4324/9781315836010

Harris, Alice. 2008. Light verbs as classifiers in Udi. *Diachronica* 25(2). 213–248. https://doi.org/10.1075/dia.25.2.05har

Hartmann, Stefan. 2018. Derivational morphology in flux: A case study of word-formation change in German. *Cognitive Linguistics* 29(1). 77–119. https://doi.org/10.1515/cog-2016-0146

Haspelmath, Martin. 2000. Periphrasis. In Booij, Geert & Lehmann, Christian & Mugdan, Joachim (eds.), *Morphologie / Morphology. Ein internationales Handbuch zur Flexion und Wortbildung / An international handbook on inflection and word-formation*, vol. 1, 654–664. Berlin; Boston: Mouton De Gruyter. https://doi.org/10.1515/9783110111286.1.9.654

Heine, Antje. 2020. Zwischen Grammatik und Lexikon. Ein forschungsgeschichtlicher Blick auf Funktionsverbgefüge. In De Knop, Sabine & Hermann, Manon (eds.), *Funktionsverbgefüge im Fokus*, 15–38. Berlin; Boston: Mouton De Gruyter. https://doi.org/10.1515/9783110697353-002

Heusinger, Klaus von & Schumacher, Petra. 2019. Discourse prominence: Definition and application. *Journal of Pragmatics* 154. 117–127. https://doi.org/10.1016/j.pragma.2019.07.025

Hoffmann, Roland. 2023. Latin support-verb constructions: A view from language typology. In Baños, José Miguel & Jiménez López, María Dolores & Jiménez Martínez, María & Tur, Cristina (eds.), *Collocations in theoretical and applied linguistics: From classical languages to Romance languages*, 21–56. Madrid: SEEC.

Hoffmann, Sebastian. 2005. *Grammaticalization and English Complex Prepositions: A Corpus-Based Study.* Florence, US: Taylor & Francis.

Hoffmann, Thomas. 2018. Creativity and construction grammar: Cognitive and psychological issues. *Zeitschrift für Anglistik und Amerikanistik* 66(3). 259–276. https://doi.org/10.1515/zaa-2018-0024

Hopper, Paul & Traugott, Elizabeth. 2003. *Grammaticalization*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139165525

Ittzés, Máté. 2022. Light Verb, Auxiliary, Grammaticalization: The Case of the Vedic Periphrastic Perfect. *Die Sprache* 54. 95–129.

Ittzés, Máté. forthcoming. Proto-Indo-European support verbs and support-verb constructions. In Fendel, Victoria (ed.), *Between lexicon and grammar: Support-verb constructions in the corpora of Greek*. Berlin: Language Science Press.

Janse, Mark. 2023. “Girl, you’ll be a woman soon”: Grammatical versus semantic agreement of Greek hybrid nouns of the Mädchen type. In Giannakis, Georgios & Filos, Panagiotis & Crespo, Emilio & de la Villa, Jesús (eds.), *Classical philology and linguistics: Old themes and new perspectives*, 263–286. Berlin; Boston: Mouton De Gruyter. https://doi.org/10.1515/9783111272887-011

Jespersen, Otto. 1954. *A Modern English grammar on historical principles. Vol. VI: Morphology*. London: Bradford and Dickens.

Jiménez López, María Dolores. 2016. On Support Verb Constructions in Ancient Greek. *Archivio glottologico italiano* 51(2). 180–204. https://doi.org/10.1400/270167

Jiménez López, María Dolores. 2021. Γίγνομαι as the Lexical Passive of the Support Verb ποιέω in Ancient Greek. In Giannakis, Georgios & Conti, Luz & de la Villa, Jesús & Fornieles, Raquel (eds.), *Synchrony and Diachrony of Ancient Greek: Language, linguistics and philology*, 227–240. Berlin; Boston: Mouton De Gruyter. https://doi.org/10.1515/9783110719192-018

Jiménez López, María Dolores & Baños, José Miguel. 2022. Translation as a mechanism for the creation of collocations (II): The alternation of operor/facio in the Vulgate. In Baños, José Miguel & Jiménez López, María Dolores & Jiménez Martínez, María & Tur Altarriba, Cristina (eds.), *Collocations in theoretical and applied linguistics: From classical to Romance languages*, 189–234. Madrid: SEEC.

Joshi, Shrikant. 2020. Affixal negation. In Déprez, Viviane & Teresa, Espinal (eds.), *The Oxford handbook of negation*, 75–90. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198830528.013.10

Kamber, Alain. 2008. *Funktionsverbgefüge – empirisch: Eine korpusbasierte Untersuchung zu den nominalen Prädikaten des Deutschen*. Tübingen: Max Niemeyer. https://doi.org/10.1515/9783484970311

Kilani, Marwan. 2023. Egyptian and its auxiliary verb constructions: Exploring a typological “chimera.” Presented at Multilingualism and structural change: Insights from past histories and present realities, Lausanne.

Koev, Todor. 2022. *Parenthetical meaning*. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780198869535.001.0001

Lakoff, George & Johnson, Mark. 1980. *Metaphors we live by*. Chicago; London: University of Chicago Press. https://doi.org/10.2307/430414

Langer, Stefan. 2004. A linguistic test battery for support verb constructions. *Lingvisticæ Investigationes* 27(2). 171–184. https://doi.org/10.1075/li.27.2.03lan

Lavidas, Nikolaos. 2009. *Transitivity alternations in diachrony: Changes in argument structure and voice morphology*. Newcastle: Cambridge Scholars.

Lavidas, Nikolaos. 2015. The Greek Septuagint and language change at the syntax-semantics interface: From null to "pleonastic" object pronouns. In Gianollo, Chiara & Jäger, Agnes & Penka, Doris (eds.), *Language change at the syntax-semantics interface*, 153–182. Berlin; Boston: Mouton De Gruyter. https://doi.org/10.1515/9783110352306.153

Ledgeway, Adam & Vincent, Nigel. 2022. Periphrasis and inflexion: Lessons from Romance. In Ledgeway, Adam & Smith, John Charles & Vincent, Nigel (eds.), *Periphrasis and inflexion in diachrony: A view from romance*, 11–60. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780198870807.003.0002

Lehmann, Christian. 1988. Towards a typology of clause linkage. In Haiman, John & Thompson, Sandra (eds.), *Clause combining in grammar and discourse*, 181–225. Amsterdam: John Benjamins. https://doi.org/10.1075/tsl.18.09leh

Lehmann, Christian. 2020. Univerbation. *Folia Linguistica Historica* 41(1). 205–252. https://doi.org/10.1515/flih-2020-0007

Loporcaro, Michele. 2022. The morphological nature of person-driven auxiliation: Evidence from shape conditions. In Ledgeway, Adam & Smith, John Charles & Vincent, Nigel (eds.), *Periphrasis and inflexion in diachrony: A view from romance*, 213–237. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780198870807.003.0009

Loprieno, Antonio. 1995. *Ancient Egyptian: A linguistic introduction*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511611865

Luraghi, Silvia. 2003a. Definite referential null objects in Ancient Greek. *Indogermanische Forschungen* 108. 167–194. https://doi.org/10.1515/9783110243482.167

Luraghi, Silvia. 2003b. *On the meaning of prepositions and cases: The expression of semantic roles in ancient Greek*. Amsterdam: John Benjamins. https://doi.org/10.1075/slcs.67

Luraghi, Silvia. 2004. Null objects in Latin and Greek and the relevance of linguistic typology for language reconstruction. In Jones-Bley, Karlene & Huld, Martin & Della Volpe, Angela & Robbins Dexter, Miriam (eds.), *Proceedings of the fifteenth annual UCLA Indo-European conference, Los Angeles, November 7–8, 2003*, 234–256. Washington, D.C.: Institute for the Study of Man.

Manolessou, Io. 2001. The evolution of the demonstrative system in Greek. *Journal of Greek Linguistics* 2. 119–148. https://doi.org/10.1075/jgl.2.05man

Marchello-Nizia, Christiane. 1996. Les verbes supports en diachronie: Le cas du français. *Langages* 30(121). 91–98. https://www.persee.fr/doc/lgge_0458-726x_1996_num_30_121_1742

Marini, Emanuela. 2005. Attivo e deponente nei veri in -fico(r) a primo elemento sostantivale. *Papers on Grammar* 9(1). 171–185.

Marini, Emanuela. 2010. L'antipassivo in greco antico: ποιείσθαι come verbo supporto in Aristotele. *Journal of Latin Linguistics* 11(1). 147–180. https://doi.org/10.1515/joll.2010.11.1.147

Marini, Emanuela. 2014. L'opposition actif vs déponent et la persistance du moyen en latin. *Langages* 194. 49–61. https://shs.cairn.info/revue-langages-2014-2-page-49?lang=fr&tab=texte-integral

Marini, Emanuela. 2018. La fonction support et ses facettes: facere [+support] [+causatif] dans le type sacra facere. In Spevak, Olga & Bodelot, Colette (eds.), *Les constructions à verbe support en latin*, 129–147. Clermont-Ferrand: Presses Universitaires Blaise Pascal.

Markopoulos, Theodore. 2009. *The future in Greek: From ancient to medieval*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199539857.001.0001

Meissner, Torsten. 2016. Archaeology and the Archaeology of the Greek Language: On the Origin of the Greek Nouns in –εύς. In Bintliff, John & Rutter, N. Keith (eds.), *The Archaeology of Greece and Rome: Studies In Honour of Anthony Snodgrass*, 22–30. Edinburgh: Edinburgh University Press.

Mel'čuk, Igor. 2004. Verbes supports sans peine. *Lingvisticæ Investigationes* 27(2). 203–217. https://doi.org/10.1075/li.27.2.05mel

Mel'čuk, Igor. 2023. *General Phraseology: Theory and Practice*. Amsterdam: John Benjamins. https://doi.org/10.1075/lis.36

Myers-Scotton, Carol. 2002. *Contact Linguistics: Bilingual Encounters and Grammatical Outcomes*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198299530.001.000

Næss, Åshild. 2007. *Prototypical transitivity*. Amsterdam; Philadelphia: John Benjamins. https://doi.org/10.1075/tsl.72

Palme, Bernhard. 2009. The Range of Documentary Texts: Types and Categories. In Bagnall, Roger (ed.), *The Oxford Handbook of Papyrology*, 358–394. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199843695.013.0016

Polenz, Peter von. 1963. *Funktionsverben im heutigen Deutsch: Sprache in der rationalisierten Welt*. Düsseldorf: Pädogogischer Verlag Schwann.

Pompei, Anna. 2006. Tracce di incorporazione in greco antico. In Cuzzolin, Pierluigi & Napoli, Maria (eds.), *Fonologia e tipologia lessicale nella storia della lingua greca. Atti del Incontro Internazionale di Linguistica Greca, Bergamo, 15–17 September 2005*, 216–237. Milan: FrancoAngeli.

Pompei, Anna. 2014. Verb-particle constructions and preverbs in Homeric Greek between telicization, incorporation and valency change. In Bartolotta, Annamaria (ed.), *The Greek verb: Morphology, syntax, and semantics. Proceedings of the 8th international meeting on Greek linguistics (Agrigento 2009)*, 253–275. Louvain-la-Neuve and Walpole, MA: Peeters.

Pompei, Anna & Grandi, Nicola. 2012. Complex -éo verbs in Ancient Greek: A case study on the interface between derivation, compounding and inflection. *Morphology* 22. 399–416. https://link.springer.com/article/10.1007/s11525-012-9204-2

Quack, Joachim. 2017. How the Coptic script came about. In Grossman, Eitan & Dils, Peter & Richter, Tonio & Schenkel, Wolfgang (eds.), *Greek influence on Egyptian-Coptic: Contact-induced change in an ancient African language*, 27–96. Hamburg: Widmaier.

Radimský, Jan. 2011. Noms prédicatifs, noms de résultat et noms concrets dans les constructions à verbe support. *Lingvisticæ Investigationes* 34(2). 204–227. https://doi.org/10.1075/li.34.2.02rad

Rainer, Franz & Buridant, Claude. 2015. From Old French to Modern French. In Ohnheiser, Ingeborg & Olsen, Susan & Rainer, Franz & Müller, Peter (eds.), *Word formation: An international handbook of the languages of Europe*, vol. 3, 1975–2000. Berlin; Boston: Mouton De Gruyter. https://doi.org/10.1515/9783110375732-024

Reintges, Christoph. 2001. Code-mixing strategies in Coptic Egyptian. In Goldwasser, Orly & Sweeney, Deborah (eds.), *Structuring Egyptian syntax: A tribute to Sarah Israelit- Groll*, 193–237. Hamburg: Widmaier.

Rosén, Hannah. 2020. Composite predicates in the layers of Latin. *Journal of Latin Linguistics* 19(2). 231–279. https://doi.org/10.1515/joll-2020-2009

Rusten, Jeffrey. 2020. τὴν ἐκβολὴν τοῦ λόγου ἐποιησάμην: Thucydides’ chronicle in the Pentekontaetia (1.97–117) is not a digression. *Histos* 14. 230–254. https://histos.org/documents/2020AA08RustenEkbole.pdf

Saeedi, Zari. 2017. Nominal predication in Persian: A functional characterization. In Nolan, Brian & Diedrichsen, Elke (eds.), *Argument realisation in complex predicates and complex events: Verb-verb constructions at the syntax-semantics interface*, 373–412. Amsterdam; Philadelphia: John Benjamins. https://doi.org/10.1075/slcs.180.13sae

Sampson, Geoffrey. 2016. Two ideas of creativity. In Hinton, Martin (ed.), *Evidence. Experiment and argument in linguistics and philosophy of language*, 15–26. Bern: Peter Lang.

Savary, Agata & Candito, Marie & Mititelu, Verginica & Bejček, Eduard & Cap, Fabienne & Čéplö, Slavomír & Cordeiro, Silvio & Cordeiro, Silvio Ricardo & Eryiğit, Gülşen & Giouli, Voula & van Gompel, Maarten & HaCohen-Kerner, Yaakov & Kovalevskaite, Jolanta & Krek, Simon & Liebeskind, Chaya & Monti, Johanna & Parra Escartín, Carla & van der Plas, Lonneke & QasemiZadeh, Behrang & Ramisch, Carlos & Sangati, Federico & Stoyanova, Ivelina & Vincze, Veronika. 2018. PARSEME multilingual corpus of verbal multiword expressions. In Markantonatou, Stella & Ramisch, Carlos & Savary, Agata & Vincze, Veronika (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 87–147. Berlin: Language Science Press. https://doi.org/10.5281/zenodo.1471591

Schutzeichel, Marc. 2014. *Indogermanische Funktionsverbgefüge*. Münster: Monsenstein und Vannerdat OHG. https://nbn-resolving.de/urn:nbn:de:hbz:6-41379596963

Schwarz, Christian. 2004. *Die tun-Periphrase im Deutschen*. Munich: Ludwig-Maximilians University MA thesis. https://d-nb.info/1122919689/34

Sheinfux, Livnat & Greshler, Tali & Melnik, Nurit & Winter, Shuly. 2019. Verbal multiword expressions: Idiomaticity and flexibility. In Parmentier, Yannick & Waszczuk, Jakub (eds.), *Representation and parsing of multiword expressions*, 35–68. Berlin: Language Science Press. https://doi.org/10.5281/zenodo.2579035

Slade, Benjamin. 2013. Diachrony of light and auxiliary verbs in Indo-Aryan. *Diachronica* 30(4). 531–578. https://doi.org/10.1075/dia.30.4.04sla

Squeri, Elena. forthcoming. χράομαι as a support verb in the medical jargon of the Hippocratic Corpus. In Fendel, Victoria (ed.), *Support-verb constructions in the corpora of Greek: Between lexicon and syntax?* Berlin: Language Science Press.

Storrer, Angelika. 2009. Corpus-based investigations on German support verb constructions. In Fellbaum, Christiane (ed.), *Idioms and collocations: Corpus-based linguistic and lexicographic studies*, 164–187. London: Continuum. https://hal.science/hal-01103474/document

Tribulato, Olga. 2015. *Ancient Greek verb-initial compounds: Their diachronic development within the Greek compound system*. Berlin; Boston: Mouton De Gruyter. https://doi.org/10.1515/9783110415827

Tronci, Liana. 2016. Sur le syntagme prépositionnel en N du grec ancien: Syntaxe et lexique de ses combinaisons avec le verbe ékhein “avoir.” In Marque-Pucheu, Christina & Kakoyanni-Doa, Fryni & Machonis, Peter & Ulland, Harald (eds.), *À la recherche de la prédication: Autour des syntagmes prépositionnels*, 141–158. Amsterdam: John Benjamins. https://doi.org/10.1075/lis.32.08tro

Tronci, Liana. 2017a. At the lexicon-syntax interface Ancient Greek constructions with ἔχειν and psychological nouns. In Georgakopoulos, Thanasis & Pavlidou, Theodossia-Soula & Pechlivanos, Miltos & Alexiadou, Artemis & Androutsopoulos, Jannis & Kalokairinos, Alexis & Skopeteas, Stavros & Stathi, Katerina (eds.), *Proceedings of the ICGL12, Vol. 2*, 1021–1033. Berlin: Edition Romiosini/ CeMoG. https://www.cemog.fu-berlin.de/en/icgl12/offprints/tronci/index.html

Tronci, Liana. 2017b. Quelques remarques pour une reconsidération des verbes latins en -isso/-izo/-idio. *Pallas* 103. 293–300.

Tronci, Liana & Logozzo, Felicia. 2022. Pseudo-coordination and serial verbs in Hellenistic Greek? Some insights from the New Testament and the Septuagint. *Journal of Greek Linguistics* 22. 72–144. https://doi.org/10.1163/15699846-02201003

Tutin, Agnès. 2016. Comparing morphological and syntactic variations of support verb constructions and verbal full phrasemes in French: A corpus based study. *PARSEME COST Action. Relieving the pain in the neck in natural language processing: 7th final general meeting*. Dubrovnik, Croatia. https://shs.hal.science/halshs-01377956/document

Van Camp, Bruno. 2005. À propos de λόγον διδόναι, formule-clé de la dialectique platonicienne. *Revue belge de Philologie et d'Histoire* 83(1). 55–62. https://www.persee.fr/doc/rbph_0035-0818_2005_num_83_1_4909

Vendler, Zeno. 1967. *Linguistics in Philosophy*. Ithaca, NY: Cornell University Press. https://doi.org/10.7591/9781501743726

Vincent, Nigel & Wheeler, Max. 2022. Layering and divergence in Romance periphrases. In Ledgeway, Adam & Smith, John Charles & Vincent, Nigel (eds.), *Periphrasis and inflexion in diachrony: A view from romance*, 93–122. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780198870807.003.0004

Vives Cuesta, Alfonso & Madrigal Acero, Lucía. 2022. Support-verb constructions in Postclassical Greek and sociolinguistics: A diachronic study of εὐχὴν ποιέω as a level-of-speech marker. In Jiménez López, María Dolores & Jiménez Martínez, María & Tur Altarriba, Cristina & Baños, José Miguel (eds.), *Collocations in theoretical and applied linguistics: From classical languages to Romance languages*, 305–334. Madrid: SEEC.

Vivès, Robert. 1983. *Avoir, prendre, perdre: Constructions à verbe support et extensions aspectualles*. France: Paris 8 Thèse de 3e cycle.

Wittenberg, Eva & Levy, Roger. 2017. If you want a quick kiss, make it count: How choice of syntactic construction affects event construal. *Journal of Memory and Language* 94. 254–271. https://doi.org/10.1016/j.jml.2016.12.001

Wittenberg, Eva & Snedeker, Jesse. 2014. It takes two to kiss, but does it take three to give a kiss? Categorization based on thematic roles. *Language and Cognitive Processes* 29(5). 635–641. https://doi.org/10.1080/01690965.2013.831918

Wittenberg, Eva & Trotzke, Andreas. 2021. Semantic incorporation and discourse prominence: Experimental evidence from English pronoun resolution. *Journal of Pragmatics* 186. 87–99. https://doi.org/10.1016/j.pragma.2021.09.019

Zakrzewska, Ewa. 2017. Complex verbs in Bohairic Coptic: Language contact and valency. In Nolan, Brian & Diedrichsen, Elke (eds.), *Argument realisation in complex predicates and complex events: Verb-verb constructions at the syntax-semantics interface*, 213–243. Amsterdam; Philadelphia: John Benjamins. https://doi.org/10.1075/slcs.180.08zak

Zilliacus, Henrik. 1956. Zur Umschreibung des Verbums in spätgriechischen Urkunden. *Eranos* 54. 160–166.

Contact information:

Victoria Beatrix Fendel
University of Oxford
e-mail: victoria.fendel@classics.ox.ac.uk
ORCID: https://orcid.org/0000-0001-6302-3726

Finnish Romani during the 20th century: Development and decay of a language

Kimmo Granqvist
University of Helsinki

Abstract

This paper discusses the history of Finnish Romani during the 20th century. The 20th century was a period of decreasing use of Finnish Romani and internal language erosion, resulting in the co-existence of inflecting Romani retaining its morpho-syntactic framework, and a Para-Romani-like variety. Numerous attested changes in Finnish Romani were attributed to the contact of the speakers of Romani with Finnish. This paper is based on a corpus comprising most written documents of Finnish Romani from the 20th century, comprising religious texts, textbooks, wordlists and dictionaries. This paper follows the theoretical framework as commonly used in European and Finnish Romani Linguistics during the 1990s and 2000s. Starting points for explaining language variation and change are functional-typological. The focus is on language-internal changes, innovations vs. conservative features, as well as on contact-induced changes. In addition, this paper deals with variation and linguistic attrition, which characterized Finnish Romani of the 20th century. This paper shows that tendencies of language-internal change both simplified the language structure and significantly increased the amount of linguistic variation. The paper also shows that matter replication is later and occurs more rarely in FR than pattern replication.

Keywords: Finnish Romani, history, variation, language change

1 Introduction

This paper discusses the development of Finnish Romani (FR) during the 20th century. Its focus is on language-internal changes, innovations vs. conservative features, as well as on contact-induced changes. During this period, the structure of FR has shown multiple tendencies towards simplification. As a result of many competing ongoing tendencies of language change, limited use of the language and the bilingualism of the Finnish Roma, FR has exhibited abundant variation in both nominal and verb morphology. Some of these changes seemingly coincide with Northeastern (NE) Romani (some of the Romani dialects spoken in Poland, the Baltic countries and Russia), but have probably been caused by natural linguistic processes or typological similarities between the contact languages of FR and NE Romani.

Abandoning the morpho-syntactic frame of Romani has been a gradual process, but it has accelerated since the beginning of the 20th century. Many speakers of FR imitate the syntactic framework of Finnish mostly using Indo-Aryan resources, without exhibiting large-scale borrowing of the Finnish morphological matter. FR has become tightly connected to Finnish used by the Roma for interaction outside their own group and even among themselves. The borrowing of phonological rules is common in language contact

VERTAISARVIOITU
KOLLEGIALT GRANSKAD
PEER-REVIEWED
www.tsv.fi/tunnus

Finnish Journal of Linguistics 37 (2024), 73–104
DOI: 10.61197/fjl.143388

situations. Syntax is one of the levels of the language that mostly converges with Finnish. This is not surprising since, in European Romani dialects, syntax is generally prone to contact-induced influences affecting their structure.

Idiolects show significant variation in FR. Among speakers of all ages, there are nowadays competent and semi-competent speakers as well as those with very weak proficiency in Romani. Differences in Romani competence are mostly reflected in verb morphology and the ability to produce complex numerals. Code-mixing phenomena are frequent in the speech of the Finnish Roma. In FR, code-mixing with Finnish is largely a compensation strategy, by means of which the Roma fill gaps in their Romani competence. In addition, code-mixing is connected to the efficiency of lexical retrieval, which is often faster in the dominating language, Finnish. (Granqvist 2000; Kovanen 2013; Salo 2021.) Kovanen (2013) further suggests that code-switching has social or interactional functions to mark taboos or difficult topics, to outline the discourse and to seek attention.

1.1 Previous studies on the history of FR

Granqvist (2010) divided the history of Finnish Romani linguistics in three periods: (i) a historical perspective represented by Pertti Valtonen especially in the 1960s, (ii) a Fennistic paradigm during the 1980s and 1990s, and (iii) the modern period since the beginning of the 21st century. In the modern period, new methods of linguistics began to be applied to the core study of the language focusing on the relationship between Romani and Old and Middle Indo-Aryan languages. Research shifted to focus on phonetics, phonology and morphosyntax, reflecting the interests and training of contemporary linguists. However, the connection to the traditional historical perspective of Finnish Romani linguistics was maintained. Since 2010, the historical development of Finnish Romani has been explored in numerous conference presentations and articles (Granqvist 2012b, 2013a, 2013b). A comprehensive monograph on the history of Finnish Romani has recently been published (Granqvist 2024).

1.2 Material and method

1.2.1 Material

In order to study the development of FR during the 20th century, we need to examine data from the period. Only a few documents in FR are available from the first half of the 20th century. Most of the existing materials come from Oskar Jalkio, the founder of the Mustalaislähetys (Gypsy Mission) NGO. Jalkio's data are currently archived at the Institute for the Languages of Finland and Romani Mission NGO. Jalkio mastered Romani but was not a native speaker of it. Not all his materials were translated by him; instead, they comprised language produced by different Roma in different times. In additions to Jalkio's data, articles published by the Estonian professor Paul Ariste contain Romani from the first decades of the 20th century.

The latter half of the 20th century is the first period from which materials in both spoken and written Romani are available, along with a number of scholarly writings elaborating the view of its structure and development. Spoken Romani has been tape-recorded since the 1960s, when Pertti Valtonen collected it for his academic theses (1964, 1968). Simultaneously, Matti Leiwo, professor emeritus of Finnish, and Pekka Sammallahti, professor emeritus of Saami languages, also became interested in Romani

and conducted their own data collections. These were very limited; nevertheless, they have not yet been fully utilized for research. In 1984, Yrjö Temo, a Finnish Rom who had written a Romani dictionary with his brother Jussi Peltosalmi and translated the Gospel of John into Romani (both published in 2014 (Granqvist 2014a)), was interviewed for two hours. Thereafter, about 15 hours of recordings of Romani were obtained during a Romani language seminar organized by the Research Institute for the Languages of Finland in 1995. At the beginning of the 21st century, Hellevi Hedman-Valentin, a planner at the Research Institute, recorded and transcribed a total of 45 hours of speech from 89 Roma (46 woman and 43 men between 16 and 89 years of age) from 32 localities all over the country. The size of the resulting corpus was 168,000 words. The most recent data collection was carried out in 2013–2014 as part of a University Helsinki project called "Finnish Romani and other Northern Romani dialects in the Baltic Sea area" (2013–2017, PI Kimmo Granqvist).

The written sources of FR from the period 1950–2013 are summarized in Table 1. Until the 1980s, religious texts constituted most of the written sources. Among the few exceptions were some newspaper articles in *Romano Boodos* – issued by Mustalaislähetys, later *Romano Missio* – Axel Kronqvist's (1871–1956) vocabulary and Pertti Valtonen's (1972) etymological dictionary. The Romani language and culture has received more attention in the national Roma political debate since the 1980s. This resulted in the publication of dictionaries and textbooks intended mainly for comprehensive schools and vocational training, later also for the university. A few children's books have been published during the past few years.

In addition, some web pages of Finnish authorities have been translated in Romani. Radio news in Romani has been transmitted weekly since 1995 by Radio Suomi and Yleisradio 1; some of the new manuscripts have been obtained and saved into a corpus.

Table 1. Central written sources of FR 1950–1999

A. Religious texts				
	Year	Translator/author	Title	Size (pp.)
Non-Roma translators/ authors	1970	Pertti Valtonen	*Markusesko Evankeliumos: kaalengo tšibbaha*	
Roma translators/ authors	1970	Viljo Koivisto	*Deulikaane tšambibi*	96
	1971	Viljo Koivisto	*Johannesesko Evankeliumos*	72
	1971–	Viljo Koivisto and other authors	*Romano boodos* newspaper	–
	1980s	Jussi Peltosalmi and Yrjö Temo	*Evankeliumis pale Johannesksese*	Manuscript

B. Textbooks

	Year	Author	Title	Size (pp.)
Roma authors	1982	Viljo Koivisto	*Drabibosko ta rannibosko byrjiba*	108
	1987	Viljo Koivisto	*Rakkavaha romanes. Kaalengo tšimbako sikjibosko liin*	280
	1995	Miranda Vuolasranta	*Romani tšimbako drom*	150
	1996	Henry Hedman	*Sar me sikjavaa romanes*	243

C. Vocabularies and dictionaries

	Year	Translator/author	Title	Size (pp.)
Non-Roma authors	1950s	Axel Kronqvist		Manuscript
	1972	Pertti Valtonen	*Romanikielen etymologinen sanakirja*	
	2010	Kimmo Granqvist	*Suomen romanikielen käänteissanakirja*	111
Roma authors	1980s	Jussi Peltosalmi and Yrjö Temo	Suomi–romani -sanakirja	Manuscript
	1994/2005	Viljo Koivisto	*Romano-finitiko-angliko laavesko liin = Romani-suomi-englanti sanakirja = Romany-Finnish-English dictionary*	324

1.2.2 Method

This paper follows the theoretical framework as commonly used in European and Finnish Romani Linguistics during the 1990s and 2000s. The starting points for explaining language variation and change are functional-typological (Martinet 1962; Greenberg 1966; Anttila 1972; Coseriu 1974; Givón 1985a, 1985b); the paper further discusses the relationship between form and function and language complexity.

Its focus is on language-internal changes, innovations vs. conservative features. In FR, language internal variation is mainly caused by innovations that have simplified the structure of Romani dialects. These include the loss of case agreement of adjective attributes, loss of subject clitics, changes in the inventory of verbal derivation suffixes (e.g. the loss of productivity of *-ar-*), reduction of *-v-* in verb forms (e.g., *čēr-av-a* > *cēr-a-a* [do-PRS.1SG-IND/FUT] ‘I do’) and syncretism phenomena in person inflections.

In addition to language-internal changes, this paper discusses contact-induced language changes. The lexicon of FR has been influenced in particular by the Germanic languages Middle Low German and Middle High German, Danish and Swedish. On the other hand, not many lexemes have been borrowed from Finnish, which is the most important current close contact language of FR, but a large number of loans are based on Finnish patterns. The transfer of Finnish phonological rules into FR has been extensive, but most Finnish rules had already been borrowed by the end of the 19th century. One focus in this paper is on morphological borrowing and related universal constraints (e.g., Moravcsik 1978; Thomason & Kaufmann 1988; Thomason 2001; Winford 2003). In this paper, I distinguish between the borrowing of morphemes, i.e. *matter replication*, and of grammatical models, i.e. *pattern replication*, of which the former is rare but the latter frequent in FR (Matras 2007; Matras & Sakel 2007; Sakel 2007). In some cases, Finnish morphosyntactic patterns even constitute an obligatory part of FR grammar. For instance, the morphosemantic functions of cases and the case government of verbs are largely borrowed from Finnish, likewise the inherited prepositions have been substituted with postpositions that trigger genitive complement similar to Finnish.

In addition, this paper deals to some extent with variation and linguistic attrition. At its present stage, FR permits a high degree of variation. Permissiveness to variation and large idiolectal differences in the amount of variation are often associated with language death (Vuorela & Borin 1998: 69, and the references therein). The attrition is so extensive, that some scholars refer to FR as an obsolescent language (Vuorela & Borin 1988: 69; Pirttisaari 2003, 2004a, 2004b: 178). Along with the attrition and the gradual loss of its own grammatical framework, FR has become increasingly symbiotic with Finnish, or Para-Romani[1]-like in some respects, while the number of its non-speakers has perhaps not increased. Hancock (1992), Vuorela and Borin (1998: 68) and Granqvist (2013: 184–185) even refer to a Finnish Para-Romani called Fennoromani, analogously to Angloromani and Scandoromani.

1.3 Outline

This paper is divided into three main sections followed by a section summarizing the key findings. First, I discuss the proficiency of the Roma in FR and its domains of use. It draws on surveys carried out by academic scholars and authorities periodically since the 1950s, mostly to survey the living conditions of the Roma and secondarily only the linguistic situation of the Roma. Thereafter, I deal with language-internal changes. The focus is on morpho-syntax: the simplification of the case system and noun classes, the simplification of the verb classes and person inflections and the development of the so-called 'new infinitive'. Finally, I discuss the changes FR has undergone during the 20th century in contact with Finnish. Here also, the emphasis is on morphological features. Most of the phonological contact-induced changes had already taken place during the 19th century or earlier. The discussion is divided into pattern replication – typological changes comprising the loss of definite articles, the change into a postposition language and the formation of new types of analytical tenses – and matter replication.

[1] The term Para-Romani, often coined with Cortiade (1991), refers to varieties with (some) Romani-based lexicon incorporated in the morphosyntactic frame of another language, e.g. Spanish (in case of Caló), Portuguese (Calão), Basque (Errumantxela), English (Angloromani), Norwegian or Swedish (Scandoromani). On Para-Romani varieties, see also Bakker (2020).

2. Surveys on proficiency in Romani and its domains of use

2.1 Language proficiency in Romani

The competence of the Roma in the Romani language has been surveyed several times, beginning in the 1950s, by means of self-assessment. In 1953, the Finnish government established a work group to investigate Roma issues, chaired by Social Counselor Paavo Mustala. Mustala's committee tasked the Social Investigations Bureau to study the Roma and their living conditions. The data for this study were collected in 1954 through Social Boards, so that all persons older than 16 years of age with a Roma background (and even younger orphan Roma children) from all municipalities were included. A total of 2,074 questionnaires were returned by June 1954; eventually the data covered information on 3,596 Roma and persons living with them. The Roma were fairly equally distributed throughout the country, though up to 25% of the people interviewed lived in Vaasa county. The survey covered their belonging to the church, age, level of education, reading and writing skills, confirmation, school attendance, marital status, family size and housing conditions, but also competence in Romani.

The empirical part of Raino Vehmas's PhD thesis (1961), dealing with the group character and cultural acculturation of the Roma, was based on interviews with 89 Roma living in rural areas and 88 Roma living in cities. The rural areas were represented by the Saarijärvi-Viitasaari region and the city areas by Helsinki. The interviews were carried out by social secretaries in the countryside and by social workers in Helsinki. The questionnaire contained 67 questions, two of which dealt with usage of and competence in the Romani language. Its focus was on the group behaviour, experiences, social participation, cognitive activity and attitudes of the Roma. The data that Vehmas (1961) published were based on the earlier survey carried out by the Social Bureau of Research in 1954 (Granqvist 2010).

In 1978–1980, the Helsinki Welfare Office investigated the social and educational conditions of the Roma by interviewing the heads of Roma households. The goal was to gather information to follow and develop the living conditions of the Roma. The survey covered the education, livelihood and other social conditions of the Roma, as well as canvassing the wishes of the Roma as to what could be done to improve their living conditions. Between May 1, 1978 through October 10, 1979, 185 interviews were accomplished; the total number of Roma households was 286 on July 15, 1978. 550 persons belonged to these households. Most interviews were carried out in Welfare Offices or at the interviewees' homes. The interviewees were mostly heads of the households, but in the case of mixed marriages, they were the Roma spouses.

Henry Hedman's (2009) study on Finnish Romani, its status in its speech community, usage and language attitudes of the Roma is currently the most extensive survey of the competence into Romani and its domains of use. For it, 306 Roma (164 women and 142 men) were interviewed throughout Finland and in Sweden.

Obviously, these surveys are not commensurable, due to differences in sample sizes and their quality and in self-assessment criteria. The 1954 survey studied speaking and listening comprehension skills in Romani. Vehmas (1961) used a four-grade scale, starting at "masters perfectly" and ending at "does not understand Romani". The Helsinki Welfare Office made use of a three-grade scale based on communicative competence ("gets along with elderly Roma" – "gets along in everyday conversations" – "knows only a few

words"). Hedman (2009) used a five-grade scale ("excellent" – "good" – "satisfactory" – "weak" – "not at all"). Hedman (2009: 24) defined satisfactory competence as ability to partly understand Romani and to be able to at least partly reply in Romani. Weak language skills meant that the person could not get along in Romani (Hedman 2009: 24).

In 1954, the elderly generation in particular tended to have mastered Romani; 85% of the Roma over 65 years of age spoke Romani. Of adult Roma, 81% at least understood Romani. According to the statistics published by Vehmas (1961: 188), 60% of adult Roma mastered Romani perfectly or well according to their own self-assessment, and 89% of the interviewees could get along in Romani. Vehmas did not observe significant differences in Romani competence between countryside and city. According to the survey by the Helsinki Welfare office (1979) more than half of the household heads had mastered Romani well enough that they could at least get along in everyday conversations and 37% so well that considered themselves able to get along with elderly Roma. However, up to 88% of Roma household heads, whose spouse was also of Roma background, was able to get along in everyday conversations. According to Hedman (2009), on the other hand, no more than 28% or the interviewed Roma claimed that he/she had mastered Romani excellently or well. But 62% of them had at least satisfactory competence.

Figures 1–3 compare the surveys. A central observation is that only excellent or good competence in Romani seems to have clearly decreased. The proportion of satisfactory competence has not changed significantly, and the amount of Roma not knowing Romani is nowadays rather lower than it used to be. However, Hedman's (2009: 24) survey indicates that almost one-third of the Roma have weak competence in Romani.

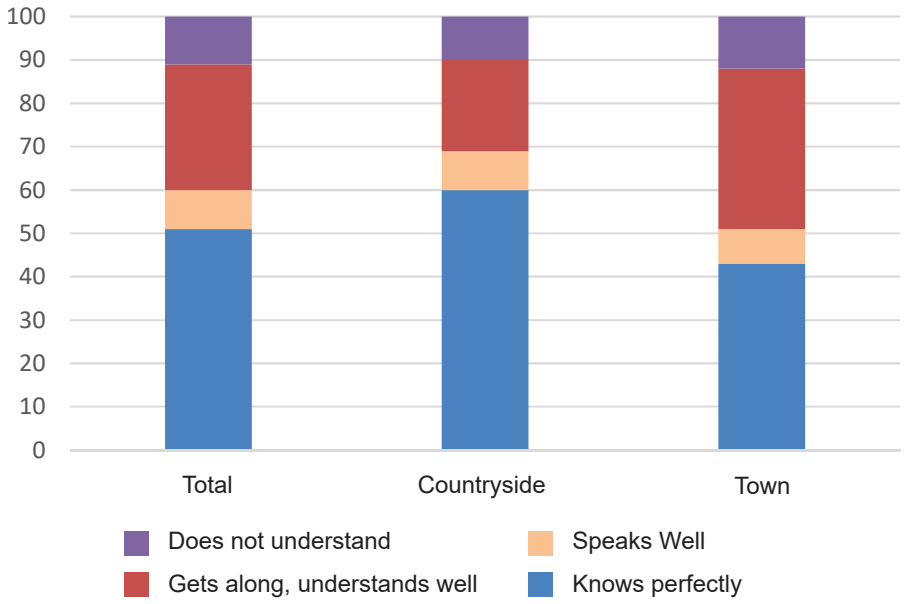


Figure 1. Competence in FR according to Vehmas (1961)

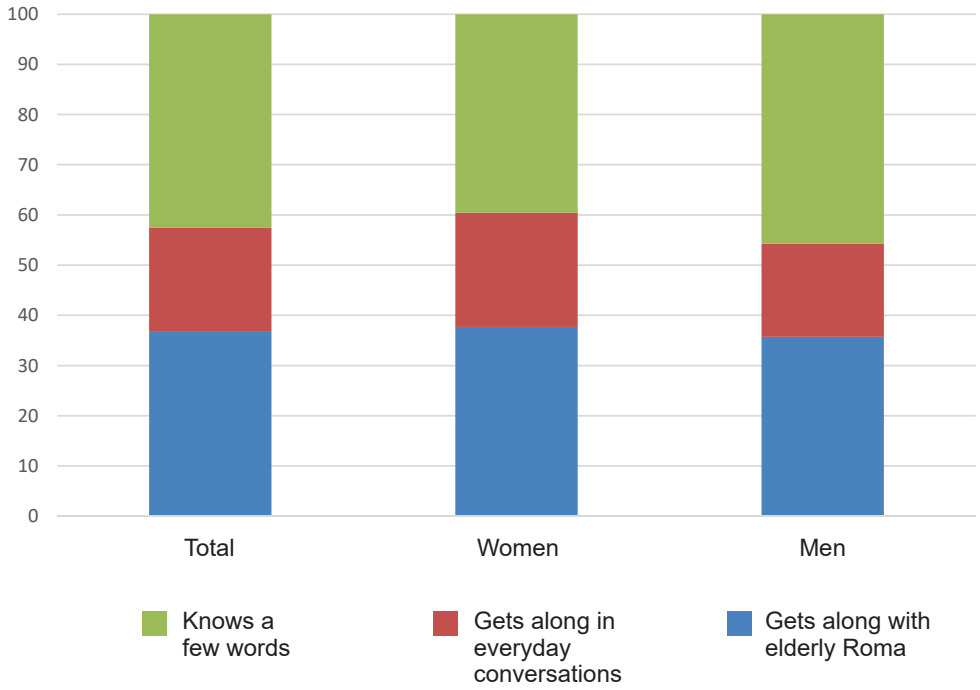


Figure 2. Competence in FR according to Helsinki welfare office survey (1979)

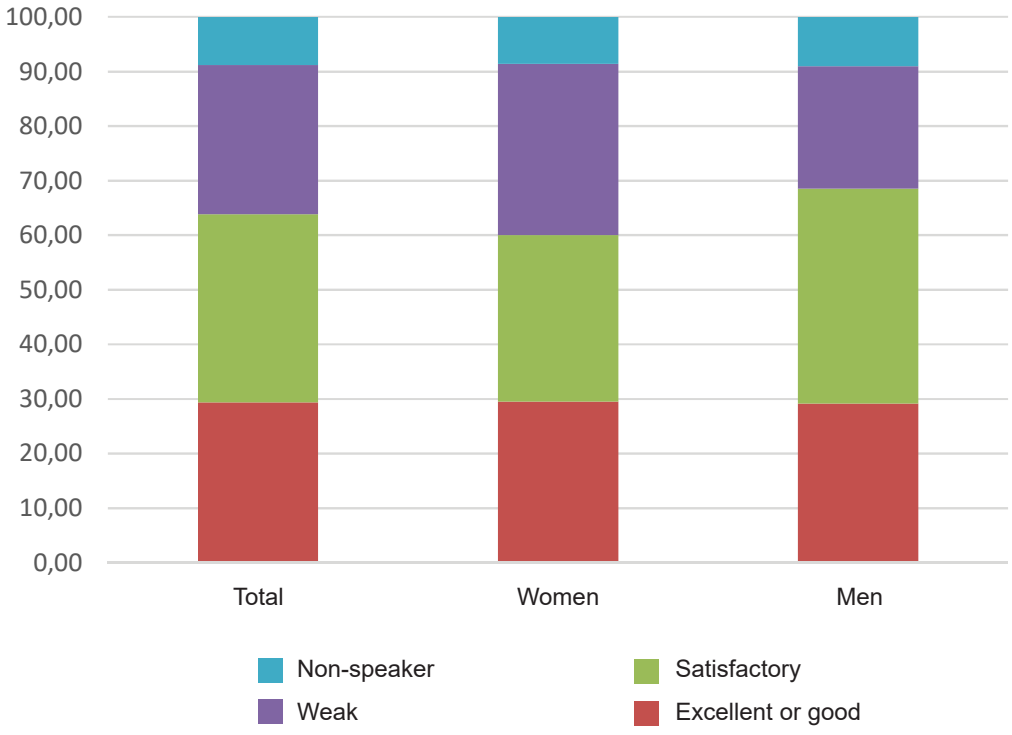


Figure 3. Competence in FR according to Hedman (2009)

Three surveys present quantitative results on the Romani competence by age-groups; the results are summarized in Figures 4–6. The survey carried out by the Social Investigation of Office in 1954 compared the oral skills of three age groups in Romani. Both the Helsinki Welfare Office and Hedman (2009) divided the informants into five age groups but used different age categories. The differences between age groups have increased over time, in particular when it comes to good competence in Romani. All three surveys repeat the view that the older Roma have better Romani competence than the younger. The big differences between age groups are related to the late acquisition of Romani along with growing up to adulthood and socialization in the Roma community (Borin & Vuorela 1998: 60; Borin 2000: 75). The differences between age groups might also reflect the gerontocratic hierarchy of the Roma community. The young Roma must pay respect to the older and aim at preserving the face of the older Roma; they must show that the older Roma are wiser than they (Granqvist 2009). Because of this, it is difficult for the young Roma to characterize their competence in Romani very strongly, to avoid the risk of exalting themselves above the older Roma. Self-assessments are, in addition, biased by the pursuit of ideal Romaniness; some Roma consider mastering of Romani important in this. Granqvist (2013a) has suggested that variation in Romani competence manifests itself more aptly through comparison of language-internal variables than using interview techniques based on self-assessment. Comparison of results based on language-internal variables and self-assessments also reveals variation in the notion of good Romani competence over time and person.

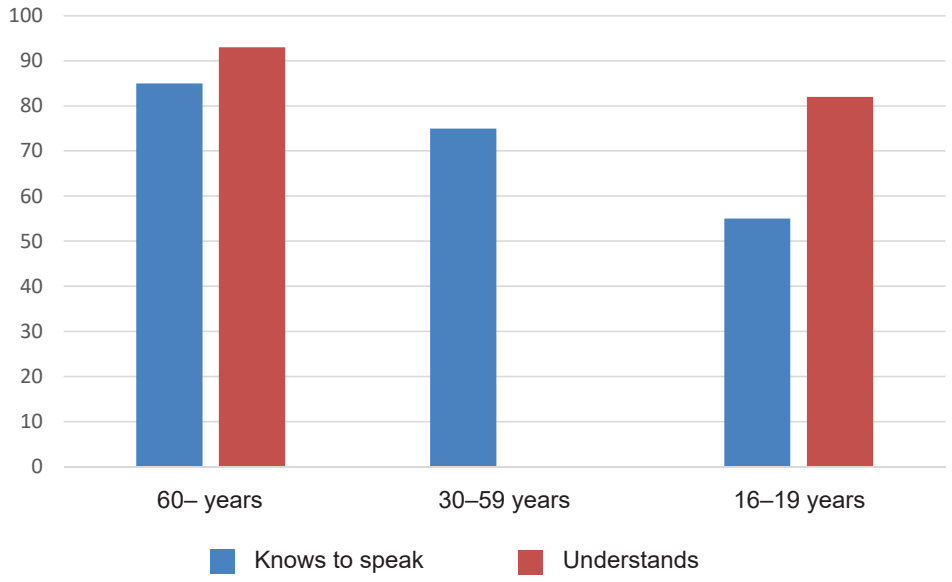


Figure 4. Competence in FR by age group according to Social Yearbook (1954). Statistics on understanding FR are missing for the age group 30–59 years.

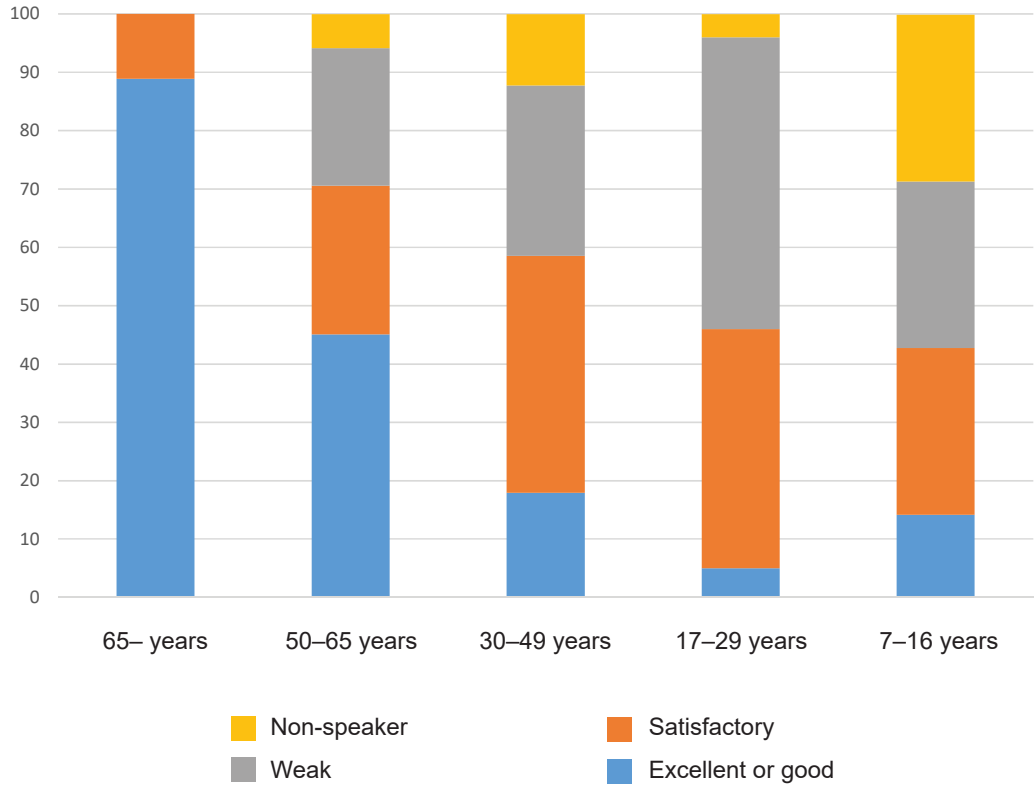


Figure 5. Competence in FR by age group according to Helsinki welfare office survey (1979)

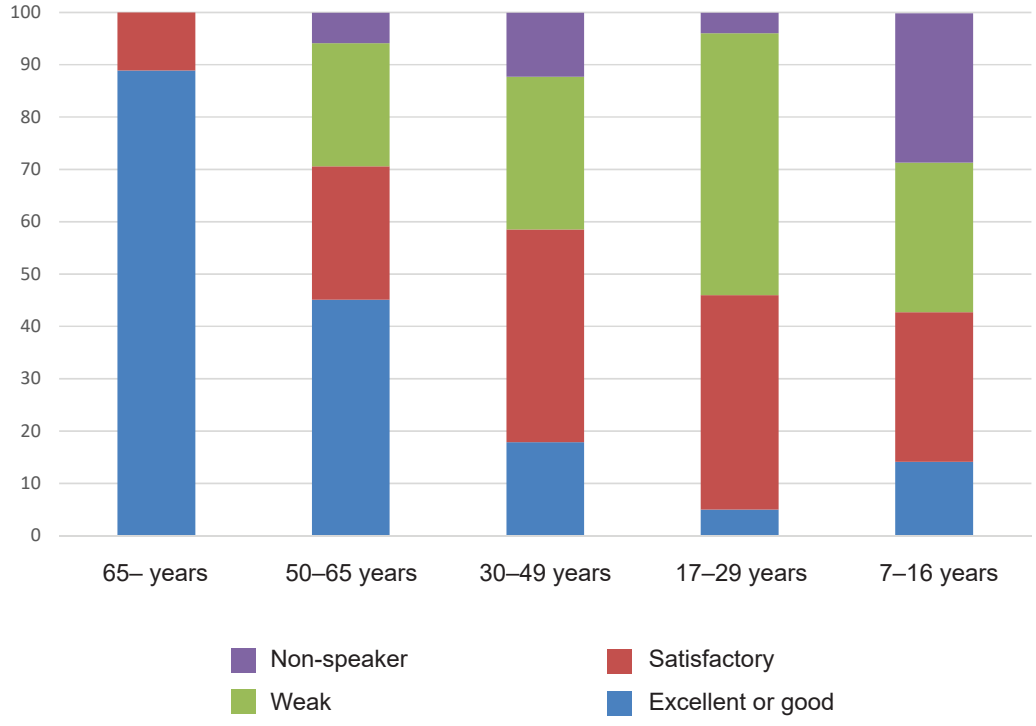


Figure 6. Competence in FR by age group according to Hedman (2009)

2.2 Domains of use

By the beginning of the 20th century, FR was no longer the primary language of everyday conversations of Roma families (Thesleff 1899). In the 1950s, 81% of the interviewed Roma used Finnish mostly or exclusively within the family; the use of Romani was slightly more widespread in the countryside than in cities. At the beginning of the 2000s, Finnish (occasionally parallel to Swedish) was the only home language of approximately 60% of the interviewees. About 40% of the interviewees declared the use of Romani in parallel to at least some extent. (Hedman 2009: 31–32.) As for other private domains of use, Hedman (2009: 32–33) mentions that the Roma also use Romani outside the home when they meet other Roma, do car and horse business, at marketplaces and shops, and at spiritual meetings. Figures 7–9 compare domains of use.

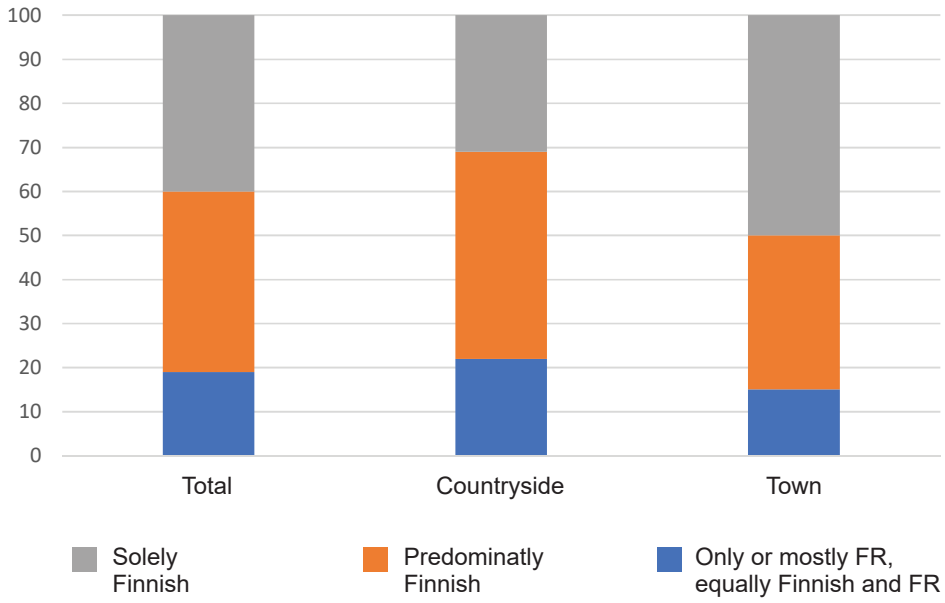


Figure 7. FR as language of conversations (Vehmas 1961)

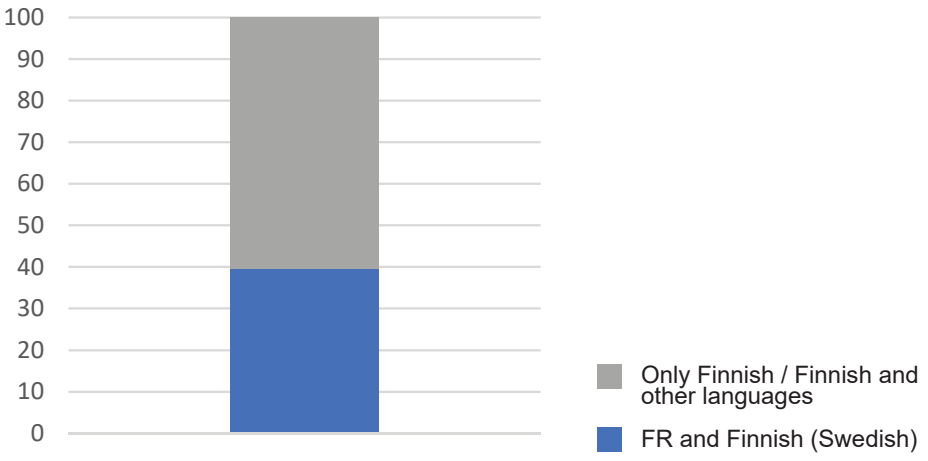


Figure 8. Home languages of the Roma (Hedman 2009)

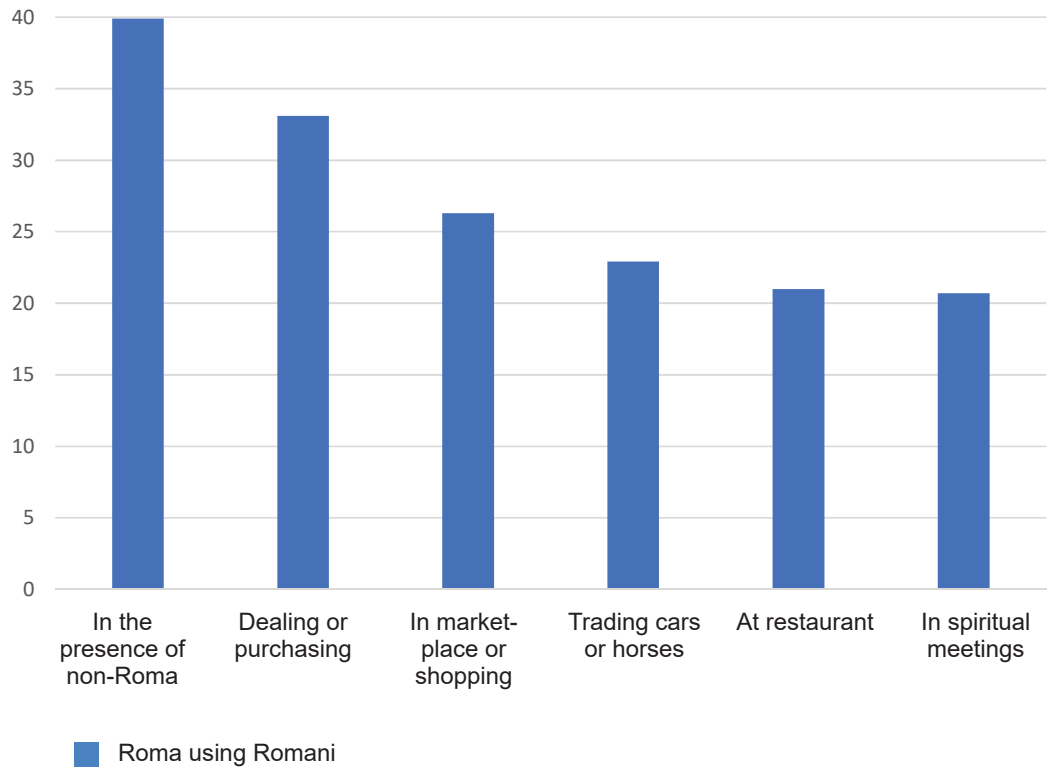


Figure 9. Other private domains of FR (Hedman 2009)

FR started to be used in public domains during the latter half of the 20th century, in school teaching, textbooks and dictionaries, religious circles and, more limitedly, as a language of administration. The attitudes of Roma towards the public use of Romani have been divided, though many Roma welcomed Romani language newspaper and journal articles and books, according to both Helsinki Welfare Office's (1979) and Hedman's (2009: 57–58) surveys. Generally, the attitudes of the Roma have also been positive towards Romani language teaching. In 1979, up to 73.5% of the interviewed Roma supported it, and at the beginning of the 2000s, approximately 80% of the Roma were positively inclined. Negative attitudes were reported among the elderly and the least educated Roma. The elderly Roma who had themselves mastered Romani fluently were not yet in 1979 concerned about the future of the language and believed that Romani was transferred to the next generations naturally without formal school teaching. Some Roma thought that Romani language would be taught by non-Roma at schools and were therefore negatively inclined.

3. Language-internal changes during the 20th century

3.1 Simplification of the case system

Many present-day Romani dialects have a three-layer case system comprising the fusional primary cases nominative and oblique as *layer I*; the agglutinative secondary cases dative, locative, ablative, instrumental and genitive as *layer II*; and the open class of analytical adpositions as *layer III*; outside these layers, many Romani dialects have vocative (Masica 1993; Matras 2002). FR belongs to those Romani dialects that preserve the conservative case inflection, including both primary cases, most of the secondary cases and analytical

adpositions. However, the last traces of the vocative are probably in Oskari Jalkio's religious texts dating back to the beginning of the 20th century. FR distinguishes between short and long genitive, of which only the short ones show possessions, while the long ones are rather derived nouns and adjectives. The markers of the secondary cases are summarized in Table 2; they have undergone only minor phonological changes in FR, mostly during the 18th–19th centuries, including the elision of final /r/ in the ablative marker *-tar* > *-ta* and the assimilation of the postnasal voiced stop into the preceding nasal (e.g., *mandar* > *manna* [I.OBL.ABL] 'from me') and the generalization of *-ha* in instrumental singular as a result of /s/ > /h/ sound change, which is an Early Romani option selection, i.e. an in-situ selection of modern Romani dialects from Early Romani variation.

Table 2. Oblique forms and secondary case markers in FR

	Singular				Plural	
	Masculine		Feminine		Masc./fem.	
	Obl.	Sec. case	Obl.	Sec. case	Obl.	Sec. case
Dative	*-es-*	*-ke*	*-(j)a-*	*-ke*	*-(j)jen-*	*-ge*
Ablative	*-es-*	*-ta*	*-(j)a-*	*-ta*	*-(j)en-*	*-na*
Locative	*-es-*	*-te*	*-(j)a-*	*-te*	*-(j)en-*	*-ne*
Instrumental	*-e-, -es-*	*-ha -sa*	*-(j)a-*	*-ha*	*-(j)en-*	*-sa*
"short" genitive / "long" genitive	*-es-*	*-k-o, -i, -e / -ker-o, -i, -e*	*-(j)a-*	*-k-o, -i, -e / -ker-o, -i, -e*	*-(j)e-n-*	*-g-o, -i, -e / -ger-o, -i, -e*

Two very important morpho-syntactic developments should be noted. The loss of the locative is still on-going, while the Suffixaufnahme (case stacking) of genitives has been lost. Both are innovations that simplify the language. The locative started to be rare in spoken Romani throughout the 20th century (on locative, see Baló 2021). Valtonen (1968: 166) perceives the loss of the locative as a process that took place between the 19th century and the 1960s and involved idiolectal variation. According to Valtonen (1968: 166), the locative was still in use in the "upper style" as the case of prepositional complements, but in "lower style" independently as prepositions were already being omitted in Jalkio's times. By the 1960s, the nominative had started to replace the locative in the "lower style". But the loss of the locative is still not complete; examples can be found of its use, both triggered by prepositions (1a) and independently (1b), in plural sometimes merged with the dative (1c) and genitive (1d). Compare *komu-jen-ne* [people-OBL.PL-DAT] 'to people' and *gräij-en-ne* [horse-OBL.PL-GEN] 'of horses'.

(1) a. *Maxkar* *men-ne*
among we.OBL-LOC
'Among us'

b. *Nās* *līn-en-ne* *sēni*
NEG book-OBL.PL-LOC nowhere
'Was not registered anywhere'

c. *me* *hin* *phen-j-ommas* *tern-e* *komu-jen-ne*
I be.PRS.3SG say-PRET-1PL young-PL people-OBL.PL-DAT
'I have said to young people'

d. *čer-d-e* *gräij-en-ne* *čyöp-i*
do-PRET-PL horse-OBL.PL-GEN trade-NOM.PL
'(they) did horse trade'

The Romani genitive is a boundary case between derivation and inflection. Similar to Indic languages, it is connected with interpretational dilemmas caused by its adjectival agreement with the head nouns, which distinguishes it from the rest of the case paradigm. In (2a) *kent-os,* is a masculine, so that the modifying genitive *sikjibosko* ends in *-o* like adjectives in masculine, but in (2b) *stranna* is a feminine, so that the genitive *xyönoski* governed by it ends in *-i*. The rest of the examples represent masculine NOM.PL (2c), feminine NOM.PL (2d), OBL.SG (2e) and OBL.PL (2f).

(2) a. *sikjib-os-k-o* *kent-os*
teaching-OBL.SG-GEN-M.NOM.SG child-NOM.SG
'disciple'

b. *xyön-os-k-i* *strann-a*
sea-OBL.SG-GEN-F.NOM.SG beach-NOM.SG
'seaside'

c. *sikjib-os-k-e* *kent-i*
teaching-OBL.SG-GEN-PL child-NOM.PL
'disciples'

d. *patri-en-g-e* *sāl-a-k-e* *fest-i*
blade-OBL.PL-GEN-NOM.PL hall-OBL.SG-GEN-PL feast-NOM.PL
'Feasts of Tabernacles'

e. *sikjib-os-k-e* *kent-os-ke*
teaching-OBL.SG-GEN-M.OBL.SG child-OBL.SG-DAT
'to the disciple'

f. *sikjib-os-k-e* *kent-en-ge*
teaching-OBL.SG-GEN-PL child-OBL.PL-DAT
'to the disciples'

The status of the Romani genitive has been controversial because of its adjectival agreement. Pott (1844–1845) speaks about the "so-called genitive" (*sogenannter Genitiv*). Likewise, Sampson (1926: 85) points out that, in line with many modern Indic languages, Romani lacks a true genitive, but instead has adjectives that agree with their head nouns in gender, number and primary cases (nominative or oblique). The genitives could be thus regarded as mixed categories (Haspelmath 1996). Some scholars see the genitive as a secondary case and interpret the adjectival agreement as a result of Suffixaufnahme. Moravcsik (1995: 452) defines Suffixaufnahme as: "a pattern, where an attributive nominal carries two distinct case markers: one appropriate to its own function as an attributive, and the other appropriate to the function of the NP that includes both the attributive and the head" (see also Koptjevskaja-Tamm 2000).

The Suffixaufnahme was regular until the mid-1900s but has de facto disappeared from modern FR. Valtonen (1968: 163) pointed out as early as the 1960s that the genitives used to generally agree with their head nouns in gender and number, but at that time usually in number only. During the latter half of the 20th century, differences between idiolects were significant, and this was also reflected in written Romani. Whereas the genitives showed almost regular gender, number and case agreement with their head nouns in the manuscripts by Peltosalmi and Temo (in the 1970s), no more than one-fourth of the genitives showed gender agreement in Koivisto's (1994) dictionary, and number agreement was completely lost. Figure (10) compares the endings of the Genitives that modify feminine singular nouns in written sources from different eras. The data include Kemell's and Reinholm's lemmas in Thesleff's (1901) dictionary; Thesleff's (1901) own lemmas; Jalkio's journal articles and religious songs; the lemmas extracted from vocabularies by Kronqvist, and Peltosalmi and Temo; and Koivisto's (1994) lemmas. Figure (11) compares the endings of genitives that modify the plural forms of nouns in the same sources except Kronqvist.[2] *-e* and *-i* vary as feminine singular endings, while plural forms of genitives also end in *-e* and *-i*.

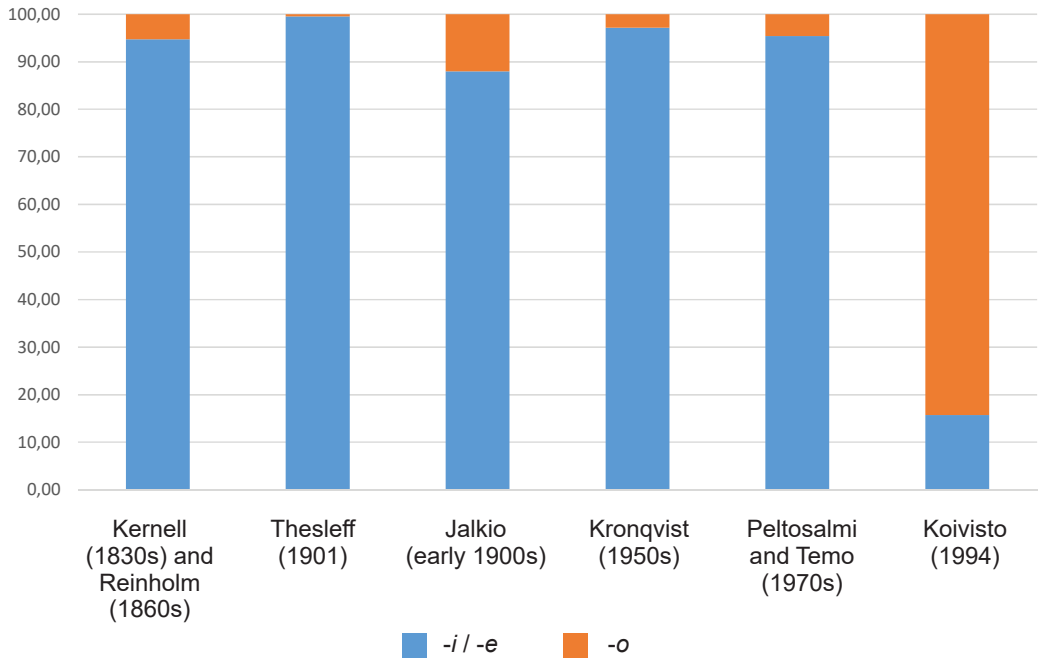


Figure 10. Suffixaufnahme in written Romani 1860s–1994, genitives modifying feminine nouns

[2] Kronqvist's data contain no relevant examples.

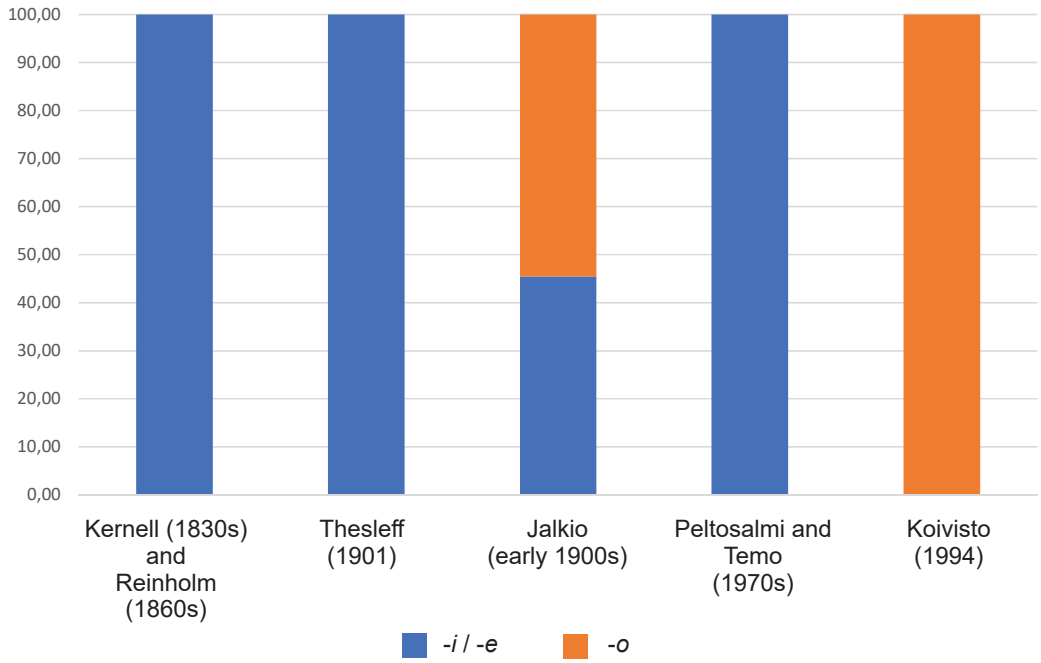


Figure 11. Suffixaufnahme in written Romani 1860s–1994, genitives modifying plural nouns

The loss of Suffixaufnahme in FR has resulted in an important typological change, causing the genitive to become a "true" case, encoded by the fossilized monomorphemic suffix *-ko*, cf. *daij-es-ko phāl* [mother-OBL.SG-GEN brother(.M)] 'mother's brother', *lesko phēn* [he.GEN sister(.F)] 'his brother', *bib-ja-ko čāv-e* [aunt-OBL.SG-GEN son-NOM.PL] 'aunt's son', in which the suffix showing adjectival agreement has become decategorialized and reanalyzed as part of the genitive marker. Thus, the genitive became morphologically analogical to the rest of the secondary cases.

3.2 Simplification of the noun classes

Signs of analogical levelling of nominal classes have been visible since the latter half of the 19th century, when the most prominent tendencies of intraparadigmatic levelling were formed: the collapse of athematic morphology and the expansion of the thematic *-o*-masculine paradigm. Similar tendencies have been attested in German Sinti and Welsh Romani (Elšík 2000: 23–24; Matras 2002: 84). In FR, the earliest examples of these tendencies are in Reinholm's (1860) extensive notes on Finnish Romani dating back to the 1860s: *kouv-a-tar* > *kouv-es-ta* [quarrel-OBL.SG-ABL] 'from the quarrel' and *čibb-a* > *ꞑibb-e* [language-NOM.PL] 'languages' and *dju-ja* > *dju-je* [woman-NOM.PL] 'women'. In Jalkio's data from the beginning of the 20th century, the analogical levelling of nominal classes was extensive and manifested itself primarily as the expansion of thematic formants into athematic nouns, e.g., in OBL.SG *-os*, *-is* > *-es: lyön-os-tar* > *lyön-es-tä* [salary-OBL.SG-ABL] 'from the salary', *onn-os-k-o* > *onn-es-k-o* [spirit-OBL.SG-GEN-M.SG] 'of the spirit', *komun-is* > *komun-es* [human_being-OBL.SG] 'person'; in NOM.PL *-i* > *-e*: *gong-i* > *gong-e* [time-NOM.PL] 'times', *kent-i* > *kent-e* [child-NOM.PL] 'children', *ōsn-i* > *oosn-e* [donkey-NOM.PL] 'donkeys', *valgōs-i* > *valgoos-e* [shepherd-NOM.PL] 'shepherds'. In NOM.PL, the suffixes *-e* substituted also the suffixes *-a* of thematic nouns, e.g., *māl-a* > *maal-e* [friend-NOM.PL] 'friends', *thān-a* > *thaan-e* [place-NOM.PL]

'places' (but *dabb-a* [wound-NOM.PL] 'wounds', *vast-a* [hand-NOM.PL] 'hands'). FR still retained distinct OBL.SG forms for both genders, but there were some exceptions such as *hārni-ja-k-o* > *häärn-es-k-o* [star-NOM.SG-GEN-M.SG] 'of the star', *rigg-a-tar* > *rigg-es-ta* [side-OBL.SG-ABL] 'from the side'.

The analogical levelling of nominal classes became more widespread and diversified during the second half of the 20th century. All tendencies that had begun during the 19th century continued and intensified; new tendencies emerged. Figure 12 illustrates some tendencies of analogical levelling of nominal classes based on religious texts. As a new tendency, the suffix *-a-* of OBL.SG of athematic feminine nouns also began to substitute for *-os-* in abstract nouns, e.g. *bolib-os-* > *bolib-a-* [world-OBL.SG] (cf. *skool-a-* [school-NOM/OBL.SG]). Thus, *-os-* replacing the suppletive suffix *-mas-* became itself an intermediate stage of the historical development of abstract noun inflection.

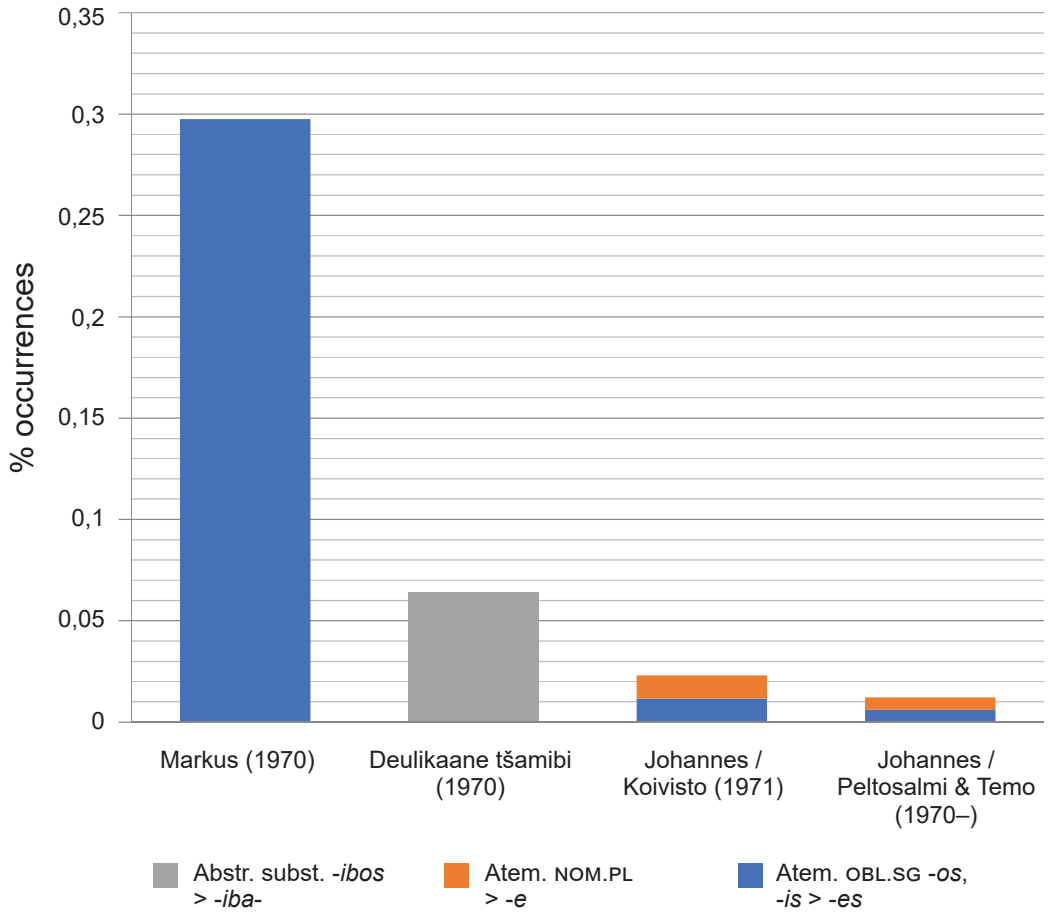


Figure 12. Tendencies of analogical levelling of nominal classes in religious texts (Granqvist 2012)

In modern spoken FR, the spread of certain nominal suffixes and the disappearance of othershas resulted in a large number of new paradigms. Granqvist (2007: 369–371) documents 31 new nominal paradigms. Most of the changes affecting the nominal classes can be perceived as a result of four tendencies of change and their combined effects; all these tendencies involve the relations between oblique and nominative plural markers with the highest frequencies of use (Figure 13).

Figure 13. Primary tendencies of analogical leveling of nominal paradigms

3.3 Simplification of verb derivation

Derived verbs constitute two main groups in FR: transitives and intransitives. It is common to FR and Sinti that the derivative morpheme modifies valence and adapts loan verbs. Unlike Northeastern Romani dialects, there are no separate loan verb markers such as *-in-*, e.g. *dum-in-* 'think'. The derivational morphology of verbs in FR has undergone different tendencies of change, resulting in a decreased number of verb classes (Pirttisaari 2002, 2003, 2004, 2005). At the 20th century, only three verb classes remained productive: the primary verbs (1/10 of Koivisto's verb lemmas), transitives derived in *-av-* (2/3 of verb lemmas) and intransitives in *-uv-* (1/4 of verb lemmas) (Valtonen 1968: 127–129; manuscript.; Granqvist 2002, 2005, 2007: 285; Pirttisaari 2003).

a) Changes in primary verbs
Primary verbs constitute a volatile class. Already in Sinti, a significant part of inherited primary verbs had moved to the classes of derived transitives: *bikin-* 'sell' > *bikin-*, *bikr-*, *bikerv-*; *inger-* 'carry' > *ligêrv-*, *ligêr-*; *inker* 'hold' > *rikêrv-*, *riker-*, *rik-*; *xač-* 'burn' >

xačêr-, *xačêrv-*; *vraker-* 'speak' > *rakêr-*, *rakêrv-* (Sinti forms Romlex), and they are FR formed in *-av-*: *biknav-*, *(r)igav-*, *(r)ikkav-*, *xačav-* and *rakkav-*. More primary verbs have moved to transitives as a result of devel-opments in FR. In modern FR, for instance *phord-* 'blow' ~ *phordav-*, *stāv-* 'walk' ~ *stāvav-*, *sterd-* 'pull' ~ *sterdav-* vary (Pirttisaari 2002: 511).

b) Changes in transitives
Many derivative suffixes that form transitives have lost their productivity in favor of *-av-*. In FR, the set of historical derivative suffixes forming transitives comprised *-av-*, *-ev-*, *-iv-*, *-arv-*, *-erv-* and *-alv-*. The suffix *-av-/-ev-*[3] cognate with Western Sinti (Barbara Schrammel-Leber, p.c. August 29, 2015) remained productive in FR as *-av-* (and *-äv-* as a result of suffix harmony); only a few traces of *-ev-* are found in language documents, and *-iv-* occurs sporadically. *-av-* is found in all FR documents since Ganander (1780), e.g., *drabaw-a* [read-PRS.1SG] 'I read' (Ganander, *drabav-es* [read-PRS.2SG] 'you read'. *anjav-* 'bathe', *garav-* 'hide', *rakkav-* 'speak', *praatav-* 'babble', *undrav-* 'wonder'.

The suffixes *-ar-/-arv-/-er-/-erv* cognate to Eastern Sinti (Barbara Schrammel-Leber, p.c. August 29, 2015) began to pass by early. Ganander (1780) had the syncopated form *aker-av-a* > *akr-aw-ān* [speak-PRS.1SG-FUT/1SG] 'I speak', from which initial *r-* was then elided as a result of an FR own innovation. Reinholm (1860s) had *akker-av-a* [speak-PRS.1SG-FUT] 'I speak'. Thesleff (1901) and Jalkio still retained *čingarv-* 'hurt', *čungarv-* ~ *čungerv-* 'spit', *igerv-* 'carry' and *phagarv-* ~ *phagerv-* 'break'; Kronqvist (1950s) only had *čingarv-* 'hurt'; Peltosalmi and Temo (1970s) no longer contained forms in *-arv-* and *-erv-*; *čingrav-*, *čungrav-*, *(r)igav-* and *phagav-* were used instead. The suffix *-alv-* is limited to verbs derived from nouns and secondary adjectives, e.g., *džōr* 'strength' > *džorjalv-* 'strengthen', bar 'mark' > *barvalo* 'rich' > *barvalv-* 'enrich'.

c) Changes in intransitives
FR has retained the class of derived intransitives, unlike Sinti. Mainly denominal and deadjectival verbs are derived using the suffixes *-uv-/-ov-*. Unlike in many other Romani dialects, there is no synthetic passive (Sampson 1926: 214). In addition, FR lacks productive means to form deverbal intransitives, while there are some verb pairs such as *naxx-* 'lose' – *naxuv-* 'escape', *stār-* 'fish' – *starjuv-* 'hook', *traxx-* 'fear' – *traxuv-* 'scare', *xunn-* 'hear' – *xunjuv-* 'be heard'. Historically, intransitives have been formed in FR by means of three suffixes *-uv-* (*-ov-*, also *-yv-*, *-öv-* as a result of vowel harmony), *-urv-* and *-ulv-*. Currently only *-uv-* is productive[4]; *-urv-* and *-ulv-* are analogical with *-arv-* and *-alv-* and have mostly been replaced by *-uv-*. As a new tendency, intransitives and transitives have begun to merge, e.g., *byrjuv-* ~ *byrjav-* 'begin', *orkuv-* ~ *orkav-* 'endure', *vandruv-* ~ *vandrav-* 'wander'.

3.4 Syncretism in person inflections

The person inflections of verbs have manifested tendencies of analogical leveling both as language-internal and contact-induced change. In the indicative and conditional, present tense singular 1st and 2nd persons tend to be syncretic, (3a), e.g. *me rakkavā* 'I speak',

[3] Holzinger (1993: 108) only mentions this causative marker.

[4] The vowels of the intranstive marker show variation *-o-* ~ *-u-*, due to vowel harmony also *-o-* ~ *-ö-*, *-u-* ~ *-y-*, e.g., *hajov-* ~ *hajuv-* 'understand', *yltov-* ~ *yltöv-* 'reach' (Thesleff 1901), *byrjuv-* ~ *byrjyv-* 'begin'. *-o-*forms occur frequently in the speech of the Ostrobothnian Roma (Henry Hedman, p.c. March 14, 2003).

vs. *tu rakkavā* ‘you speak’. The earlier evidence of this only dates back to the latter half of the 20th century. Due to contact with Finnish, the 3rd person plural is often syncretic with the 3rd person singular (3b), e.g. *jou čērela* ‘he does’ vs. *jōn čērela* ‘they do’; this is an old tendency in FR. More arbitrarily, inflectional homonymy is attested on the 2nd person singular and the 1st person plural (3c), *tu rakkaveha* ‘you speak’ vs. *ame rakkaveha* ‘we speak’.

(3)

a.	SG	PL
1	*-a-*	*-ah-*
2		*-en-*
3	*-el-*	

b.	SG	PL
1	*-a-*	*-ah-*
2	*-eh-*	*-en-*
3	*-el-*	

c.	SG	PL
1	*-a*	*-eh-*
2	*-eh-*	*-en-*
3	*-el-*	

3.5 New infinitive

In modern FR, the indicative and subjunctive are primarily distinguished from each other by the presence of the future marker *-a* in the indicative, e.g. *rakkav-el-a* [speak-PRS.3SG-FUT/IND] ‘he/she speaks’ and its absence in the subjunctive *rakkav-el* [speak-PRS.3SG] ‘he/she speaks’. In addition, the person inflections of indicative and subjunctive started differentiating from each other during the latter half of the 20th century. The first step of the simplification of person inflection was the loss of the separate first-person plural marker *-as* in the subjunctive even from the most conservative idiolects by the 1970s–1980s, but from a majority of idiolects much earlier. As a result, the suffixes *-en/-n* were extended to the entire plural. See example (4a). The paradigm of the subjunctive was further simplified during the latter half of the 20th century. In many idiolects, the third person singular suffix *-el/-l-* was generalized to the entire singular (occasionally though, the second person singular suffix *-es/-s*) (4b). The last example (4c) represents the so-called “new infinitive” (Elšík & Matras 2006: 127–130), in which person inflection of modal finite verb complements have been reduced to the point that the third person singular suffix has been generalized into the entire paradigm (cf. Brandt-Taskinen 2001: 56–57). However, the new infinitive does not occur in any written sources of FR, but first in spoken FR data recorded at the end of the 20th century, and rarely even then. The new infinitive has been documented in some other European Romani dialects as well (Boretzky 1996; Matras 2002: 161). Also in these, the third person singular functions as infinitive, sometimes second or third person plural (Boretzky 1996; Matras 2002: 161).[5]

(4)

a.	SG	PL
1	*-a*	*-en/ -n*
2	*-es/-s*	
3	*-el/-l*	

b.	SG	PL
1	*-el/-l*	*-en/ -n*
2		
3		

c.	SG	PL
1		*-el/ -l*
2		
3		

Another set of phenomena distinguishing indicative and subjunctive are contractions limited to indicative only[6] and new kinds of systematic homonymy that emerged in indicative inflections. In Jalkio’s data in particular, the plural suffixes *-ēn-a/-en-a-*

[5] Cech & Heinschink (2001) mention in Slovenian and Istrian Dolenjski Roma another type of infinitives in *-i*, e.g., *vakeri* ‘speak’.

[6] Jalkio has one contracted subjunctive form: *kamm-en-a te būro-n* [want-PRS.3P-IND COMPL live-PRS.3PL] ‘want to live’.

were extended to the third person singular, e.g., *do-uva na hyöv-en-a* [it-NOM.SG NEG need-PRS.3P-IND] 'it is not needed', *v-en-a rōligib-a* [come-PRS.3P-IND peace-NOM.SG] 'peace will come'.

The complementizer *te* used to be obligatory until the beginning of the 1900s (e.g., *Me chamm-a tej cha-w* [I want-PRS.1SG COMPL eat-PRS.1SG] 'I want to eat' (Ganander), *fedde hin te d-el sar te l-en* [better is COMPL give-PRS.3SG than COMPL get-3SG] 'it's better to give than to get' (Reinholm), but became optional or was lost by the mid-1950s, e.g., *v-ēl-a fā̈rdav-el* [come-PRS.3SG-IND travel-PRS.3SG] 'will travel' (Valtonen), *ame l-ah-as d-en svaariba* [we get-PRS.1PL-COND give-PL answer] 'we would be allowed to give an answer' (Koivisto).

4. Contact-induced changes

The earlier contact influences of German, Danish, and Swedish are visible in FR, predominantly in lexical domains and to a limited extent in phonology (the vowels /ü, ø, æ/ borrowed with Germanic loanwords, the ties of stress and quantity; Prokosch's law (Venneman 1988)). However, by the end of the 19th century, the Roma had already forgotten their active knowledge of Swedish (Thesleff 1899) and even earlier, their knowledge of German. Their contact with Swedish was re-established as a result of subsequent migration of many Roma to Sweden since the 1960s; however, this has triggered no visible late influences on FR. The sole close contact language of FR thus remained Finnish.

Finnish lexical items have been borrowed since early times. Ganander's essay (1780) contained a few Finnish loan words, and a more significant number of Finnish borrowings are in Reinholm's notes and other 19th-century sources. Despite its early onset, there has only been limited lexical transfer from Finnish. This might be attributed to the secret language functions of FR, or to the linguistic dominance of Finnish among the Roma since the 19th century and subsequent preference for codeswitching over loan adaptation as a strategy to fill lexical gaps. Written sources of FR from the late 18th and 19th century suggest that most of the transfer of Finnish phonological principles and rules had already taken place or were productive by the end of the 19th century: this includes the polarization of voiced stops into voiceless, long vowel diphthongization, vowel harmony and svarabhakti vowels (Granqvist 2013b). The simplification of initial consonant clusters is probably the only phonological change triggered by contact with Finnish that is first documented after the 1950s, e.g. /stranna/ > [ranːa] 'strand', /drann-/ > [ranː-] 'bite'; for a comprehensive treatment of the phenomenon, see Granqvist (2007: 194). Finnish has had a profound influence on the syntax of FR, as word order and many morpho-syntactic patterns tend to be copied from Finnish (Granqvist 2014b). In this article, the discussion on contact-induced changes will concentrate on morphological borrowing in FR (Granqvist 2014b).

4.1 Pattern Replication

I have divided the discussion on morphological borrowing in two main sections following the sub-division by Matras & Sakel (2007) into *pattern replication* and *matter replication*; the subdivision is illustrated in Figure 14. Pattern replication refers to the borrowing of morpho-syntactic patterns and replicating using Romani's own resources,

e.g., morphemes that are inherited from Indo-Aryan or belong to later layers of historical contact languages, such as Byzantine Greek, Slavonic lan-guages, Hungarian, Middle Low German, Middle High German, Danish, Late Old Swedish or Swe-dish. Heath (1984: 367) speaks about "pattern transfer". Pattern replication is extremely frequent in FR: it has resulted in morpho-syntactic patterns that constitute obligatory parts of the structure of FR. Early 19th century (documented in 1817–1830s) tendencies of pattern replication involved the development of case-licensing principles in FR and Finnish-like passive constructions after the collapse of the conservative analytical passive found in most European Romani dialects. Matter replication refers to borrowing of Finnish morphological markers into Romani. It is predominantly late and rare. Interestingly, and against the universal constraints of morphological borrowing, inflectional morphemes in FR are more prone to borrowing from Finnish than derivational ones; this will be discussed in Section 4.2.

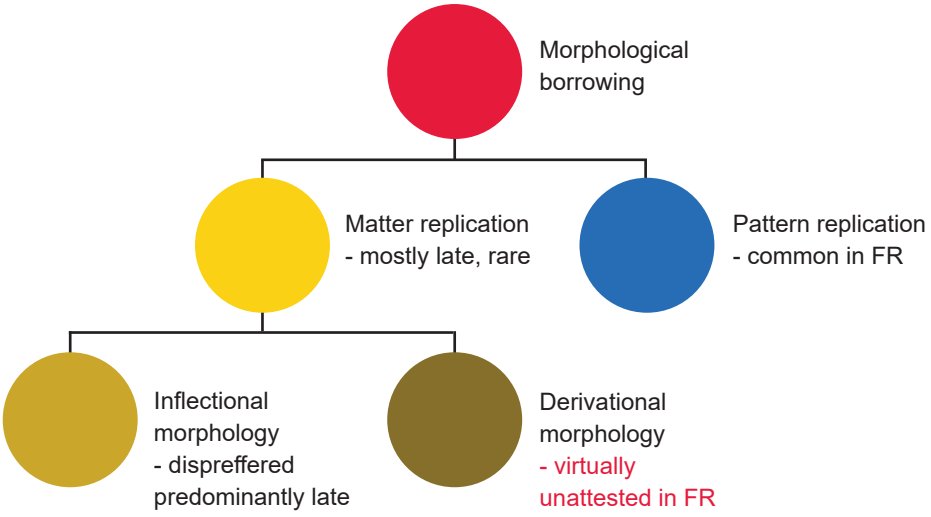


Figure 14. Typology of morphological borrowing in FR (cf. Matras 2007)

As examples of pattern replication, I will briefly discuss in the following sections 1) the loss of definite articles, 2) the typological change of FR into a postposition language and the development of ambipositions similar to those in Finnish, and 3) the development of the analytical past tenses perfect and pluperfect and the analytical future. The first signs of all three phenomena discussed here date back to the 19th century, but most of the significant developments took place during the 20th century.

4.1.1 Loss of definite articles

FR had a category of prenominal definite articles that resembled the articles in Sinti as well as the reconstructed articles of Northeastern Romani (Table 3). The masculine nominative singular form was *o*; in feminine nominative singular, two forms competed: *e* similar to Northeastern Romani and *i* similar to Sinti, e.g., *e ɥae* 'the girl', *e touverissa* 'the ax' (Reinh. = Henrik August Reinholm's notes), *ibibi* 'the aunt', *ijak* 'the eye' (Gan.). In nominative plural, *o* and *i* similar to Northeastern Romani and Sinti were in free variation, e.g. *o gaje* 'the non-Roma', *i buttia* 'the jobs' (Reinh.), *e* was used in all forms of oblique, e.g. *e gresta* 'from the horse', *e ɥaejinge* 'to girls' (Reinh.).

Table 3. Forms of definite article

		Finnish Romani (Reinholm)		Sinti (Holzinger 1993: 46)		NE dialects (reconstruction Boretzky 2000; ref. Tenser 2009: 110)	
		Nominative	Oblique	Nominative	Oblique	Nominative	Oblique
Sg.	Masc.	*o*	*e*	*o*	*i*	*O*	*e*
	Fem.	*e ~ i*	*e*	*i*	*i*	*E*	*e*
Pl.		*o ~ i*	*e*	*i*	*i*	*O*	*e*

While the first signs of the decay of the definite articles were visible in the written sources from the latter half of the 19th century, definite articles were already virtually lost by the first half of the 20th century, except for prepositions.[7] In Jalkio's materials (early 1900s), remnants of articles occurred with prepositions, e.g., (masc. *o*): *ar-o onnos* [in-ART.M.SG spirit-NOM.SG] 'in the spirit', *ar-o deul-en-ne* [in-ART.M.PL heaven-OBL.PL.LOC] 'in the heavens'; (fem. *i*): *ap-i himl-a* [on-ART.F.SG sky-NOM.SG] 'in the heaven'; (plural *e*): *kaj-e pes-k-e sikjiboskekent-en* [near-ART.PL REFL.OBL.SG-GEN-PL disciple-OBL.PL] 'near his disciples', *naal-e komu-ja* [in.front.of-ART.PL people-NOM.PL] 'in front of the people'. Ariste's slightly later materials (ca. 1940s) were characterized by a masculine takeover: the feminine and plural forms of articles were lost even with prepositions: e.g. *aro butt-i* [in-ART.M.SG work-NOM.SG] 'at work', *ar-o xlitt-a* [in-ART.M.SG sleigh-NOM.SG(.F)] 'in the sleigh', *ar-o tšheer-en-ne* [in-ART.M.SG house-OBL.PL-LOC] 'in the houses' , *ap-o phuu-ja* [on-ART.M.SG countryside-NOM.PL] 'in the country-side'.

A similar kind of decay and loss of definite articles has also taken place in North Russian, Baltic and Polish Northeastern Romani dialects, triggered by contract with the articleless Slavonic languages Polish and Russian. (Matras 1999: 9–10; Boretzky 2000: 34; Wentzel 1964; Tenser 2005; see also Baló & Bodnárová 2023 on masculine takeover and the weakening of gender opposition on Romani in Hungary). According to Tenser (2008: 110), definite articles have been completely lost in Latvian Fandari and Pškov Xaladytka, both belonging to the so-called Belarus sub-group of Northeastern Romani dialects, and in Lotfitka and Estonian Romani, both belonging to the so-called Latvian sub-group of Northeastern Romani dialects. Of the Northeastern dialects, only Russian and Polish Xaladytka and Ukrainian Ghympeny retain definite articles (Tenser 2008: 112). In some of the Northeastern Romani dialects, bound remnants of articles are retained with prepositions as in FR during the first half of the 20th century, e.g. *and-o kher* [in-ART.M.SG house(.M)] 'in the house', *and-e fabryk-a* [in-ART.F.SG factory-NOM.SG] 'in the factory', *pal-e man* [near-OBL.I.OBL] 'near me' (Tenser 2008: 110). A simultaneous development with the decay of definite articles was, in FR, the weakening of the gender opposition in nouns.

[7] The sole exception in Jalkio's data: *tserav-el i khangar-i sonak-es-ta* [do-PRS.3SG ART church-NOM.SG gold-OBL.SG-DAT] 'to make a church out of gold'. Actually, the idiolects show a significant amount of variation. Even at the end of the 1980s, definite articles were in use to a limited extent by some Roma.

4.1.2 Change into a postposition language and development of ambipositions

Finnish Romani has undergone an important typological change from a preposition language into a (mainly) postposition / ambiposition language. The old spatial / temporal prepositions remained in use in FR until the beginning of the 20th century and sporadically in conservative idiolects, at least until the 1970s, e.g., *angla mange* 'in front of me', *naāli puorta* 'in front of the gate', *pach taffla* 'at the table', *prāli kuorma* 'on the carriage' (Reinh.); *pālo pāluno* 'afternoon', *teli falda* 'dependent, subordinate' (Thesl.); *maǩkar penne* 'among them', *naalo maan* 'in front of me', *paali čyöpibosko taffla* 'behind the sales desk', *praalo phuu* 'on Earth', *trystalo boliba* 'around the world', etc. (Jalk.). Valtonen (1968: 167) lists the prepositions *aŋglo* 'in front of, before', *nālo* 'in front of, before', *paxo* 'near' and *prālo* 'on(to)'. Even in the translation of John's Gospel to Romani by Peltosalmi and Temo, examples such as *angla leste* 'in front of him' are attested, but in free variation with postposition phrases.

The first documents of postpositions assigning genitive, replicating the Finnish pattern, are in Reinholm's notes; in *ʒaki ēstä* 'for the girl', the Finnish postposition *ēstä* (< Finnish *edestä* 'from the front') is preceded by the genitive *ʒa-k-i* [girl.OBL.SG-GEN-F.NOM.SG] 'for the girl's sake'. The use of postpositions was generalized by Ariste's times (1930s–1940s): there are plenty of examples of pattern replication, such as: *komu-jen-go maxkar* [people-OBL.PL-GEN among] 'among people', *tukko nālal* [you.GEN in.front.of] 'in front of you', *kuti tī-a-ko pālal* [little time-OBL.SG-GEN after] 'after a little time', *lesko trystal* [he.GEN around] 'around him', *vondr-os-ko tēlal* [bed-OBL.SG-GEN under] 'under the bed'.

Furthermore replicating the Finnish pattern, the apdositions *džinom ~ đženom* 'through', *maxkar* 'among', *nāl* 'in front of; before', *nēr* 'near', *perdal* 'by', *prāl* 'on(to)' and *trystal(o)* 'around' have been developed into ambipositions having their complements in the genitive as postpositions and in the nominative or oblique as preposition, e.g., *bolib-os-ko trystal* [world-OBL.SG-GEN around] '(in a circle) around the world' and *trystalo bolib-a* [around world-NOM.SG] 'around the world'; thus the different cases of complements give the PPs different meanings: in *bolib-os-ko trystal* 'around' has to be understood very literally as a circle with the world as its centre, while in *trystalo bolib-a* 'around' is a vague region. The distinction in boundedness is similar to Finnish (Hakulinen et al. 2004: § 1498).

The only words that function solely as prepositions are *aro* 'in(to)', *apo* 'on(to)' and *kajo* 'towards', which are historically but not synchronically bimorphemic, consisting of the preposition itself and a remnant of a definite article (e.g. *ap o*, *ap i*, *ar e* etc.).

4.1.3 Analytical tenses

Due to the great restructuring of the verb morphology in FR (18th–19th centuries, e.g. Granqvist 2012, 2013a, 2013b), involving the collapse of the morphological aspect and the merger of the preterit and the pluperfect, FR was left with a single past tense, which morphosemantically covered the functions of the imperfect, preterit and pluperfect and also corresponded to the Finnish perfect. This is illustrated by Examples (5a–c), taken from Reinholm's materials and translated based on Reinholm's glosses in Swedish.

(5) a. preterit/imperfect:

me	*dján-id-om(-)as*
I	know-PRET-1SG(-REM)

'I knew'

b. perfect:

Me	*presa-d-omm(-)as*	*bút.*
I	pay-PRET-1SG(-REM)	a_lot

'I have paid a lot'

c. pluperfect:

P-ies	*p-es-k-i*	*vett-a.*
drink-3SG	REEL-OBL.SG-GEN-F.NOM.SG	sanity-NOM.SG

'He drank away his sanity.'

The past tense system was complemented with an analytical Perfect and Pluperfect by the end of the 19th century. The pattern was replicated from Finnish forms consisting of olla 'to be' and a participle (*ole-n puhu-nut* [be-PRS.1SG speak-PTCP] 'I have spoken', *ol-i-tte puhu-neet* [be-PST-2PL speak-PTCP] 'you had spoken'). Thesleff's song texts (approx. the 1890s) were the first data containing the new phrasal verbs, which consisted of the auxiliary *s-/h-* 'to be' and an athematic participle of the matrix verbs, as in (6). However, such analytic tenses first became frequently used in the materials dating back to the beginning of the 20th century.

(6)

Voj	*s-am*	*han-ime*	*ta*	*lis-ime.*
oh	be-PRS.1PL	miss-ATHEM.PTCP	and	suffer-ATHEM.PTCP

'Oh, we have missed and suffered.'

In Jalkio's materials, two types of constructions competed, one with an athematic and one with a thematic participle of the matrix verb, as shown in Table 4.

Table 4. Types of analytical past tenses in Oskari Jalkio's materials

	s-/h- + ATEM.PTCP		*s-/h-* + TEM.PTCP	
PERFECT	*s-om*	*āh-imen*	(no examples)	
	be-PRS.1SG	be-ATHEM.PTCP		
	'I have been'			
	hin	*par-imen*		
	be.PRS.3SG	change-ATHEM.PTCP		
	'has been changed'			
PLUPERFECT	*s-omm-as*	*tryst-imen*	*s-as*	*au-l-o*
	be-1SG-REM	go.around-ATHEM.PTCP	be-PRET.3SG	come-PRET-M.SG
	'I had travelled around'		'(he) had come'	
	s-as	*rakk-imen*	*s-as*	*trād-id-e*
	be-PRET.3SG	speak-ATHEM.PTCP	be-PRET.3PL	drive-PRET.-PL
	'I had spoken'		'(they) had driven'	

The opposition of diathesis was neutralized in Jalkio's perfect and pluperfect, because Romani participles can be both active and passive. Examples (7a–b) are in active and (7c) in passive. The passive was distinguished from the active form by the lack of an explicit subject, as in Finnish passive constructions (on Finnish, see Hakulinen et al. 2004: 1254–1255). However in Finnish, the implicit subject is plural, e.g., *Ennen oltiin vähään tyytyväisiä* [before be.PASS.IMPF little.ILL satisfied.PART.PL] 'Before we used to be satisfied with little' but in FR, the implicit subject was singular, as indicated by the participle phello (Hakulinen et al. 2004: 1256).

(7) a. *s-om* *drab-imen*
be-PRS.1SG read-ATHEM.PTCP
'i have read'

b. *Dād* *s-as* *d-īl-o.*
father be-PRET.3SG give-PRET-M.SG
'the father had given.'

c. *so* *s-as* *phel-l-o* *len-ge* *kent-os-ta*
what be-PRET.3SG say-PRET-M.SG they.OBL-DAT child-OBL.SG-ABL
'what was told them about the child'

The modern analytical tenses perfect and pluperfect had become stable by the second half of the 20th century. The auxiliary was *s-/h-* 'to be' as in the previous period, but the preterit replaced athematic participles in the matrix verbs. The new forms are thus no longer direct replications of the Finnish pattern, but FR's own innovations. In perfect the auxiliary is in present tense and in past tense in pluperfect.

Two types of constructions compete in the analytical tenses: in the more complex one, both the auxiliary and the matrix verb inflect in persons (8a), and in the simpler one, the auxiliary is generic (3SG always), but the matrix verb is person-inflected (8b).

(8) a.

	SG		PL	
1	*som*	*rakkadom*	*sam*	*rakkadam*
2	*sal*	*rakkadal*	*san*	*rakkade*
3	*hin*	*rakkadas*	*hin*	*rakkade*

b.

	SG		PL	
1		*rakkadom*		*rakkadam*
2	*hin*	*rakkadal*	*hin*	*rakkade*
3		*rakkadas*		*rakkade*

In addition to the analytical past tenses, a periphrastic future *v-* 'come' + subjunctive had emerged by the beginning of the 20th century, replicating the Finnish verbal phrase consisting of *tulla* 'to come' as an auxiliary and infinitive III of the matrix verb, e.g. *tulen tekemään* [come.PRS.1SG do.3INF] 'I will do' as in Examples (9a–b).

(9) a. *Messias* *v-el-a* *te* *staav-el* *teele.*
Messias come-PRS.3SG-IND COMP step-PRS.3SG down
'Messias will step down.' (Jalk.)

b. *Oosn-e* *v-el-a* *panna* *te* *stakrav-el* *teele*
donkey-NOM.PL come-PRS.3SG-IND still COMP step-PRS.3SG down

t-i *rajiskiduitu,* *Bi* *ap-i* *da* *grubb-os* *v-el-a*
your-F.SG Court But on-F.SG this crip-NOM.SG come-PRS.3SG-IND

te *čerav-el* *i* *khangar-i* *sonak-es-ta…*
COMP do-PRS.3SG ART.F.NOM.SG church-NOM.SG gold-OBL.SG-ABL
'Donkeys will trample palace, but upon this crib, a church will be built.'

4.2 Matter replication

Universal constraints on morphological borrowing have been proposed in a number of typological studies (e.g. Moravcsik 1978; Thomason & Kaufmann 1988; Thomason 2001; Winford 2003). The generalizations are: frequency-based hierarchies (a majority, e.g., Haugen 1950; Heath 1984; Thomason & Kaufmann 1988; van Hout & Muysken 1994 etc.); implicational hierarchies (e.g. Moravcsik 1978; Matras 1998, 2002; Field 2002; Elšík & Matras 2006); or based on both frequency-based and implicational observations (Stolz & Stolz 1996; Ross 2001). The existence of these constraints has been rejected by some other scholars, such as Campbell (1993). Some relevant generalizations about the borrowability of morphology are summarized in Table 5.

Table 5. *Generalization on morphological borrowing* vs. FR.

Generalizations (Moravcsik 1978; Field 2002)	In FR
Unbound > Bound morphemes	Unbound > Bound morphemes (except clitics)
Derivational morphology > Inflectional morphology	Inflection > Derivation

Borrowing of Finnish inflectional morphology or replication of Finnish morphological matter is chronologically far later and rarer than pattern replication. Granqvist (2014) suggests that borrowing of Finnish inflectional morphology follows the timeline illustrated in Figure (15).

		Borrowed markers	Earliest source
Frequent ↕ Rare	Bound morphemes	Clitics	Ganander (1779)
		Noun inflection/ local cases (Ex. 1)	Reinholm (1860)
		Noun inflection/ grammatical cases (Ex. 2a-b)	[2nd half of the 20th century]
		Verb inflection/ persons endings (Ex. 3b-c)	
		Noun inflection/ modus markers (Ex. 4a-c)	
Frequent ↕ Rare	Unbound morphemes	Negation particle	Ariste (1940)
		Negation verb	[2nd half of the 20th century]
		Modals and auxiliaries	

Figure 15. Borrowing of Finnish inflectional morphology (Granqvist 2014)

One of the central findings of Granqvist (2014) is that contrary to the universal constraints on morphological borrowing, borrowing of derivational morphology is in FR far more infrequent than that of inflectional morphology. Code-switching is instead used as a compensation strategy to fill gaps (Granqvist 2000). Sporadic examples of comparatives formed with the Finnish morpheme -mpi have been attested since Reinholm's notes (1860, e.g., *terne-mpi* [young-COMP] 'younger'). A few exceptions fill lexical gaps: FR lacks its own means to derive frequentative verbs, but *minhu-il-v-* 'tease' is formed in resemblance of the Finnish frequentative *vittu-ill-a* 'bully'. Double causatives are avoided in FR. Some Finnish verbs are borrowed with Finnish derivative morphology and adapted into Romani: *kasvatt|av-* 'educate, grow, bring up' < FI *kasvattaa*, *kulett|av-* 'transport' < FI *kuljettaa*. A small number of Finnish loanwords borrowed with Finnish morphology to the structure of FR are documented but no longer in use: *kukkasa* (Reinh. 1860) 'with flowers' < FI *kukka* (+ Suff. *-sa*), *nahgist* (Gan. 1780) 'diligently' < FI *nahkiasti*. Compounding is rare in FR; however, there are items with a compound modifier borrowed from Finnish: *ābislīn* Th. 'abc-book' < FI *aapis-* + *līn* 'book', *aikadžēno* 'grown-up man' < FI *aika-* and *džēno* 'man' and with a head borrowed from Finnish: *auripäi* 'outward' < *auri* 'out' + FI *päin* 'toward'.

5. Conclusion

In section 2, I compared surveys of competence in FR. I concluded that the surveys repeat the view that older Roma master Romani better than younger individuals. The differences between age groups are related to the late acquisition of Romani along with growing up to adulthood and socialization in the Roma community (Borin & Vuorela 1998: 60; Borin 2000: 75), but possibly also reflect the gerontocratic hierarchy of the Roma community. The surveys indicate a decrease of excellent and good competence in Romani over time, but virtually no change in the proportion of non-speakers. Apparently, the notion of good Romani competence varies over time and from person to person.

Section 3 dealt with language-internal changes. Five phenomena were discussed: the simplification of the case system and noun classes, the simplification of the verb classes and person inflections, and finally, the development of the so-called 'new infinitive'. While all these tendencies discussed aim at simplifying the language structure, they yield a significant amount of variation at the current intermediary stage of the language; this manifests itself clearly e.g. in the large number of noun classes attested in modern FR. Some of the changes result in added morphotactic transparency but increase phonological complexity by resulting in heavier forms.

Section 4, discussing contact-induced changes, showed that matter replication is later and occurs more rarely in FR than pattern replication, which is in some cases even obligatory. Matter replication in FR predominantly follows universal tendencies: free morphemes are more prone to be borrowed than bound morphemes (Moravcsik 1978; Field 2002); however, against the universal tendencies, Finnish inflectional markers are more easily borrowed to FR than derivational suffixes.

Abbreviations

1	1st person
2	2nd person
3	3rd person
ABL	ablative
ART	article
ATHEM	athematic
DAT	dative
F	feminine
FUT	future
GEN	genitive
ILL	illative
IMPF	imperfect
INF	infinitive
IND	indicative
INSTR	instrumental
LOC	locative
M	masculine
NOM	nominative
OBL	oblique
PART	partitive
PASS	passive
PL	plural
PRET	preterite
PRS	present tense
PTCP	participle
REEL	reflexive
SG	singular

Data sources

Granqvist, K. 1999. *Suomen romanikielen käänteissanasto. Reverse Lexicon of Finnish Romani.* (Kotimaisten kielten tutkimuskeskuksen julkaisuja 111). Helsinki: Kotimaisten kielten tutkimuskeskus.

Granqvist, K. (Ed.) 2014a. *Juho Peltosalmi ja Yrjö Temo. Suomi–romani-sanakirja ja Johanneksen evankeliumi.* Helsinki: Suomen Romaniyhdistys.

Hedman, H. 1996. *Sar me sikjavaa romanes. Romanikielen kielioppiopas.* Jyväskylä: Opetushallitus.

Koivisto, V. 1970. *Deulikaane tšambibi. Hengellisiä lauluja.* Forssa: Mustalaislähetys.

Koivisto, V. 1971. *Johannesesko Evankeliumos.* Helsinki: Suomen Pipliaseura.

Koivisto, V. 1982. *Drabibosko ta rannibosko byrjiba.* Helsinki: Ammattikasvatushallitus – Kouluhallitus.

Koivisto, V. 1987. *Rakkavaha romanes. Kaalengo tšimbako sikjibosko liin.* Helsinki: Ammattikasvatushallitus – Valtion painatuskeskus.

Koivisto, V. 1994/2005. *Romano-finitiko-angliko laavesko liin. Romani-suomi-englanti sanakirja. Romany-Finnish-English Dictionary.* (Kotimaisten kielten tutkimuskeskuksen julkaisuja 74.) Helsinki: Painatuskeskus.

Thesleff, A. 1901. *Wörterbuch des Dialekts der finnländischen Zigeuner.* (Acta Societatis Scientiarum Fennicae 29(6).) Helsinki: Finnische Litteratur-Gesellschaft.

Valtonen, P. 1970. *Markusesko evankeliumos.* Helsinki: Kristillisen Kirjallisuuden Seura.

Vuolasranta, M. 1996 [1995]. *Romani tšimbako drom.* 2nd edn. Jyväskylä: Opetushallitus.

Vuolasranta, M. & Hagert, A. & Majaniemi, P. & Huttu, H. 2003. *Romani tšimbako buttiako liin 1*. Helsinki: Opetushallitus.

References

Anttila, R. 1972. *Historical and Comparative Linguistics.* (Current Issues in Linguistic Theory 6). Amsterdam: John Benjamins.

Ariste, P. 1940. Über die Sprache der finnischen Zigeuner. *Õpetatud Eesti Seltsi Aastaraamat, Annales Litterarum Societatis Esthonicae* 1938/2. 206–221.

Bakker, P. 2020. Para-Romani Varieties. In Matras, Y. & Tenser, A. (eds.), *The Palgrave Handbook of Romani Language and Linguistics*, 353–386. Palgrave Macmillan.

Boretzky, N. 1996. The 'new infinitive' in Romani. *Journal of Gypsy Lore Society* 6. 1–51.

Boretzky, N. 2000. The definite article in Romani dialects. In Elšík, Viktor & Matras, Yaron (eds.), *Grammatical Relations in Romani: The Noun Phrase*, 31–63. Amsterdam: John Benjamins.

Borin, L. 2000. A corpus of written Finnish Romani texts. In O Croinin, Donncha (ed.), *LREC 2000. Workshop Proceedings. Developing language resources for minority languages: Reusability and strategic priorities*, 75–82. Athens: ELRA.

Brandt-Taskinen, P. 2001. *Suomen romanikielen verbikomplementit*. University of Helsinki. (Master's thesis.)

Campbell, L. 1993. On Proposed Universals of Grammatical Borrowing. In Aertsen, Henk & Jeffers, Robert J. (eds.), *Historical linguistics* 1989, 91–109. Amsterdam: John Benjamins.

Cech, P. & Heinschink M. 2001. A dialect with seven names. *Romani Studies* 11(2). 137–184.

Cortiade, M. 1991. Romani versus Para-Romani. In Bakker, Peter & Cortiade, Marcel (eds.), *In the Margin of Romani: Gypsy languages in contact*, 1–15. Amsterdam: Institute for General Linguistics.

Coseriu, E. 1974. *Synchronie, Diachronie und Geschichte*. (Internationale Bibliothek für allgemeine Linguistik. Band 3). München: Wilhelm Fink.

Dressler, W. U. 1977. Grundfragen der Morphonologie. Wien: Österreichische Akademie der Wissenschaften.

Elšík, V. 2000. Romani nominal paradigms: Their structure, diversity, and development. In Elšík, V. & Matras, Y. (eds.), *Grammatical Relations in Romani: The Noun Phrase*, 9–30. Amsterdam: John Benjamins.

Elšík, V. & Matras, Y. 2006. *Markedness and Language Change: The Romani Sample.* Berlin: Mouton de Gruyter.

Field, F. 2002. *Linguistic Borrowing in Bilingual Contexts*. Amsterdam: John Benjamins

Givón, T. 1985a. Language, function and typology. *Journal of Literary Semantics* 14(2). 83–97.

Givón, T. 1985b. Iconicity, isomorphism and non-arbitrary coding in syntax. In Haiman, John (ed.), *Iconicity in Syntax*, 187–220. Amsterdam: John Benjamins.

Granqvist, K. 2000. Intrasentential Codeswitching in the Speech of Finnish Roma: A Case Study. (Paper presented at the 5th International Conference of Romani Linguistics, Sofia, 14−17 Septem-ber 2000.)

Granqvist, K. 2002. Finnish Romani Phonology and Dialectology. *SKY Journal of Linguistics* 15. 61–84.

Granqvist, K. 2005. ROMTWOL: An implementation of a two-level morphological processor for Finnish Romani. In Schrammel, Barbara & Halwachs, Dieter & Ambrosch, Gerd (eds.), *General and Applied Romani Linguistics. Proceedings from the 6th International Conference on Romani Linguistics*, 150–162. München: Lincom Europa.

Granqvist, K. 2007. *Suomen romanin äänne- ja muotorakenne.* Helsinki: Yliopistopaino.

Granqvist, K. 2009. Mikä on erilaista romanien diskurssissa. In Idström, Anna & Sachiko, Sosa (eds.), *Kielissä kulttuurien ääni*, 206–222. Helsinki: Suomalaisen Kirjallisuuden Seura.

Granqvist, K. 2010. Two hundred years of Romani Linguistics. In Karttunen, Klaus (ed.), *Anantaṁ śāstram. Indological and Linguistic Studies in Honour of Bertil Tikkanen*. (Studia Orientalia 108), 245–265 Helsinki: Finnish Oriental Society.

Granqvist, K. 2012. Romanikielen historiaa Suomessa. In Blomster, Risto & Rekola, Tuula & Tervonen, Miika & Viljanen, Anna Maria (eds.), *Suomen romanien historia*, 272–287. Helsinki: Suomalaisen Kirjallisuuden Seura.

Granqvist, K. 2013a. Attritiosta Suomen romanikielessä. In Granqvist, Kimmo & Rainó, Päivi (eds.), *Rapautuva kieli: Kirjoituksia vähemmistökielten kulumisesta ja kadosta*, 103–148. Helsinki: Suomalaisen Kirjallisuuden Seura.

Granqvist, K. 2013b. Finnish Romani during the 1800s. In Schrammel-Leber, Barbara & Tiefenbacher, Barbara (eds.), *Romani V. Papers from the Annual Meeting of the Gypsy Lore Society, Graz 2011*. (Grazer Romani Publikationen 2), 13–28. Graz: Grazer Linguistische Monographien.

Granqvist, K. 2014b. Mixed morphologies: Morphological borrowing in Finnish Romani. (Paper presented at the 41st Finnish Conference of Linguistics, Turku, 10 May, 2014.)

Granqvist, K. 2024. *Suomen romanikielen varhaiset vuodet Gananderista toiseen maailmansotaan.* Suomen Romaniyhdistys. (https://www.suomenromaniyhdistys.fi/wp-content/uploads/2024/02/1800-l_valmis_opt.pdf)

Hakulinen, A. & Vilkuna, M. & Korhonen, R. & Koivisto, V. & Heinonen, T. R. & Alho, I. 2004. *Iso suomen kielioppi.* (Suomalaisen Kirjallisuuden Seuran Toimituksia 950.) Helsinki: Suomalaisen Kirjallisuuden Seura.

Hancock, I. F. 1992. The Hungarian student Vályi István and the Indian connection of Romani. *Roma* 36. 46–49.

Haspelmath, M. 1996. Word-class-changing inflection and morphological theory. In Booij, G. & van Marle, J. (eds), *Yearbook of Morphology 1995*, 43–66. Dordrecht: Kluwer.

Haugen, E. 1950. The analysis of linguistic borrowing. *Language* 26(2). 210-231.

Heath, J. 1984. Language contact and language change. *Annual Review of Anthropology* 13. 367–384.

Hedman, H. 2009. *Suomen romanikieli: Sen asema yhteisössään, käyttö ja romanien kieliasenteet.* (Kotimaisten kielten tutkimuskeskuksen verkkojulkaisuja 8.) Helsinki: Kotimaisten kielten tutkimuskeskus. (http://scripta.kotus.fi/www/verkkojulkaisut/julk8/)

Helsinki welfare office. 1979. *Helsingin mustalaisväestön sosiaaliset ja sivistykselliset olot.* (Moniste.) Helsingin kaupunki: Huoltovirasto.

Holzinger, D. 1993. *Das Romanes: Grammatik und Diskursanalyse der Sprache der Sinte.* (Innsbrucker Beiträge zur Kulturwissenschaft 85.) Innsbruck: Verlag des Instituts für Sprachwissenschaft der Universität Innsbruck.

Koivisto, V. 1992. *Tutkimus Suomen romanikansan ammateista ja niissä tapahtuneista muutoksista.* University of Helsinki. (Master's thesis.)

Koptjevskaja-Tamm, M. 2000. Romani genitives in cross-linguistic perspective. In Elšík, Viktor & Matras, Yaron (eds.), *Grammatical relations in Romani: The noun phrase*, 123–149. Amsterdam: John Benjamins.

Kovanen, P. 2013. Koodinvaihtelu romanikielessä. In Granqvist, Kimmo & Salo, Mirkka (eds.), *Romanikieli ja sen tutkimusalat*, 195–216. Helsinki: Suomalaisen Kirjallisuuden Seura.

Martinet, A. 1962. *A Functional View of Language*. Oxford: Oxford University Press.

Matras, Y. 1999 s/h alternation in Romani: An historical and functional interpretation. *Grazer Linguistische Studien* 51. 99–129.

Matras, Y. 2002. *Romani: A linguistic introduction*. Cambridge: Cambridge University Press.

Matras, Y. 2007. The borrowability of grammatical categories. In Matras, Y. & Sakel, J. (eds.), *Grammatical borrowing in cross-linguistic perspective*, 31–74. Berlin, New York: Mouton de Gruyter.

Matras, Y. & Sakel, J. 2007. Investigating the mechanisms of pattern-replication in language conver-gence. *Studies in Language* 31(4). 829–865.

Moravcsik, E. 1978. Universals of language contact. In Greenberg, Joseph H. (ed.), *Universals of Human Language*, 94–122. Stanford: Stanford University Press.

Moravcsik, E. 1995. Summing up Suffixaufnahme. In Plank, Frans (ed.), Paradigms: The economy of inflection, 451–484. Berlin: Mouton de Gruyter.

Pirttisaari, H. 2002. *Suomen romanin partisiippien morfologiaa*. University of Helsinki.

Pirttisaari, H. 2003. Muutos ja variaatio Suomen romanin verbien taivutustyypeissä. *Virittäjä* 4. 508–528.

Pirttisaari, H. 2004. Variation and change in the verbal morphology of Finnish Romani. In Nenonen, Marja (ed.), *Papers from the 30th Finnish Conference of Linguistics, Joensuu, May 15–16, 2003*, 195–216. Joensuu: University of Joensuu.

Pirttisaari, H. 2005. A functional approach to the distribution of participle suffixes in Finnish Romani. In Schrammel, Barbara & Halwachs, Dieter & Ambrosch, Gerd (eds.), *General and Applied Romani Linguistics. Proceedings from the 6th International Conference on Romani Linguistics*, 114–127. München: Lincom Europa.

Pott, A. F. 1844–1845. *Die Zigeuner in Europa und Asien I–II: Ethnographisch-linguistische Untersuchung, vornehmlich ihrer Herkunft und Sprache, nach gedruckten und ungedruckten Quellen*. Halle: Heynemann.

Ross, M. D. 2001. Contact-induced changes in Oceanic languages in North-west Melanesia. In Aikhenvald, A. Y. & Dixon, R. M. W. (eds.), *Areal diffusion and genetic inheritance: Problems in comparative linguistics*, 134–160. Oxford: Oxford University Press.

Salo, M. 2021. *Romanikieliset elementit romanien suomenkielisellä verkkokeskustelupalstalla*. Helsinki: Unigrafia.

Sakel, J. 2007. Types of loan: Matter and pattern. In Matras, Yaron & Sakel, Jeanette (eds.), *Grammatical Borrowing in Cross-Linguistic Perspective*, 15–30. New York: De Gruyter Mouton.

Sampson, J. 1926. *The Dialect of the Gypsies of Wales being the Older Form of British Romani Preserved in the Speech of the Clan of Abram Wood*. Oxford: Clarendon.

Social Investigation Bureau. 1954. *Mustalaisten olot. Sosiaalinen aikakauskirja*. Helsinki: Sosiaaliministeriö.

Stolz, Chr. & Stolz, Th. 1996. Funktionswortentlehnung in Mesoamerika: Spanisch-amerindischer Sprachkontakt. Hispanoindiana II. *Sprachtypologie und Universalienforschung* 49. 86–123.

Tenser, A. 2005. *Lithuanian Romani*. Lincom Europa: Munich.
Tenser, A. 2008. *The northeastern group of Romani dialects*. University of Manchester. (Doctoral dissertation.)
Thesleff, A. 1899. *Finlands zigenare. En etnografisk studie*. Helsinki.
Thomason, S. G. 2001. *Language Contact: An Introduction*. Edinburgh: Edinburgh University Press.
Thomason, S. G. & Kaufmann, Th. 1988. *Language contact, creolization, and genetic linguistics*. Berkeley: University of California Press.
Valtonen, P. 1964. *Indoarjalaiset sanat Suomen mustalaisten kielessä.* University of Helsinki. (Master's thesis.)
Valtonen, P. 1968. *Suomen mustalaiskielen kehitys eri aikoina tehtyjen muistiinpanojen valossa*. University of Helsinki. (Licenciate thesis.)
Valtonen, P. 1972. *Suomen mustalaiskielen etymologinen sanakirja*. (Tietolipa, 69.) Helsinki: Suomalaisen Kirjallisuuden Seura.
Van Hout, R. & Myusken, P. 1994. Modelling lexical borrowability. *Language Variation and Change* 6. 39–62.
Vehmas, R. 1961. *Suomen romaniväestön ryhmäluonne ja akkulturoituminen*. University of Turku. (Doctoral dissertation.)
Vennemann, Th. 1988. *Preference Laws for Syllable Structure and the Explanation of Sound Change*. Berlin: De Mouton Gruyter.
Vuorela, K. & Borin, L. 1998. Finnish Romani. In Ò Corráin, A. & Mac Mathúna, S. (eds.), *Minority Languages in Scandinavia, Britain and Ireland*, 51–76. Uppsala: Almqvist & Wiksell International.
Wentzel, T. 1964. *Tsyganski jazyk*. Moskva: Nauka.
Winford, D. 2003. *An introduction to contact linguistics*. Malden, MA: Blackwell.

Contact information

Kimmo Granqvist
University of Helsinki
kimmo.granqvist@helsinki.fi

Deeply embedded clauses in Finno-Ugric: A pilot study on Estonian and Moksha Mordvin

Edyta Jurkiewicz-Rohrbacher
Universität Hamburg, Universität Regensburg

Petar Kehayov
University of Tartu

Abstract

Complex sentences often contain clauses embedded in clauses that themselves are embedded. The properties of such deeply embedded clauses (DECs) and their relations to other parts of the sentence are poorly studied. We address this research gap by studying printed text material from two structurally different Finno-Ugric languages: Estonian and Moksha Mordvin. We investigate the relationship between embedding depth and the type of the embedded clause, its position relative to the superordinate clause and its temporal reference. Combining these variables, we observe associations between specific depths (first-, second-, third-order embedding), clause types (complement, relative, adverbial), positions (left-, right-, center-embedding), and temporal reference (absolute, relative). We show that DECs are not entirely identical with first-order embeddings, i.e., that embedding depth is a factor influencing the grammar of subordinate clauses and conclude that assessing DECs is crucial to the description of clausal subordination in a language.

Keywords: recursion in language, complex sentences, deep clausal embeddings, complement, relative, adverbial clauses, time reference, Estonian, Moksha Mordvin

1 Introduction

1.1 Deeply embedded clauses

Clausal subordination constructions have been extensively studied in Finno-Ugric languages and belong now to the minimum of detail required in a reference grammar of a language. But grammars, and even special studies on clause combining, tend to focus on first-order embedded clauses and leave deeper embeddings unattended. The Finnish example (1) demonstrates such an embedding.

(1) [$_{C=}$*Kyse* *on* *enemmänkin* *siitä,* [$_{C1=}$*onko* *otettu* *sellaisia*
matter.NOM be.PRS.3SG more:ADD it:ELA be:POL.Q take:PPP such:PL.PRT

riskejä, [$_{C2=}$*joihin* *ei* *ole* *ollut* *valtuuksia*]]].
risk:PL.PRT REL:PL.ILL NEG.3SG be.CNG be:APP.SG authorization:PL.PRT

'It is more a question of whether risks have been taken that were not authorized.' (ISK 2008: §1154)

VERTAISARVIOITU
KOLLEGIALT GRANSKAD
PEER-REVIEWED
www.tsv.fi/tunnus

Finnish Journal of Linguistics 37 (2024), 105–133
DOI: 10.61197/fjl.144760

A deeply embedded clause (DEC) is a clause which is embedded in a clause that itself is embedded. The clause in the innermost brackets of (1) occurs at an embedding depth of two, i.e., it is a second-order embedded clause. Sentences with multiple embeddings can be formalized [C [C1 [C2 [C3 […]]]]], where C = main clause, C1 = first-order embedded clause, C2 = second-order embedded clause, C3 = third-order embedded clause, etc. A deeply embedded clause is any clause below C1, i.e., any clause at an embedding depth of two or more.

1.2 Rationale

Although DECs are poorly studied as a phenomenon on their own, they are instrumental in theoretical syntax, specifically for the idea of recursion as an essential property of human language. Recursion, defined as "embedding a constituent in a constituent of the same type" (Pinker & Jackendoff 2005: 10), where "[m]aterial introduced by any lower recursive cycle is always contained in the material introduced by the immediately higher cycle" (Karlsson 2010b: 51), was argued to allow syntactic embedding at arbitrary depths (Blasi et al. 2019: 3938).

The importance of recursion to formal linguistics explains the interest of researchers in the structural characteristics of deeply embedded clauses, such as their embedding depth, or the position of the embedded clause in relation to the superordinate clause (tail- vs. center-embedding). However, even such measurable properties remain understudied from a typological perspective. Exceptions include Karlsson (2007a; 2007b; 2010a; 2010b), who studied the embedding depth and position of clauses in seven European languages (English, German, French, Latin, Swedish, Danish, Finnish), and observed severe restrictions on recursion in actual language usage. Right-branching recursion ([*I forgot* [*that she knows* [*who I am*]]]) rarely exceeds an embedding depth of five in written, and three in spoken language. Center-embedding does not exceed the depth of two (maximum three) in written languages ([*The man* [*the boy* [*the girl kissed*] *hit*] *filed a complaint*]) and one in speech.

In fact, the recursion postulate has had a negative effect on the interest in DECs. If rerunning the same procedure leads to reproduction of the same structure, one cannot expect to find something new in DECs compared to first-order embeddings. Function- and usage-oriented research has provided additional reasons for ignoring DECs. Interactional Linguistics, for example, opposes the idea of hierarchically organized sentence structure and regards clausal subordination as an analytical construct, not as a property of language (Laury & Ono 2010; Laury et al. 2021). This is why grammars of Finno-Ugric languages, even of major ones, do not explore deep clausal embeddings. The most recent Finnish and Estonian grammars, for example, exhaust the issue by stating that *there are* sentences with DECs (cf. ISK § 883[1] and EG § 403). In Hungarian grammars the phenomenon is discussed in greater detail, but even there the focus is exclusively on formal properties of such sentences, such as their hierarchical structure and possible sub-branching configurations (Rácz 1968; Keszler 2000).[2]

[1] The situation in Finnish linguistics can be explained by the upsurge of interest in the grammar of spoken language and the influence of Interactional Linguistics on Finnish grammaticography, a good example of which is ISK, the most comprehensive grammar of Finnish to this date (Maria Vilkuna, p.c.).

[2] These issues are discussed also in dedicated works on syntax (e.g. EKS: 753–755 on Estonian), along with observed embedding depths (e.g. Ikola et al. 1989: 17–20 on Finnish) and coreference relations across clauses (e.g. Fejes 2006 on Hungarian).

Another commonly held belief that hinders interest in multiple embeddings is that they are characteristic of written and not oral language (see Mithun 1984; Karlsson 2009). Most Finno-Ugric literary languages are very young; therefore, Uralicists tend to perceive recursively embedded clauses (especially finite ones) as a recent and somewhat unnatural property of "Russian" or "Germanic" literary tradition, which does not deserve special attention. A common view in traditional descriptions of Finno-Ugric languages is that multiple embedding is not only typical of written language – it is as an invention of writing.[3] But early texts in Finno-Ugric languages also raise questions in this respect. One of the first texts in Moksha, the Short Catechism of 1861, displays finite embeddings at a remarkable depth. Example (2) contains a third-order embedded clause.[4] What do such examples imply? Could we assume that once a language is written, recursive embedding arises and rapidly becomes an automatic procedure? Would that mean that all written languages allow embedding at an equal depth? Alternatively, is it possible that multiple embedding of finite clauses existed in oral language but 'waited to be inscribed'?[5]

(2) [$_{c=}$ *Ťä* *anamaťʼ* *veĺďä* *miń* *makśisaśk* *eś*
this.NOM request:DEF.SG.GEN through we.NOM give:PRS.S1PL>O3SG own.NOM

voĺeńäkiń *kaźńiś* *Škabazťi* *i* *anatama,* [$_{c1=}$ *meźä*
freedom:GEN present:DEF.SG.NOM God:DEF.SG.DAT and ask:PRS.1PL what.

Son *ťijəĺ* *marχtəńək,* [$_{c2=}$ *meźä* *Ťejnza* *eŕavi*
he.NOM do:CONJ.3SG with:1PL.POSS what.NOM he.DAT.3SG must.PRS.3SG

i *Soń* *oću* *Jońancti,* *stańä,* *meźä* *i*
and he:GEN great.NOM mind:3SG.POSS SG.DAT so what.NOM also

Ĺijä *lomaťťńä* *mäĺ* *marχta* *ťijəĺχťʼ* *Soń*
other.NOM people:DEF.NOM desire.NOM with do:CONJ.3PL he:GEN

voĺanc *kolga* *mastərťʼ* *laŋksa,* [$_{c3=}$ *koda* *ćebäŕsta*
will:3SG.POSS.GEN according earth:DEF.SG.GEN on how properly

ťijəńďsaź *Soń* *voĺanc* *Aŋgeĺχťťńä* *meńəĺ*
do:PRS.S3PL>O3SG he:GEN will:3SG.POSS.GEN angel:DEF.PL.NOM heaven.NOM

laŋksa]]]].
on

'[With this request, we give our freedom as a gift to God and ask [that He would do with us [what He needs according to His great intention, and what other people would also properly do according to His will on earth, [just as the angels in heaven gladly fulfil His will]]]].' (Feoktistov 1976: 249)

[3] A quote from Buzakov (1973: 8–9) depicts this view: "The category of complex sentences is relatively young in the Mordvin languages. As is well known, complex sentences are more intrinsic to written than to oral language. And since the first Mordvin texts appeared in the 18th century, there is every reason to assume that complex sentences in Mordvin languages began to develop at that time" (our translation).

[4] The English translation is bracketed to make this long sentence assessable.

[5] It is also possible that the recursive structure in (2) is a parrot translation from Russian. Like other first texts in minor languages, this is a translation of a canonical text. The Short Catechism was published in Moscow in the synodal printing house and its author is unknown (see https://wikisource.org/wiki/Нюрьхкяня_катехизис/В).

We study Estonian and Moksha Mordvin – two languages which belong to different branches of Finno-Ugric. The former has a relatively old literary tradition and a large body of published texts, the latter a young and sparse literary tradition, and thus a rather small body of texts. Estonian represents a typical East-Central European national language, a result of 19th century nation-building and subsequent language planning based on German models. This process led to the development of a rich literary tradition encompassing texts across all genres (Laanekask & Erelt 2003). Moksha literary language never became a daily used and shared register of the language community. Language planning began in the 1920s and the number of texts published annually (mainly fiction and journalism) has ever since remained low (see Bartens 1999: 9–23; Feoktistov 1976: 10–69, 127–153). The body of printed texts in Moksha is hundreds, if not thousands of times smaller than that of Estonian texts.

Linguistic grounds to choose these languages include the fact that they are genealogically sufficiently distant to represent variation on a Finno-Ugric scale. Estonian is among the most SAE-like Uralic languages, while Moksha is a peripheral member of the Volga-Kama Sprachbund. During the last millennium Estonian has been in contact mainly with German, Russian, and Latvian, while Moksha with Russian, Tatar, and Chuvash.

Moksha has a richer grammatical mood system but somewhat less differentiated tense system than Estonian. Unlike Estonian, Moksha has elaborate devices for cross-referencing arguments on the verb. Both are SVO languages, but in Moksha the transition from OV to VO order has taken place in the more recent past, which is reflected in the excessive variation of word order in contemporary language (Toldova et al. 2018: 549–550, 608–615; Hamari & Ajanki 2022; Vilkuna 2022). The decrease of SOV word order (and the respective increase of SVO) in Mordvin languages can also be observed by comparing 19th century folklore transcripts with texts from the 20th century (Saarinen 1991: 50). In Estonian, the NP is more rigidly head-final than in Moksha in which an adjective modifier, a quantifier, or determiner may also follow the noun (Toldova et al. 2018: 296). Moksha is a strictly postpositional language, whereas Estonian also has prepositions and ambipositions. The two languages differ as to the degree to which they use non-finite clauses in subordination: Moksha is a more non-finite language than Estonian.[6]

These differences are interconnected with the features we monitor in our study on DECs (see §1.3). Therefore, we expect to observe differences between the languages also in sentences with DECs. Secondly, and more importantly, monitoring structurally different languages allows us to make cross-linguistic generalizations over the properties of DECs as such.

[6] Finite and non-finite dependent clauses occur in all Finno-Ugric languages, but their weight in the system of subordination varies immensely (Skribnik 2022). The finite strategy is spreading from the West, whereas the inherited non-finite strategy is in retreat. Languages can be arranged on a scale, where the Ob-Ugric languages are the most 'non-finite' ones, and South Saami is probably the most 'finite' one (Ylikoski 2022: 127). The difference between Estonian and Moksha is noticeable. In our initial sample of sentences with DECs, which included as a separate level of embedding also non-finite clauses without a subordinating conjunction (cf. §2.3), Moksha featured 28% and Estonian 16% non-finite clauses.

1.3 Research question

Basing on the above discussion, our goal is to study the properties of deeply embedded clauses (DECs) in comparison to the clauses which in §1.1 were defined as first-order embedded clauses (C1s). An important caveat here is that we compare DECs with C1s in a population of sentences with DECs; we make no claims about C1s in sentences ending at an embedding depth of one. The central research question of this study is:

> RQ1: Within a sample of sentences containing DECs, do DECs differ significantly from C1s regarding the distributional properties of their structural features?

A related question is: if they do, then *how* do they differ from C1s? How do the properties of embedded clauses change with increase in embedding depth, which features disappear and which appear at what level of embedding? A positive answer to RQ1 would mean that clausal subordination in language cannot be described and explained based only on first-order embedded clauses and their relation to the main clause. As we saw above, the implicit assumption that C1s suffice to elucidate subordination can be found even in grammars of major Finno-Ugric languages. Conversely, a negative answer to RQ1 would mean that C1s suffice to elucidate all facets of clausal subordination in language.

To provide an answer to RQ1, we monitor the interplay of the embedding depth of clauses with a) their clause type (complement, relative, adverbial), b) position relative to the superordinate clause (left-, right-, center-embedded), and c) temporal reference in relation to the speech situation (absolute) or to the time scheme of the higher clause (relative). These variables will be introduced in detail in §3. While attempting to answer the research question and monitoring the variables, we also make observations as to minor but potentially interesting issues, such as whether the two literary languages differ as to the average embedding depth of their sentences with DECs.

Why exactly variables (a–c)? They are simple categorical variables ([a] and [b] are ternary and [c] binary), unlike for example '(non-)finiteness' of clauses, which is a major feature of clausal subordination in Finno-Ugric, but which is a composite variable constructed differently for various languages. The value 'finite' is a sum of some or all of the features of tense, mood, and verbal person/number inflection on the predicate of the clause. From a quantitative perspective, studying finiteness entails examining the marking of each of these features individually and in conjunction with their interaction terms. This would necessitate a considerably larger sample than the one we could assemble. In fact, we will minimize the effect of finiteness, as it undermines our basic unit of analysis – 'clause'; see §2.3 for details.

Sentences with DECs in Finno-Ugric languages have not been studied in relation to the above variables. Furthermore, with these variables we cover both the syntax and semantics of multiple embedding. The clause type variable bears both on the syntactic function of the clause (argument, modifier, or adjunct) and on its meaning (complement, relative and adverbial clauses relate to expressions of different conceptual status and complexity). The position variable is a purely syntactic (configurational) variable, whereas the temporal reference a purely semantic variable.

As already noted, these variables are interconnected with general typological features of Estonian and Moksha. For example, the position variable is related to the basic word order and the adposition–noun order. Likewise, the position correlates in Uralic with the choice between finite and non-finite subordination (Vilkuna 2022; Kiss 2023).

Finiteness indirectly affects embedding depth; e.g., Progovac (2010) argued that small clauses with reduced tense and other features of finiteness (e.g., *us having left*, …) have limited recursion potential.

This is the first study on the given topic, and it is explorative at best. Nevertheless, it is corpus-based and, despite the relatively small data, also quantitative. The features of embedded clauses will be quantified and the differences in their distribution will be evaluated by means of statistical methods.

In §2 we describe the studied sample, the queries in corpora, and discuss the restrictive criteria we applied in filtering eligible sentences and clauses. In §3 we define our variables and show their distributional properties in the studied samples of Estonian and Moksha. Note that §3 is purely *descriptive* and not *inferential* in nature. The objective is to demonstrate the core characteristics of the sample in relation to the research question and variables selected for the study. We do not immediately provide quantitative evaluation of the relationship between the embedding depth and the studied variables, because our approach is multidimensional. This means we examine not only the association between each potentially relevant variable and embedding depth, but also consider how the strength of this association may be influenced by the value of another variable; this concept is referred to as interaction term (Oakes 1998: 37). Consequently, in §4 we employ logistic regression modelling of the embedding depth as a dependent variable, along with the structural features for which sufficient observations could be obtained from the sample. §5 discusses the results of the statistical analysis.

2 Sample compilation

2.1 Language

The compilation of comparable samples of sentences containing multiply embedded clauses, i.e., of sentences containing at least second-order embeddings, for Moksha and Estonian was a challenging task, because Moksha is a low-resource language. To ensure the high recall of sentences with DECs, we needed corpora with morphosyntactic annotation to formulate complex (yet rather imprecise) queries from which we could filter out the irrelevant hits. Moksha material was extracted from the Corpus of Contemporary Literary Moksha (CCLM)[7], the only available morphosyntactically annotated source at the time of the search. The Estonian material originates from the Estonian National Corpus (ENC 2019)[8]. These corpora differ drastically in the size and functionalities of the available interfaces; therefore, we compiled the Moksha sample first and adjusted the search design for Estonian to obtain a maximally comparable data set.

As specified in §1.2, Moksha has a rather sparse literary tradition. This is reflected in the structure of the corpus, 86.4% of which consists of press (ca. 1,500,000 words at the time of sample compilation, September 27–30, 2021). The fiction subcorpus (0.8% of CCLM) consists of texts published between 1929 and 1937 and, at the time of the search (September 24, 2021), contained only 11,519 running words. The respective figures for the Estonian subcorpora of ENC are five million words for fiction and five million for press.[9]

[7] Moksha.web-corpora.net

[8] www.sketchengine.eu

[9] The fiction subcorpora of both languages consist of original literary works, i.e., they do not contain translations.

Although Moksha data seems imbalanced regarding genre, this does not compromise our analysis. Moksha press material can be regarded as undetermined with respect to the distinction between journalism and fiction. As mentioned, Moksha has a small user community, which, we believe, has inhibited the growth of differences between genres. The same can be observed for other literary forms of Finno-Ugric languages of Russia, at least for communities comparable in size with the speakers of Moksha. Finno-Ugric prose writers often publish in the local press; Mordvin, Udmurt, or Komi newspapers and journals function as platforms for practicing the language and keeping it alive, and they publish more fiction-like texts than the press in nations with long literary traditions. While the Moksha sample is not representative with respect to genres, it is not necessarily less diverse relative to the entire body of Moksha texts than the Estonian sample (see Stefanowitsch 2020: 34–35 for the difference between *representativeness* and *diversity* of corpora). Therefore, we compare the Estonian and Moksha data without discriminating between press and fiction.[10]

2.2 Querying corpora

Because of the very limited functionalities of CCLM, the manual filtering of the obtained material was beyond our capabilities, so we restricted the number of press sources, attempting at the same time maximal coverage among the media channels represented in the corpus: newspapers (92,624 words) and web pages of TV (21,497 words) and radio (3,801 words). Thus, Moksha sentences in our sample originate from a sample of 129,441 words (117,922 words press plus 11,519 fiction). The obtained sample of eligible candidates for sentences with DECs was reduced by applying a restrictive criterion presented in §2.3 below. As a result, the final population of sentences with DECs dropped to 156 items: 8 sentences from fiction and 148 from press. Considering that each sentence contains at least two successively embedded clauses, and at least one DEC, and that each clause was to be coded for several variables, we considered this sample to be sufficient for the purposes of the study.

For Estonian, we needed a population of at least the same size. Maarja-Liisa Pilvik designed a query in Corpus Query Language comparable to the one constructed for Moksha by the corpus designer Timofey Arkhangelskiy. On October 6, 2021, she retrieved from the press and fiction subcorpora of ENC (2019) one thousand random sentences per subcorpus with deep embeddings. ENC was accessed via the SketchEngine corpus manager. We manually checked only half of the obtained concordances, leaving the rest in reserve for further studies. As in Moksha, the eligible candidates for sentences with DECs were further filtered according to our restrictive criterion. This way, we obtained a sample of 292 Estonian sentences: 149 from fiction and 143 from press.

[10] Genre seems, however, to be an important variable. The unpublished study of Sinnemäki (2004) suggested that artistically marked genres feature greater embedding depth than everyday written genres.

2.3 Restrictive criteria in filtering sentences and clauses

The final sample of sentences with DECs was compiled by applying the following criterion. We adopted a rather restrictive view of the notion of 'clause' and did not count clause-like units without a subordinating conjunction (henceforth 'subordinator') as separate levels of embedding.[11] Because our approach is quantitative and explorative, we tried to circumvent in this way the potential bias towards structures that scholars consider cases of clause-union and co-lexicalization (see Givón 2001: 39–90). These phenomena are instantiated by the frequent use of modal and light verbs; see example (3).

A unit of linguistic expression is counted in this study as a separate clause, regardless of whether its predicate is a finite or non-finite verb form (e.g. infinitive or participle), only if it contains a subordinator. This is demonstrated by examples (3) from Moksha and (4) from Estonian. In example (3), the infinitive *jumaftəms* 'lose' in C and its object are not counted as a separate clause (or, respectively, as a separate level of embedding), because the infinitive is a direct complement of *eŕavś* 'must' and there is no subordinator in the expression. On the other hand, in C1 of this sentence the infinitive *šaŕχkəďəms* 'realize' occurs with a subordinator in a purpose clause which is counted as a separate level of embedding.[12] The predicate of C2 is an adjective conjugated like a finite verb in person/number and occurring in a complement clause headed by the subordinator *koda* 'how'. Accordingly, C2 is counted as a separate clause, i.e., a separate level of embedding. Likewise, the finite predicate in C1 of example (4) *ei ole raske* 'is not difficult' and the infinitive *kirjutada* 'write' are not considered separate clauses, i.e., different levels of embedding, because there is no subordinator between them. The combination of the subordinator *kui* 'if' and the infinitive *teha* 'make' in C2, on the other hand, is counted as a separate clause, i.e., a separate level of embedding.

(3) [...] [$_{C=}$ *ťejńä* *eŕavś* *iŋgə́li* *jumaftəms* *toń,* [$_{C1=}$ *štoba*
1SG.DAT must:PST1.3SG before lose:INF 2SG.GEN in_order_to

šaŕχkəďəms, [$_{C2=}$ *koda* *ton* *ťejńä* *piťńijat*]]].
realize:INF how_much 2SG.NOM 1SG.DAT precious:PRS.2SG
'… I first had to lose you in order to understand how precious you are to me.'

(4) [$_{C=}$ *Puik* *selgitab,* [$_{C1=}$ *et* *eesti* *keelt* *ei*
Puik.NOM explain:PRS.3SG that Estonian language:PRT NEG

ole *raske* *kirjutada,* [$_{C2=}$ *kui* *teha* *lapsele*
be.PRS.CNG difficult write:INF if make:INF child:ALL

selgeks *õigekirjas* *valitsev* *süsteem*]]].
clear:TRL spelling_system:INE prevalent.NOM system.NOM
'Puik explains that it is not difficult to write Estonian if you make the prevalent spelling system clear to the child.'

[11] With the term 'subordinator' we designate complementizers, adverbializers (conjunctions introducing various types of adverbial clauses), and relativizers (relative pronouns and adverbs).

[12] In Moksha, a noun, pronoun, adjective or adverbial expression can be conjugated like a verb and function as a predicate of a finite clause (see e.g. Hamari & Ajanki 2022: 423).

By excluding from the sample clauses without a subordinator, there remained only syndetic subordinate clauses. We regard the presence of a subordinator as a sign of grammatical and prosodic autonomy of the clause, even if its predicate is a non-finite verb form or is omitted.[13] This way we ensure that we have a population of sentences with DECs even upon a very narrow understanding of 'clause' (see e.g. Kehayov 2016 for clause-status criteria in Finnic). Each sentence in the sample contains at least one clause embedded in an embedded clause. The least complex sentence meeting our operational definition of a sentence with DEC(s) is one with a main clause, a first-order embedded clause with a subordinator, and a second-order embedded clause with a subordinator. Importantly, this definition produced a very restricted but clearly defined set of eligible non-finite clauses, those fronted by a subordinator, and led to the exclusion of many infinitives, participles, converbs, and action nominals, and to a general reduction of the effect of finiteness on the distribution of features in the Estonian and Moksha sample.[14]

Within this population, we disqualified first-order embedded clauses, which do not embed a further clause. Such 'childless' C1s are excluded from the frequency counts. For example, in a sentence with the structure [C [C1 [C2]] [C1]], only the first C1 is monitored in relation to position, type, and temporal reference, and included in the respective calculations. A sentence which contains two DECs with different antecedents but occurring at the same depth, like [C [C1 [C2]] [C1[C2]]], is treated as two separate sentences with DEC.[15]

Once the samples were assembled, the sentences with DECs were annotated for the variables, stored in comma separated format and quantitatively evaluated in the R statistical environment (R Core Team 2021, version 4.3.1).

3 Variables and their distribution

3.1 Sheer embedding depth

We annotated, in total, 448 sentences with DECs, which contained 985 embedded clauses. Of these, 156 Moksha sentences with DECs contained 328 embedded clauses, and 172 DECs. The Estonian sample comprised 292 sentences with DECs, with 657 embedded clauses, and 365 DECs.

Although the number of multiple embeddings in a sentence could be operationalized as a discrete numeric variable, we prefer to treat DECs and the sentences in which they occur as observations of categorical data and study the whole distribution, rather than focusing on the central tendency. This is because the set of values is rather restricted in our sample to (C)2, (C)3, (C)4 and (C)5. While recursion could theoretically be infinite, it is highly improbable to obtain a sentence with a two-digit number of successive embeddings in actual language use. Table 1 summarizes the frequencies of the maximum embeddings observed in our data. Estonian has 228 sentences ending at an embedding depth of two (i.e., 228 sentences whose deepest clause is C2), which is 78.1% of the total

[13] In case of omission, the verb and its finiteness features are recoverable from the context.

[14] As noted in §1.2, Moksha resorts to non-finite verb forms in subordination more often than Estonian.

[15] All this involved a rigorous filtering of irrelevant sentences and clauses. Challenging cases included complex sentences containing coordinated complex sentences, and discriminating between clauses attached to complex sentences and clauses attached to clauses within these complex sentences.

of sentences with DECs. Likewise, the proportion of sentences ending at an embedding depth of three comprise 19,5% of the total of sentences with DECs. In Moksha, 91% of sentences ends with a C2.

Table 1. Number of sentences ending at a certain embedding depth and their share of the total of sentences with DECs

Depth	Estonian		Moksha	
	N=292	% of sentences with DECs	N=156	% of sentences with DECs
C2	228	78.1%	142	91.0%
C3	57	19.5%	12	7.7%
C4	5	1.7%	2	1.3%
C5	2	0.7%	0	0.0%

Sentences ending at the C2 level are by far the most frequent in both languages. The percentage of sentences featuring a depth of three or deeper is twice as high in Estonian (21.9%) as in Moksha (9%). It can be inferred that recursive embedding is somewhat less automatic in Moksha than in Estonian. The Fisher's Exact Test for count data allows us to reject the null-hypothesis that the samples are drawn from the same distribution ($p=0.002$). This provides compelling evidence to support the assertion that the Estonian sample exhibits greater embedding depth than the Moksha sample.

The maximum depth observed in the Estonian sample is five, while in the Moksha sample it reaches four. The absence of C5s in Moksha may be attributed to the limited sample size, and thus, it would be premature to conclude that the boundaries of recursive embedding differ between these languages.

Table 2 shows the aggregated frequencies of DECs at different embedding depths (beginning with C2), which we will use in §3.2–3.4 as a variable. The percentages show the share of sentences with clauses at a certain embedding depth from all sentences with DECs in the sample. For example, the Estonian sample comprises 292 sentences with DECs, which contain 365 DECs altogether. The sentences containing a third-order embedded clause (C3) are 22% (n=64) of all sentences with DECs (n=292).

Table 2. Number of clauses at different embedding depths and their share of the total of sentences with DECs

Depth	Estonian		Moksha	
	N=365 (DECs)	% of sentences with DECs	N=172 (DECs)	% of sentences with DECs
C2	292	100%	156	100%
C3	64	22%	14	9%
C4	7	2%	2	1%
C5	2	0.6%	0	0.0%

We see from Table 2 that in Estonian the share of sentences with clauses occurring at the embedding depth of three or deeper is two and a half times higher than in Moksha. Summing up the percentages in the last three rows of the table, we get 24.6% for Estonian and 10% for Moksha.

3.2. Clause type

This variable accounts for the syntactic type of the clause: complement clause (e.g. including the semantic types *that*-clause, wh-complement clause, etc.), relative clause, and adverbial clause (e.g. of time, manner, purpose, etc.).[16]

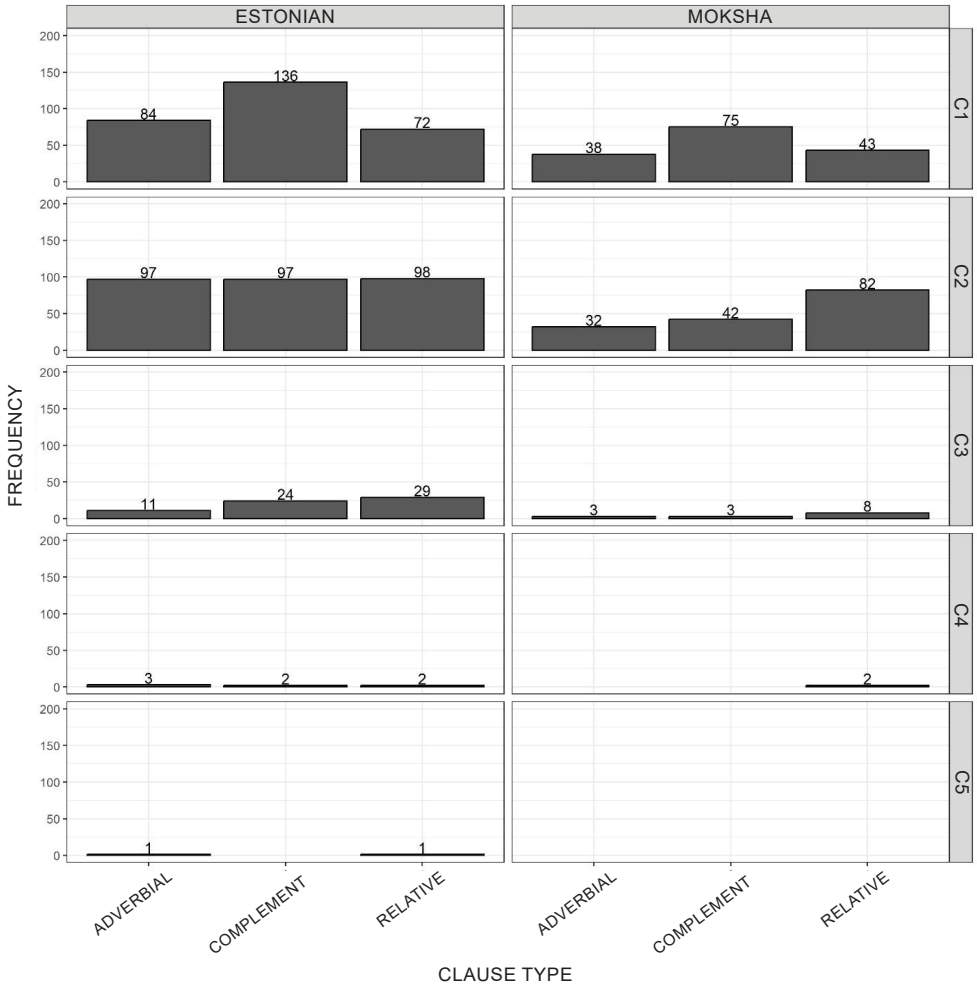


Figure 1. Types of embedded clauses at different embedding depths

The frequency counts for all embedding depths are given in Figure 1, whereas in Table 3 DECs are aggregated in one row. In Estonian, we observe all types of subordinate clauses at all levels until C4. In Moksha, only relative clauses remain until level C4.

[16] Certain functional subtypes of clauses are more likely to be non-finite and lack a subordinator. Accordingly, the decision to exclude constructions without subordinators, which had a severe effect on the share of clauses with non-finite predicates in the sample, resulted in unbalanced distribution of subtypes of clauses. For example, based on this criterion state-of-affairs complements ('I want/start/try *to go*.') and non-finite purpose clauses were expelled from the sample. But this is irrelevant here because we are concerned with major syntactic types of clauses, not with their semantic subtypes.

Table 3. Distribution of C1s and DECs across clause types

Depth	Estonian N=657			Moksha N=328		
	adverbial	complement	relative	adverbial	complement	relative
C1	84	136	72	38	75	43
DEC	112	123	130	35	45	92
Σ	196	259	202	73	120	135

In both samples, complement clauses seem to prefer to be C1s, i.e., first-order embeddings. In the Estonian sample, they are slightly more frequent in first-order embedding than in deep embedding; 52% (n=136) of the Estonian complement clauses are C1s. In Moksha, the tendency is more pronounced: 63% (n=75) of the Moksha complement clauses are C1s. Relative clauses, in contrast, tend to appear in both languages as DECs: 64% (n=130) of Estonian and 68% (n=92) of Moksha relative clauses are DECs. Adverbial clauses do not show a clear tendency. In the Estonian sample, 57% (n=112) of the adverbial clauses are DECs, while in the Moksha sample 52% (n=38) of them are C1s.

Switching the perspective from clause type to embedding depth, the figures in Table 3 suggest that, in both languages, the most common C1 is a complement clause. In the Estonian sample, the adverbial clause is the second most frequent C1, while in the Moksha sample, C1 is least frequently an adverbial clause. DECs, on the other hand, are most frequently relative clauses in both languages.

In Figure 1 above, we see how the relative frequencies of clause types change when moving deeper in the recursive structure. From level C1 to level C2, the share of complement clauses in Estonian decreases, while relative clauses and adverbial clauses become more frequent. In Moksha, relative clauses nearly double their frequency with the transition from C1 (n=43) to C2 level (n=82), whereas the frequency of complement and adverbial clauses drops. Third-order embeddings are quite infrequent; nonetheless, we have 73 observations from Estonian and 16 from Moksha. In both languages, relative clauses are most frequent at this embedding depth.

A reviewer of this article drew our attention to an unpublished master's thesis on complex right-branching clauses in Finnish (Sinnemäki 2004). Sinnemäki (2004: 49–57) arrives at a similar association between clause type and embedding depth. The share of complement clauses in his sample decreased as the depth increased, while the share of relative clauses increased. On the other hand, unlike in our data, the share of adverbial clauses increased with the increase in depth, even more so than the share of relative clauses. We leave the interpretation of this difference for further research, noting that Sinnemäki's sample is comparable to ours, as he used press and prose texts.

All in all, it seems that the distribution of clause types among first-order embeddings is similar in Estonian and Moksha, with complement clauses being the most typical C1s. We also observe an increase of the frequency of relative clauses in both languages at further embedding depths. This increase is much more rapid in Moksha. It appears that Moksha features a larger difference between the average depth of complement and relative clauses than Estonian. This observation will be evaluated in §4.

3.3 Position of the clause

Next, we studied the position of the embedded clause relative to its matrix clause, distinguishing between 'right-branching tail-embedding', 'left-branching tail-embedding', and 'center-embedding'. All examples presented so far instantiate right-embedding where the subordinate clause is postposed relative to its matrix clause. Left-embedding, producing clauses preposed to their matrix clause, is characteristic in Estonian and Moksha for conditional (*if*-), purpose (*in order to*-) and concessive (*although*-) adverbial clauses. In center-embedding, the subordinate clause is placed to the right of at least one constituent (which can be a relative pronoun) of the superordinate clause and to the left of the remaining part of this clause (cf. Karlsson 2007a: 109; 2010b: 53). In example (5) from Moksha, the relative clause C2 is center-embedded in a complement clause.

(5) [$_{C=}$*A* *piŋgś* *veši* [$_{C1=}$*štoba* *organizatsija,* [$_{C2=}$*kona*
but time:DEF.NOM demand:PRS.3SG that.IRR organization.NOM which

ťäďäń *Ĺemsta* *väťəĺ* *rabota* *iťťńəń*
mother:GEN name:ELA conduct:PST2.3SG work.NOM child:PL.DEF:GEN

jotksa], *uĺəĺ* *puraptf*]]
among be:CONJ.3SG come_together.PST.PTCP

'But the time demanded that the organization, which in mother's name was working with the children, should come together.'

Is there a dependence between the position of the clause relative to its matrix clause and its embedding depth? What is the maximal depth of right-, left- and center-embedding in the two languages?

Figure 2 presents the position of clauses in relation to the depth at which they occur. The default right-embedding persists into the deepest observed level of embedding, while the infrequent left- and center-embedding reach the depth of three in Estonian and two in Moksha. In both languages the use of center-embedding grows from C1s to C2s, while the use of left-embedding decreases in Estonian and slightly increases in Moksha. Normalized to the size of the samples, Moksha C2s are clearly more sensitive to center-embedding than Estonian C2s: center-embedding occurs 9 times in Estonian C2s which is 3% of all C2s in the Estonian sample, whereas in Moksha we have 15 occurrences at the C2 level, which amounts to nearly 10% of the C2s in the Moksha sample. However, this is the most unbalanced variable since the figures for center- and left-embedding are too small; therefore, we will not include it in the statistical analysis in §4.

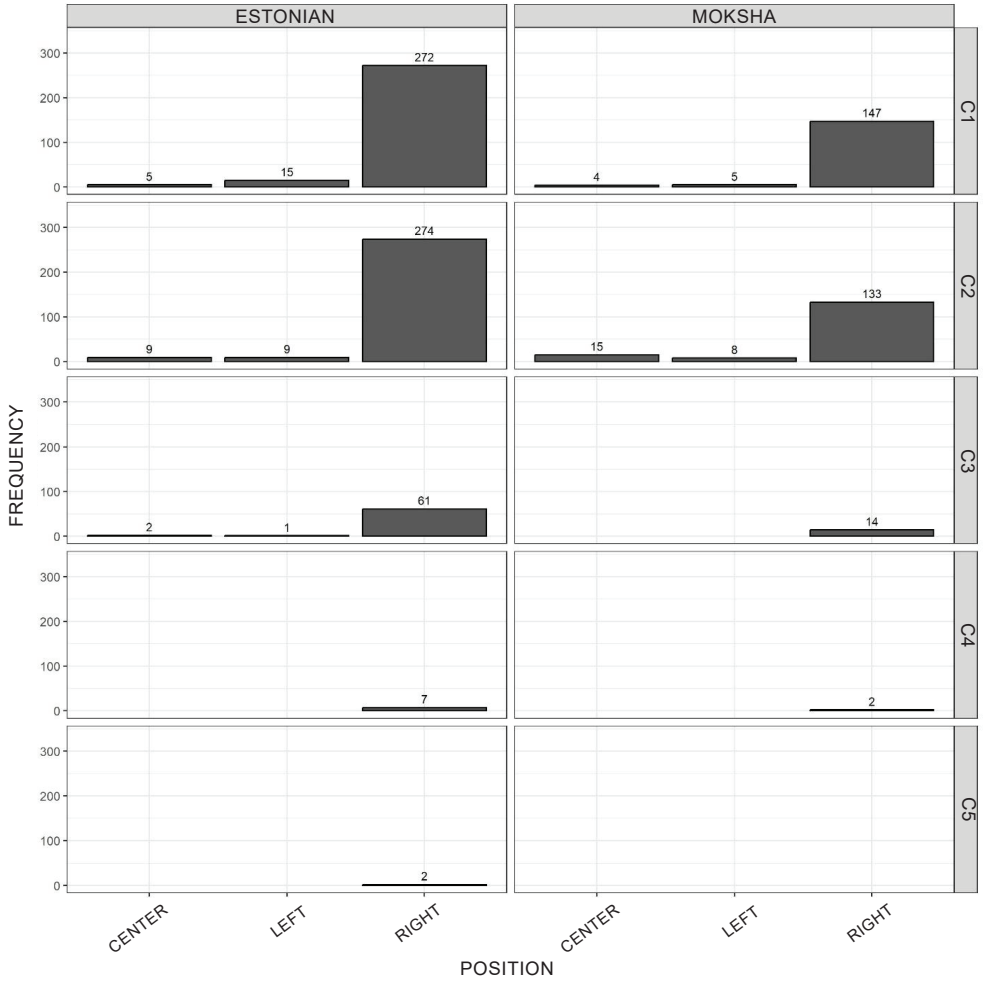


Figure 2. Position of embedded clauses at different embedding depths

We noted in §1.2 that language users' ability to recursively apply center-embedding, i.e. to *nest clauses*, is limited. Research has shown that this limitation is related to working memory. Center-embedding is a burden to the human parser because, as Auer (2005: 27) noted, the memory for form is shorter than the memory for content. The language user cannot keep the embedding procedure open for a long time and faces difficulties in producing nested discontinuous clauses (Blasi et al. 2019; Laury & Ono 2010). Listeners and readers too have limited ability to keep material in mind while waiting for further input to be linked to earlier parts of the sentence (Frank et al. 2016: 554).[17]

But this limitation does not mean that center-embedding cannot be applied at any embedding depth. The distribution in Figure 2 suggests that center-embedding is efficient also in deep embeddings, i.e., embedding depth poses no restrictions to center-embedding. The lack of occurrences at C4 and C5 levels in our samples can be related to the infrequency of fourth- and fifth-order embeddings.

3.4 Temporal reference: absolute vs. relative

The subordinate clauses in the samples were coded in relation to the distinction between absolute and relative temporal reference. We speak of 'absolute temporal reference' when the tense distinctions of a subordinate clause are determined in relation to the time of

[17] See also Gibson (1998) for a profound discussion of the memory cost of syntactic discontinuity.

the speech act (i.e. of the utterance of the speaker). Given that the reference point for the temporal location of the state of affairs (SoAs) of the clause is the moment of utterance, we speak of absolute present, absolute past, or absolute future. The temporal reference is 'relative' when the SoAs of the clause is located in time with reference to a moment (or period of time) in which another SoAs obtains. In this case, the reference point for the location of the SoAs of the clause is the time of a SoAs described in a superordinate clause, and we have relative present, relative past, or relative future (see Comrie 1985: 55; Sgall 1990; Cristofaro 2003: 63). In languages with both finite and non-finite subordinate clauses, finite clauses may have either absolute or relative temporal reference, whereas non-finite clauses usually have relative temporal reference.[18] This derives from the fact that, in such languages, finite clauses tend to express propositions, i.e. meaning units that have truth-value and are independently tensed, whereas non-finite clauses tend to express SoAs and thus have sub-propositional status, lacking truth-value and independent tense (see Dik & Hengeveld 1991; Kehayov & Boye 2016: 815–817).

In example (6) from Estonian, C1 and C2 have relative temporal reference. The situation described by the main clause C occurs in the past, and the SoAs described by C1 is located in time relative to this past moment; in this case, we have relative present (or future, which is marked in Estonian by the same non-past tense form). The present tense in C1 cannot be interpreted as absolute present, because the SoAs described in 'start ministerial term' precedes the moment of speaking. This is also the case for C2, which is simultaneous with C1. Clause C3, on the other hand, is tensed past, but this past tense does not express anteriority with reference to the SoAs of the superordinate clause. The SoAs described in C3 is simultaneous with the SoAs described in the higher clauses: Kallas wants to get rid of the chancellor *at the time* he (hypothetically) buys a car and *at the moment* he starts his ministerial term. The only adequate interpretation of the past tense in C3 is that it expresses 'past from now' (i.e. past from the moment of utterance), which means that C3 features absolute temporal reference.

(6) [$_{C=}$*Arusaadavalt* *ei* *saanud* *minister* *Kallas*
understandably NEG get:PST.CNG minister Kallas

endale *lubada,* [$_{C1=}$*et* *alustab* *ministriametit*
REFL:ALL afford.INF COMP start:PRS.3SG ministerial_TERM:PART

sellega, [$_{C2=}$*et* *ostab* *kantslerile,* [$_{C3=}$*kellest*
it:COM COMP buy:PRS.3SG chancellor:ALL who:ELA

ta *nagunii* *tahtis* *lahti* *saada,*]
3SG anyway want:PST.3SG rid get:INF

uue *auto*]]].
new.GEN car:GEN

'Understandably, Minister Kallas could not afford to start his ministerial term so that he buys a new car for the chancellor, whom he wanted to get rid of anyway.'

[18] See, however, Shagal (2018: 67–68, 73) on absolute temporal reference in Mari, Komi-Zyryan, and Tundra Nenets participial clauses.

In (7) from Moksha, C1 has an absolute and C2 a relative temporal reference. C is in present tense expressing that its SoAs occurs at the moment of speaking, and the Moksha first past tense in C1 conveys the past from this moment, i.e., we have an absolute past. C2, on the other hand, contains a non-finite verb form, which expresses future (posteriority) relative to the SoAs described in C1: the Soviet people die and *then* the fascism is defeated.

(7) [$_{C=}$*Bəta* *son* *af* *sodasi,* [$_{C1=}$*što* *jotaj*
as_if s/he.NOM NEG know:PRS.S3SG>O3SG that last

vajnasa *miń* *27* *miĺjon* *lomańəńkä*
war:INE we.GEN 27 million.NOM people:1PL.POSS.NOM

maksəz *eŕafsnən,* [$_{C2=}$*štoba* *mašftəms* *fašizmaŕ*]]].
give:PST1.S3PL>O3PL life:3PL.POSS.GEN so_that defeat:INF fascism:DEF.SG.GEN
'As if he doesn't know that in the last war, 27 million people from our land gave their lives to defeat fascism.'

Table 4 presents the frequencies of the values 'absolute' and 'relative' temporal reference of C1s and DECs. In total, 53% (345 of 657) of the Estonian embedded clauses and 46% (151 of 328) of the Moksha embedded clauses have absolute temporal reference.

In Estonian, C1s feature 68% (199 of 292) absolute temporal reference while DECs only 40% (146 of 365). In Moksha, C1s feature 51% (80 of 156) absolute temporal reference and DECs 41% (71 of 172). Thus, we observe in both languages a decrease of absolute temporal reference from the C1 level to DECs, and a corresponding increase of relative temporal reference.

Table 4. Temporal reference in C1s and DECs

Depth	Estonian N=657		Moksha N=328	
	Absolute	Relative	Absolute	Relative
C1	199	93	80	76
DEC	146	219	71	101
Σ	345	312	151	177

The distribution of absolute (T-Abs) and relative (T-R) temporal reference across embedding depths is shown in Figure 3.

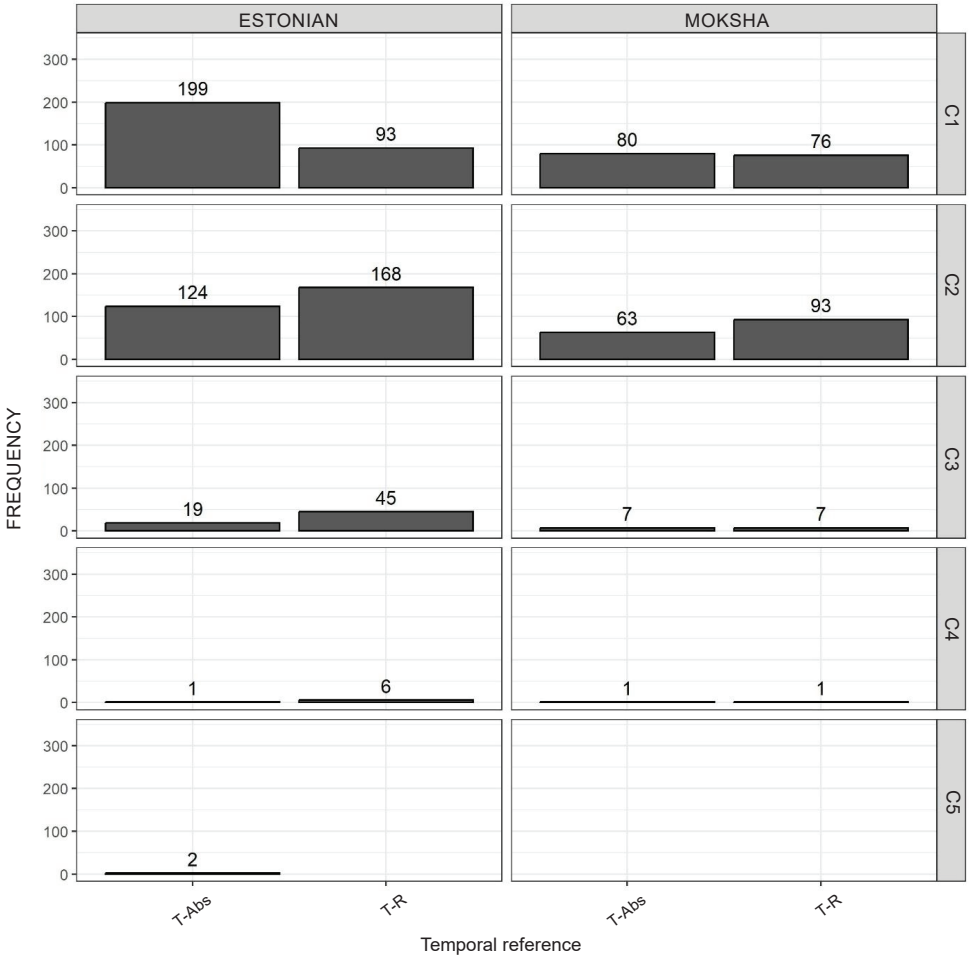


Figure 3. Distribution of temporal reference across embedding depths

Figure 3 confirms the observations from Table 4 and suggests that relative temporal reference becomes more frequent with the increase of embedding depth. The distribution of relative temporal reference at level C2 is similar in both languages (168 of 292, 57% in Estonian; 93 of 156, 60% in Moksha). In the Estonian sample, we observe an increase of clauses with relative temporal reference also from C2 to C3 level (45 of 64 C3s have relative temporal reference, i.e. 70%). In Moksha we have only 14 observations at C3 level, which are equally distributed. Note, however, that the two Estonian clauses observed at the deepest level C5 have absolute temporal reference. This means that although DECs are on average temporally more dependent on the clauses of higher order than C1s, depth is not an obstacle for absolute temporal reference. Deep syntactic embedding does not necessarily correlate with semantic embedding in the temporal structure of the higher clauses. Also, examples (6) and (7) above demonstrated that DECs can have absolute or relative reference irrespectively of the temporal reference of the superordinate clause: all possible combinations of the values T-Abs and T-R are observed in the recursive structure.

4 Quantitative evaluation of data

4.1 Method and design

In §3.2–3.4, we discussed the distribution of the variables 'clause type', 'position', and 'temporal reference' in relation to the embedding depth. We observed two important trends. First, in both languages the share of complement clauses decreases with the embedding depth, while the share of relative clauses increases. Second, the share of clauses with absolute temporal reference decreases with the embedding depth, while the share of clauses with relative temporal reference increases. This trend seems more pronounced in Estonian since relative temporal reference is in general more frequent in the Moksha than in the Estonian sample which, as noted above, has to do with the frequency of non-finite clauses (with subordinators) in the two samples. The distribution of clause position relative to its matrix clause is very unbalanced, with only a few observations for center- and left-embedding, but moving from C1 to C2 level, the use of center-embedding increases in both samples.

These observations suggest that the answer to RQ1 formulated in §1.3 is positive with respect to all structural features. In sentences with deeply embedded clauses, C1s differ from DECs regarding the distributional properties of clause type, position, and temporal reference.

In this section, we statistically evaluate the strength of the observed tendencies by means of logistic regression analysis (cf. McCullagh & Nelder 1989; Dobson 1990; Hastie & Pregibon 1992). We have few observations for embedding depths C3 and deeper, therefore we aggregate all DECs to one level of variable, as we did in Tables 3 and 4. In this way, we treat the depth of embedded clauses as a binary dependent variable with DEC as 1 and C1 as 0. In the logistic regression modelling, we monitor how the probability of obtaining deeply embedded clauses changes in different sets of conditions.

The variables are summarized in Table 5. We treat 'clause type' and 'temporal reference' as independent variables (also called explanatory or predictor variables), the former with three levels of variation, the latter with two. The feature position, which we described in §3.3, has a very unbalanced distribution: left- and center-embeddings are too infrequent to be included in the analysis.

Although the general trends were similar in both languages, the retrograde of changes between C1s and DECs seemed to have different dynamics. Therefore, we include the independent variable 'language' to control the strength of language-specific effects.

Table 5. Variable description in the regression models

Variable	Levels	Type of variable
Depth of embedded clause	1 – DEC 0 – C1	Dependent
Clause type	Adverbial Complement Relative	Independent
Temporal reference	Absolute Relative	Independent
Language	Estonian Moksha	Independent

To answer our main research question RQ1, we must formulate the following questions concerning each independent variable studied in the available samples:

RQ2. Is any clause type more compatible with any embedding depth?
RQ3. Is any embedding depth more compatible with any type of temporal reference?
RQ4. Are there significant structural differences between Estonian and Moksha regarding the embedded clause architecture?

Our research questions are operationalized in the form of a set of null hypotheses listed below:

$H_{0.1}$: C1s and DECs have the same structural properties regarding the studied variables.
$H_{0.2}$: Adverbial, relative and complement clauses are equally compatible with C1s and DECs.
$H_{0.3}$: Absolute and relative temporal reference is equally compatible with C1s and DECs.
$H_{0.4}$: Moksha and Estonian have the same architecture of embedded clause.

4.2 Results

Our approach to logistic regression modeling is information theoretic, as recommended by Burnham and Anderson (2002) for observational data, rather than based on statistical hypothesis testing. We do not assume that there is only one correct model; instead, our goal is to select the most parsimonious model that adequately explains distributions in the data. We constructed in total seventeen models (see Table 6), considering all possible combinations of variables and their interactions. Thus, we compared very simple models with just one independent variable (see rows 10, 13, 16 of Table 6) with very complex, saturated models with all three independent variables and all possible interaction terms (row 6 of Table 6).

We used the Akaike Information Criterion (AIC) corrected for small sample size (AICc, Sugiura 1978; Hurvich & Tsai 1989; 1991) as criterion for the best-fitting model. Just like AIC, AICc takes into account two measures: the maximized log-likelihood function (LL) and a penalty for high number of estimated parameters (k). However, the AICc score includes an additional term that penalizes the high number of parameters in the case of small samples.[19] For our study, this means that AICc is more restrictive and therefore more reliable. We utilized the 'glm' and 'effects' functions (Venables & Ripley 2002) for model fitting, and the 'AICmodavag' package (Mazerolle 2023) for AICc calculation.

All models are listed in Table 6 along with their AICc values, the number of estimated parameters, and log-likelihood values. The Δ AICc column indicates the differences in AICc scores. The relative likelihood of a model (given in the AICc Weight column) can be further used to assess how confident we can be that a particular model outperforms the other models in the comparison.

[19] Compare the formulas: $AIC = 2k - 2\ln(\hat{L})$, $AICc = 2k * \frac{n}{n-k-1} - 2\ln(\hat{L})$, where k denotes the number of para-meters, n denotes the sample size, and $\hat{L}$ denotes the maximum of the likelihood function for the model.

Table 6. Comparison of models based on AICc. The interaction term is marked with a star.

No.	Model	K	AICc	Δ AICc	AICc Weight	Cum.Wt	LL
1.	Clause type + Language + Temporal reference + Temporal reference * Language	6	1254.18	0.00	0.64	0.64	-621.05
2.	Clause type + Language + Temporal reference + Temporal reference * Language+ Type of embedded clause * Language	8	1256.87	2.70	0.17	0.80	-620.36
3.	Clause type + Language + Temporal reference + Temporal reference * Language + Temporal reference * Type of embedded clause	8	1257.53	3.35	0.12	0.92	-620.69
4.	Clause type + Temporal reference + Language	5	1261.35	7.17	0.02	0.97	-625.64
5.	Clause type + Language + Temporal reference + Type of embedded clause * Language	7	1262.01	7.83	0.01	0.98	-623.95
6.	Clause type + Language + Temporal reference + Type of embedded clause * Language + Type of embedded clause * Temporal reference + Temporal reference * Language +Type of embedded clause * Language * Temporal reference	12	1263.64	9.46	0.01	0.99	-619.66
7.	Clause type + Temporal reference	4	1263.70	9.52	0.01	1.00	-627.83
8.	Clause type + Language + Temporal reference + Temporal reference * Type of embedded clause + Type of embedded clause*Language	9	1264.75	10.57	0.00	1.00	-623.28
9.	Clause type + Temporal reference + Type of embedded clause * Temporal reference	6	1266.45	12.27	0.00	1.00	-627.18
10.	Temporal reference	2	1314.34	60.17	0.00	1.00	-655.17
11.	Temporal reference + Language	3	1314.45	60.28	0.00	1.00	-654.21
12.	Clause type + Language + Temporal reference + Type of embedded clause * Temporal reference	7	1320.97	66.79	0.00	1.00	-653.43
13.	Clause type	3	1329.74	75.56	0.00	1.00	-661.86
14.	Clause type + Language	4	1329.81	75.64	0.00	1.00	-660.89
15.	Clause type + Language + Type of embedded clause * Language	6	1330.13	75.95	0.00	1.00	-659.02
16.	Language	2	1360.60	106.43	0.00	1.00	-678.30
17.	Temporal reference + Language + Temporal reference * Language	4	1363.59	109.41	0.00	1.00	-677.77

The best fitting Model 1 (row 1 in Table 6) has the lowest AICc value=1254.18 and accounts for 64% of the cumulative model weight. It models the probability of a deeply embedded clause using all three potentially available independent variables: clause type, temporal reference, and language. It also includes an interaction term between temporal reference and language. Model 2, which is the second-best (row 2 of Table 6), has an AICc value of 1256.87. Since differences greater than two indicate that the model with the lower AICc score should be preferred (Burnham & Anderson 2002: 70), we have good reasons to believe that Model 1 fits the data better. Model 2 is very similar to Model 1, but it includes an additional interaction term between clause type and language, which is not a significant predictor in the model itself. The evidence ratio calculated from the AICc weights of Model 1 and Model 2 is 3.55. It shows that the poorer fit of Model 2 could be partly due to the sample used, and that some of the variance included in Model 1 could be explained by the interaction between the clause type and language in other samples. Nonetheless, we will focus on the factors included in Model 1, as they significantly contribute to modeling the differences between DEC and C1.

Table 7 summarizes the estimation results. The coefficients are calculated for the base reference condition 'Moksha adverbial clause with absolute temporal reference', which is almost equally probable to appear in DECs and C1s (see Figure 4 below). All computed effects significantly change the probability of a clause being deeply embedded. Complement clauses are less likely to occur as DECs compared to adverbial clauses; this is visible from the negative sign of the estimates. Since the coefficient for relative clauses is positive, we can deduct that relative clauses have the highest chances of occurring as DECs from all three types. Since pure language effect is insignificant ($p=0.504$), we can assume that the clause type has similar impact on probability of a clause being deeply embedded in both languages. The positive coefficient 0.5901 ($p=0.011$) indicates that the relative temporal reference increases the probability of a clause to be deeply embedded. However, from the positive and significant interaction term, we observe that this effect is much stronger in Estonian (0.8712, $p<0.01$).

Table 7. Coefficients of the best-fit model. Base of reference: Moksha adverbial clauses with absolute temporal reference.

Level	Estimate	Std. Error	Pr(>|z|)
Intercept	-0.2786	0.2095	0.186
Complement Clause	-0.5814	0.1705	<0.01 *
Relative Clause	0.6747	0.1776	<0.01 *
Estonian	-0.1364	0.2041	0.504
Relative Reference	0.5901	0.2332	0.011 *
Relative Reference*Estonian	0.8712	0.2874	<0.01 *

Since a detailed comparison of the exact estimates' coefficient for the twelve possible conditions (language × clause type × temporal reference) is hard to follow, we transform the coefficients of the chosen model into the predicted probabilities of a clause being a DEC for each condition. The obtained probabilities and their confidence intervals are plotted in Figure 4, such that the mutual impact of all effects is clearly visible.

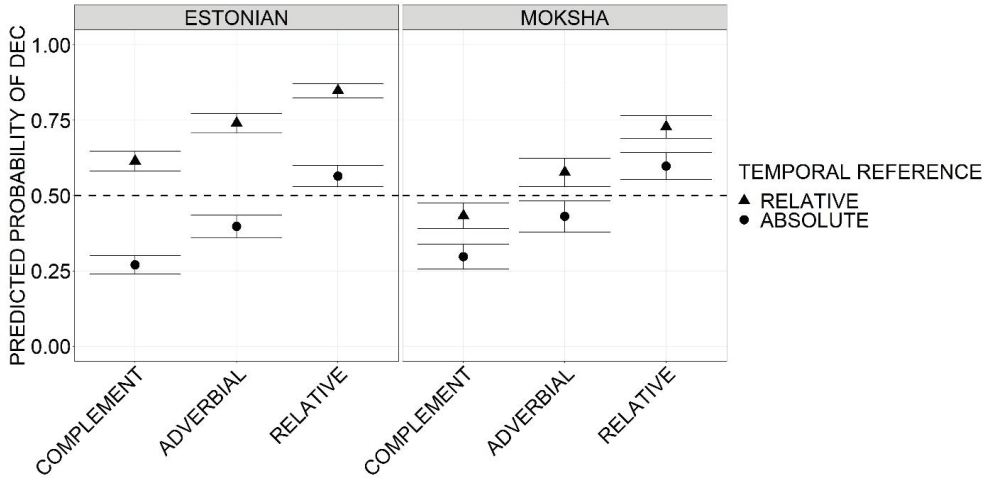


Figure 4. Predicted probabilities of DEC according to the best fitted model

In Figure 4, the dashed horizontal line at the level p=0.50 indicates equal predicted probabilities of a clause being a DEC and C1. Therefore, the points below the dashed line can be interpreted as more likely to occur in C1s, while those above as more likely to occur in DECs.[20] The three types of clauses clearly below the dashed line are complement and adverbial clauses with absolute temporal reference in both languages, and Moksha complement clauses with relative temporal reference.

The probability of complement clauses with absolute temporal reference being a DEC is p=0.27 for Estonian, and p=0.30 for Moksha. These clauses are least compatible with deep embedding in both languages, and they are roughly twice as probable to be a C1 as DEC.[21] Adverbial clauses with absolute temporal reference are 13% more likely to occur as DECs (p=0.40 for Estonian and p=0.43 for Moksha) than complement clauses with absolute temporal reference. In Moksha this probability equals the probability of a complement clause with relative temporal reference to appear as DEC.

In Estonian, the strong effect of relative temporal reference increases the probability of a complement clause with relative temporal reference being a DEC well above 0.5 (p=0.61). In fact, all clause types are more likely to occur as DECs in Estonian, provided their temporal reference is relative. The probabilities, in this case, are p=0.74 for adverbial clauses and p=0.85 for relative clauses. The latter are most compatible with DECs of all clause types. In other words, relative clauses with relative temporal reference are 5.5 times more likely to occur as DECs than as C1s. In comparison, the odds ratio is 2.84 for Estonian adverbial clauses and 1.56 for complement clauses.

In Moksha, the differences are smaller, but like in Estonian, both adverbial and relative clauses with relative temporal reference are more likely to occur as DECs than as C1s. For relative clauses (p=0.72), the probability of occurring as DEC is 2.5 times higher than the probability of occurring as C1, while for adverbial clauses (p=0.58), the odds ratio is only slightly above 1, that is, 1.32.

Finally, the generally high compatibility of relative clauses with deep embedding seems a common feature in the discussed languages, as even relative clauses with absolute temporal reference are somewhat more likely to occur as DECs than as C1s. The

[20] The probability of a counter-event, i.e., that a clause occurs in C1 is simply 1-p.

[21] We come to this conclusion by checking the proportion: p/(1-p), that is, odds ratio.

probability of this event is modelled as p=0.60 for Moksha and as p=0.56 for Estonian. These probabilities are similar to the probability of adverbial clauses with relative temporal reference to occur as DECs.

5 Discussion

Applying a logistic regression model, we detected significant structural differences in the data regarding the embedded clauses of the first and deeper orders in complex sentences. Therefore, we have good reasons to reject null-hypothesis H0.1 and assume the validity of the hypothesis that C1s do differ significantly from DECs regarding the distributional properties of their structural features. This gives a positive answer A1 to our main research question RQ1. Regarding the other null-hypotheses and research questions we find out that:

A2. The clause type is relevant for the embedding depth. Complement clauses seem to prefer first order embedding, while relative clauses seem more compatible with deep embedding. Adverbial clauses do not show clear behavior in this respect; their behavior seems, therefore, to be driven by their temporal reference (see A3 below). We can only speculate about the factors conditioning the observed difference in the distribution of complement and relative clauses. One possible explanation relates to the information-structuring in sentences with multiple embeddings. The great majority of complement clauses occur as objects of the governing verb, and a great majority of complement-taking verbs belong to one of the following classes: cognitive verbs ('know', 'think', 'believe', etc.), verbs of perception ('see', 'hear', 'feel', etc.), and utterance verbs ('say', 'ask', etc.). We assume that these verbs are likely to occur in the ultimate main clause, because they are often associated with the speech act. Laury and Helasvuo (2016; 2020) observed that the most common cognitive verbs in Finnish, *tietää* 'know', *ajatella* 'think', and *muistaa* 'remember', tend to occur in the first-person singular form. Considering that a 1SG participant often coincides with the source of the illocutionary force conveyed by the main clause, it is not surprising to see that complement clauses are the most common clause type at an embedding depth of one. Relative clauses, on the other hand, restrict (or comment on) specific referents, occurring at any embedding depth, and their information value for the message of the entire sentence is lower. It can be assumed that, in the flow of information, the object of narrator's cognition, perception or report would be more important and precede expressions specifying facts about individual referents (cf. Sinnemäki 2004: 81–85). This would account for the observed correlation between clause type and embedding depth.

A3. Our findings indicate that the relative temporal reference is a significant factor in determining the probability of a clause being deeply embedded. It is crucial to highlight that, in quantitative terms, clause type and temporal reference act as predictors of the probability of embedding depth independently of each other. The AICc-based ranking does not favor models with an interaction term between temporal reference and clause type, which we would expect if a particular type of temporal reference constituted a significant difference between C1 and DECs only in presence of a particular syntactic type of clause. In fact, such models (rows number 7, 9, 10 and 13 of Table 6) were ranked lower than the models with two or three independent effects (rows number 5 and 8 of Table 6).

Furthermore, the best model extended by the interaction term between clause type and temporality was ranked third. The likelihood ratio test used to compare the goodness of fit of the model ranked first and the model ranked third yielded insignificant results ($p=0.71$). Consequently, we have strong evidence for the conclusion that temporal reference and clause type are independent factors in clause depth prediction.

Our study provides a piece of evidence that DECs are generally less temporally independent than C1s in sentences with multiply embedded clauses. Note, however, that this is a trend, not an absolute rule, since the only observations of the rare fifth-order embedding are clauses with an absolute temporal reference.

A4. Although the distributions of temporal reference and clause type are similar in both languages, we observe some language-specific effects. While the differences in the probabilities of clauses with the absolute temporal reference to be deeply embedded do not differ significantly between the two languages, the probability of a clause with relative temporal reference to be deeply embedded is much higher for all types of Estonian clauses than for Moksha. This strong effect indirectly influences the patterns of clause types. In Moksha, complement clauses appear generally more compatible with C1s and relative clauses more compatible with DECs. In Estonian, we can generalize only over relative clauses, but not over complement clauses, as their preference is clearly related to the type of temporal reference.

What are the possible theoretical implications of the main result of this study, i.e., the positive answer to RQ1? Recursion in sentential embedding is typically understood as hierarchical organization that allows clauses of the same type to occur inside each other; e.g., the occurrence of a relative clause inside a relative clause (e.g. van der Hulst 2010: xxiv). Because of the limits of the sample, we did not study the reiteration of specific features, such as the frequency of embedding of complement clauses in complement clauses, or of left-embedded clauses in left-embedded clauses, or of clauses with absolute temporal reference in clauses with absolute temporal reference. Instead, we studied how the distribution of formal and functional features of clauses changes with the increase of embedding depth. We did not observe categorical differences between first-order and deeply embedded clauses but only probabilistic ones, demonstrating, however, that the distribution of properties of DECs in sentences with multiple embeddings is not entirely predictable from the distribution of properties of C1s.

The idea of recursion has been advocated by Generative Grammar, whose competence-centered deterministic view of language does not involve interest in the probability of feature-distribution in language performance. Therefore, definitions of recursion in clausal embedding (cf. e.g. Hollebrandse 2020) do not directly entail that the distribution of available properties is reiterated along the embedding cycle, i.e., that different embedding depths display exactly the same distributions of properties. Nonetheless, the mere idea of recursion as an infinite procedure which lacks termination conditions (Karlsson 2010b) easily leads to the assumption that the probability of a feature remains constant upon repeated application of the procedure.

The generative tradition concedes the existence of soft, syntax-external constraints on recursive embedding: observed limitations on recursion are typically explained in terms of processing load (e.g., Karlsson 2010a; Blasi et al. 2019: 3938). In this study, we focused on the variables ‘clause type’ and ‘time reference’, whose distribution is not related

to processing load, at least not in an obvious way.[22] We demonstrated that the variable 'embedding depth' is not entirely independent from the studied variables. Considering that the idea of recursion in natural language is accepted and often presupposed far beyond its original theoretical framework (see Karlsson 2007a for an overview), our answer to RQ1 deserves to be assessed and accommodated by prospective debates on recursion in clausal embedding.

6 Conclusions

The quantitative analysis of the structural properties of sentences with deeply embedded clauses revealed important distributions, which would have remained unnoticed by purely qualitative research. Specific findings of the study include:

i. Estonian seems to feature greater average embedding depth than Moksha. More extensive research will seek to validate and explain this observation, considering also extra-linguistic circumstances. We are tempted to assume that the age of the literary language, the spread of literacy in it among the population, the number and size of printed texts, and possibly the model languages (German and Russian, respectively) on which their language planning has been based, is reflected in the average depth of clausal embedding.

ii. We found statistically significant influence of the type of embedded clause on the depth of embedding. Complement clauses tend to be embedders (i.e., first-order embedded clauses embedding a further clause), while relative clauses tend to be DECs (i.e., embedded in an embedded clause).

iii. Smaller embedding depth correlates with a more frequent occurrence of absolute temporal reference, greater embedding depth with a more frequent occurrence of relative temporal reference.

iv. Estonian and Moksha differ in the relative strength of the factors clause type and temporal reference. In Moksha, the differences between first-order and deep embeddings are best explained by the type of clause (C1s complement vs. DECs relative), while the rest of variation is explained by temporal reference. In Estonian, the order is the opposite. The differences in embedding depths are best explained primarily by the temporal reference, and secondarily by the clause type, since only relative clauses in our Estonian data show a clear preference for DECs.

v. Due to the limited amount of data, we did not arrive at definite conclusions concerning the position variable. It seems that the use of center-embedding increases with the embedding depth, but this requires investigation based on larger corpus. Center-embedding causes processing difficulties; discontinuity is a challenge to language users' working memory, it poses limits to the size of center-embedded clauses and is likely to produce shorter clauses than tail-embedding. At the same time, Blasi et al. (2019) observed that depth correlates with clause length reduction: deeper embeddings appear to be shorter in the number of words. Therefore, center-embeddings appear suitable for use as DECs.

[22] Unlike 'position of clause' with constraints on center-embedding, or 'length of clause'.

It should be reiterated that our results say nothing about the syntax of C1s in sentences ending at an embedding depth of one, i.e., in sentences with only first-order embeddings. We hope that our findings will be tested in the future against a sample consisting of such sentences.

Recapitulating, we argue that DECs make an independent contribution to the study of clausal subordination. This study showed that in sentences with deeply embedded clauses, DECs are not entirely identical with first-order embeddings, i.e., that depth is a factor influencing the grammar of subordinate clauses. This, we believe, should be acknowledged in debates over the status of recursion in clause combining.

Acknowledgements

This research has been supported by Estonian Research Council grant STP2 “Exploring deep clausal embeddings in Finno-Ugric”. We are also grateful to Natalia Abrosimova, Timofey Arkhangelskiy, Jesse Holmes, Gwen Janda, Liina Lindström, Maarja-Liisa Pilvik, Maria Vilkuna, and to the anonymous reviewers for their valuable comments and recommendations. This study is dedicated to Remco van Pareren, the late researcher of Mordvin languages.

Abbreviations

ADD	additive focus
ALL	allative case
APP	active past participle
C	clause
CNG	connegative form
COMP	complementizer
CONJ	conjunctive mood
DAT	dative
DEC	deeply embedded clause
DEF	definite declination
GEN	genitive
H_0	null hypothesis
ELA	elative case
ILL	illative case
INE	inessive case
INF	infinitive
NEG	negative
NOM	nominative
O	object
PL	plural
POL	polar (question) marker
POSS	possessive suffix
PPP	passive past participle
PRS	present
PRT	partitive case
PST	past

PST1	first past tense
PST2	second past tense
PTCP	participle
Q	question
REFL	reflexive
REL	relative pronoun
RQ	research question
S	subject
SG	singular
T-ABS	absolute temporal reference
T-R	relative temporal reference
TRL	translative case

Data sources

CCLM = Arkhangelskiy, Timofey. 2019. *Corpus of contemporary Literary Moksha*. (Moksha.web-corpora.net) (Accessed 2021-09).

ENC 2019 = *Estonian National Corpus*, version 2019. Accessible in SketchEngine: (www.sketchengine.eu) (Accessed 2021-10-06).

References

Bartens, Raija. 1999. *Mordvalaiskielten rakenne ja kehitys*. (Mémoires de la Société Finno-Ougrienne 232). Helsinki: Suomalais-Ugrilainen Seura.

Blasi, Damian E. & Cotterell, Ryan & Wolf-Sonkin, Lawrence & Stoll, Sabine & Bickel, Balthasar & Baron, Marco. 2019. On the distribution of deep clausal embeddings: A large cross-linguistic study. In Korhonen, Anna & Traum, David & Màrquez, Lluís (eds.), *Proceedings of the 57th annual meeting of the Association for Computational Linguistics*, 3938–3943. Florence: Association for Computational Linguistics. https://doi.org/10.18653/v1/p19-1384

Burnham, Kenneth P. & Anderson, David R. 2002. *Model selection and multimodel inference: A practical information-theoretic approach*. New York: Springer. https://doi.org/10.1007/b97636

Buzakov, Ivan S. 1973. *Složnoe predloženie v mordovskix jazykax* [The complex sentence in Mordvin languages]. Saransk: Mordovskoe knižnoe izdatel'stvo.

Comrie, Bernard. 1985. *Tense*. Cambridge: Cambridge University Press.

Cristofaro, Sonia. 2003. *Subordination*. Oxford: Oxford University Press.

Dik, Simon C. & Hengeveld, Kees. 1991. The hierarchical structure of the clause and the typology of perception-verb complements. *Linguistics* 29(2). 231–259.

Dobson, Annette J. 1990. *An introduction to generalized linear models*. London: Chapman and Hall.

EG = Metslang, Helle & Erelt, Mati & Habicht, Külli & Hennoste, Tiit & Kasik, Reet & Teras, Pire & Viht, Annika & Asu, Eva Liina & Lindström, Liina & Lippus, Pärtel & Pajusalu, Renate & Plado, Helen & Rääbis, Andriela & Veismann, Ann. 2023. *Eesti grammatika* [Estonian grammar]. Tartu: Tartu Ülikooli kirjastus.

EKS = Erelt, Mati & Metslang, Helle (eds.). 2017. *Eesti keele süntaks* [Estonian Syntax]. (Eesti Keele Varamu 3). Tartu: Tartu Ülikooli kirjastus.

Fejes, Katalin B. 2006. Koreferencia-viszonyok a két- és többtagú összetett mondatban [Coreference relations in two- and multipart compound sentences]. *Nyelvtudomány* 2. 9–19.

Feoktistov, Aleksandr P. 1976. *Očerki po istorii formirovanija mordovskix pis'menno-literaturnyx jazykov (rannij period)* [Essays on the history of formation of Mordvin literary languages (early period)]. Moscow: Nauka.

Frank, Stefan L. & Trompenaars, Thijs & Vasishth, Shravan. 2016. Cross-linguistic differences in processing double-embedded relative clauses: Working-memory constraints or language statistics? *Cognitive Science* 40(3). 554–578. https://doi.org/10.1111/cogs.12247

Gibson, Edward. 1998. Linguistic complexity: Locality of syntactic dependencies. Cognition 68. 1–76. https://doi.org/10.1016/S0010-0277(98)00034-1

Givón, Talmy. 2001. *Syntax: An introduction, Volume II*. Amsterdam: John Benjamins. https://doi.org/10.1075/z.syn2

Hamari, Arja & Ajanki, Rigina. 2022. Mordvin (Erzya and Moksha). In Bakró-Nagy, Marianne & Laakso, Johanna & Skribnik, Elena (eds.), *The Oxford guide to the Uralic languages*, 392–431. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780198767664.003.0023

Hastie, Trevor J. & Pregibon, Daryl. 1992. Generalized linear models. Chapter 6. In Chambers, John M. & Hastie, Trevor J. (eds.), *Statistical models in S*. New York: Wadsworth & Brooks/Cole.

Hollebrandse, Bart. 2020. Indirect recursion: The importance of second-order embedding and its implications for cross-linguistic research. In Amaral, Luiz & Maia, Marcus & Nevins, Andrew & Roeper, Tom (eds.), *Recursion across Domains*, 35–47. Cambridge: Cambridge University Press.

van der Hulst, Harry. 2010. Re Recursion. In van der Hulst, Harry (ed.), *Recursion and human language* (Studies in Generative Grammar 140), xv–liii. Berlin & New York: De Gruyter Mouton. https://doi.org/10.1515/9783110219258

Hurvich, Clifford M. & Tsai, Chih-Ling. 1989. Regression and time series model selection in small samples. *Biometrika* 76. 297–307.

Hurvich, Clifford M. & Tsai, Chih-Ling. 1991. Bias of the corrected AIC criterion for underfitted regression and time series models. *Biometrika* 78. 499–509.

Ikola, Osmo & Palomäki, Ulla & Koitto, Anna-Kaisa. 1989. *Suomen murteiden lauseoppia ja tekstikielioppia* [Syntax and text grammar of the Finnish dialects] (Suomalaisen Kirjallisuuden Seuran Toimituksia 511). Helsinki: Suomalaisen Kirjallisuuden Seura.

ISK = Hakulinen, Auli & Vilkuna, Maria & Korhonen, Riitta & Koivisto, Vesa & Heinonen, Tarja Riitta & Alho, Irja. 2008. *Ison suomen kieliopin verkkoversio* [The comprehensive grammar of Finnish, online version] (Kotimaisten kielten tutkimuskeskuksen verkkojulkaisuja 5). (http://scripta.kotus.fi/visk/etusivu.php) (Accessed 2023-05-23).

Karlsson, Fred. 2007a. Constraints on multiple initial embedding of clauses. *International Journal of Corpus Linguistics* 12(1). 107–118. https://doi:10.1075/ijcl.12.1.07kar

Karlsson, Fred. 2007b. Constraints on multiple center-embedding of clauses. *Journal of Linguistics* 43(2). 365–392. https://doi.org/10.1017/S0022226707004616

Karlsson, Fred. 2009. Origin and maintenance of clausal embedding complexity. In Sampson, Geofrey & Gil, David & Trudgill, Peter (eds.), *Language complexity as an evolving variable*, 192–202. Oxford: Oxford University Press.

Karlsson, Fred. 2010a. Multiple final embedding of clauses. *International Journal of Corpus Linguistics* 15(1). 88–105. https://doi:10.1075/ijcl.15.1.04kar

Karlsson, Fred. 2010b. Recursion and iteration. In van der Hulst, Harry (ed.), *Recursion and human language* (Studies in Generative Grammar 140), 43–68. Berlin & New York: De Gruyter Mouton. https://doi.org/10.1515/9783110219258.43

Kehayov, Petar. 2016. Complementation marker semantics in Finnic (Estonian, Finnish, Karelian). In Boye, Kasper & Kehayov, Petar (eds.), *Complementizer semantics in European languages* (Empirical Approaches to Language Typology 57), 449–497. Berlin & Boston: De Gruyter Mouton. https://doi.org/10.1515/9783110416619-015

Kehayov, Petar & Boye, Kasper. 2016. Complementizer semantics in European languages: Overview and generalizations. In Boye, Kasper & Kehayov, Petar (eds.), *Complementizer semantics in European languages* (Empirical Approaches to Language Typology 57), 809–878. Berlin & Boston: De Gruyter Mouton. https://doi.org/10.1515/9783110416619-015

Keszler, Borbála. 2000. A többszörösen összetett mondatok elemzése. In Keszler, Borbála (ed.), *Magyar grammatika* [Hungarian grammar], 542–554. Budapest: Nemzeti Tankönyvkiadó.

Kiss, Katalin É. 2023. The (non-)finiteness of subordination correlates with basic word order: Evidence from Uralic. *Acta Linguistica Academica* 70(2). 171–194.

Laanekask, Heli & Erelt, Tiiu. 2003. Written Estonian. In Erelt, Mati (ed.), *Estonian language* (Linguistica Uralica, Supplementary Series / Volume 1), 273–342. Tallinn: Estonian Academy Publishers.

Laury, Ritva & Helasvuo, Marja-Liisa. 2016. Disclaiming epistemic access with 'know' and 'remember' in Finnish. In Lindström, Jan & Maschler, Yael & Pekarek Doehler, Simona (guest eds.), *Grammar and negative epistemics in talk-in-interaction: Cross-linguistic studies* (Special issue of *Journal of Pragmatics* 106), 80–96. https://doi.org/10.1016/j.pragma.2016.07.005

Laury, Ritva & Helasvuo, Marja-Liisa. 2020. The emergence and routinization of complex syntactic patterns formed with ajatella 'think' and tietää 'know' in Finnish talk-in-interaction. In Maschler, Yael & Pekarek Doehler, Simona & Lindström, Jan & Keevallik, Leelo (eds.), *Emergent Syntax for Conversation: Clausal patterns and the organization of action* (Studies in Language and Social Interaction 32), 55–85. Amsterdam: John Benjamins.

Laury, Ritva & Ono, Tsuyoshi. 2010. Recursion in Conversation: What speakers of Finnish and Japanese know how to do. In van der Hulst, Harry (ed.), *Recursion and human language* (Studies in Generative Grammar 140), 69–92. Berlin & New York: De Gruyter Mouton. https://doi.org/10.1515/9783110219258.69

Laury, Ritva & Ono, Tsuyoshi & Suzuki, Ryoko. 2021. Questioning the clause as a crosslinguistic unit in grammar and interaction. In Ono, Tsuyoshi & Laury, Ritva & Suzuki, Ryoko (eds.), *Usage-based and typological approaches to linguistic units* (Benjamins Current Topics 114), 123–160. Amsterdam & Philadelphia: John Benjamins. https://doi.org/10.1075/bct.114.06lau

Mazerolle, Marc J. 2023. *AICcmodavg: Model selection and multimodel inference based on (Q)AIC(c). R package version 2.3.3.* (https://cran.r-project.org/package=AICcmodavg) (Accessed 2024-03-08).
McCullagh, Peter & Nelder, John A. 1989. *Generalized linear models*. London: Chapman and Hall.
Mithun, Marianne. 1984. How to avoid subordination. In Dahlstrom, Amy & Macauley, Monica (eds.), *Papers selected from the parasession on subordination* (Berkeley Linguistics Society 10), 493–509. Berkeley: University of California. https://doi.org/10.3765/bls.v10i0.1937
Oakes, Michael P. 1998. *Statistics for corpus linguistics*. Edinburgh: Edinburg University Press. http://doi.org/10.1515/9781474471381
Pinker, Steven & Jackendoff, Ray. 2005. The faculty of language: What's special about it? *Cognition* 95(2). 201–236. https://doi.org/10.1016/j.cognition.2004.08.004
Progovac, Ljiljana. 2010. When clauses refuse to be recursive: An evolutionary perspective. In van der Hulst, Harry (ed.), *Recursion and human language*, 193–211. Berlin & New York: De Gruyter Mouton. https://doi.org/10.1515/9783110219258.193
R Core Team 2021. *R: A language and environment for statistical computing. R Foundation for Statistical Computing*, Vienna, Austria. (URL https://www.R-project.org/) (Accessed 2023-06-18).
Rácz, Endre. 1968. A többszörösen összetett mondat. In Rácz, Endre (ed.), *A mai Magyar nyelv* [Contemporary Hungarian language], 443–446. Budapest: Tankönyvkiadó.
Toldova et al. 2018 = Toldova, Svetlana Ju. & Xolodilova, Marija A. & Tatevosov, Sergej G. & Kaškin, Egor V. & Kozlov, Aleksej A. & Kozlov, Lev S. & Kuxto, Anton V. & Prizivinceva, Marija Ju. & Stenin, Ivan A. (eds.), *Ėlementy mokšanskogo jazyka v tipologičeskom osveščenii* [Elements of the Moksha language in a typological context]. Moskva: Buki Vedi.
Saarinen, Sirkka. 1991. Typological differences between the Volgaic languages. *Yearbook of the Linguistic Association of Finland* 4. 43–52.
Sgall, Petr. 1990. Absolutes und relatives Tempus. In Wagner, Karl Heinz & Widgen, Wolfgang (eds.), *Studien zur Grammatik und Sprachtheorie* (Bremer Linguistisches Kolloquium 2), 57–64. Bremen: Milde Multiprint.
Shagal, Ksenia. 2018. Participial systems in Uralic languages: An overview. *Journal of Estonian and Finno-Ugric Linguistics (JEFUL)* 9(1). 55–84. (https://ojs.utlib.ee/index.php/jeful/article/view/jeful.2018.9.1.03) (Accessed 2024-03-12).
Sinnemäki, Kaius. 2004. *Complex right-branching clauses*. University of Helsinki. (Unpublished master's thesis).
Skribnik, Elena. 2022. Clause combining. In Bakró-Nagy, Marianne & Laakso, Johanna & Skribnik, Elena (eds.), *The Oxford guide to the Uralic languages*, 996–1017. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780198767664.003.0053
Stefanowitsch, Anatol. 2020. *Corpus linguistics: A guide to the methodology* (Textbooks in Language Sciences 7). Berlin: Language Science Press.
Sugiura, Nariaki. 1978. Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics – Theory and Methods* 7(1). 13–26.
Venables, William N. & Ripley, Brian D. 2002. *Modern applied statistics with S* (4th ed.). New York: Springer.
Vilkuna, Maria. 2022. Word order. In Bakró-Nagy, Marianne & Laakso, Johanna & Skribnik, Elena (eds.), *The Oxford guide to the Uralic languages*, 950–960. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780198767664.003.0049
Ylikoski, Jussi. 2022. South Saami. In Bakró-Nagy, Marianne & Laakso, Johanna & Skribnik, Elena (eds.), *The Oxford guide to the Uralic languages*, 113–129. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780198767664.003.0008

Contact information:

Edyta Jurkiewicz-Rohrbacher
Universität Hamburg
Institute for Slavic Studies
e-mail: edyta.jurkiewicz-rohrbacher@uni-hamburg.de

Petar Kehayov
University of Tartu
Institute of Estonian and General Linguistics
email: petar.kehayov@ut.ee

suuri, iso vai *kookas? tärkeä, keskeinen* vai *merkittävä?* Miten ensikielisyys vaikuttaa lähisynonyymisten adjektiivien valintaan kaunokirjallisuuskontekstissa

Niina Kekki
Turun yliopisto

Ilmari Ivaska
Turun yliopisto

Abstrakti

Tarkastelemme tässä artikkelissa monimuuttuja-sekamallin avulla, minkä adjektiivin lähisynonyymisista pesueista *suuri/iso/kookas* ja *tärkeä/keskeinen/merkittävä* S2- ja S1-puhujat valitsevat kaunokirjallisessa tekstikontekstissa. Edistyneiden kielenoppijoiden lähisynonyymien käytön ja opettamisen ongelmat on tunnistettu jo varhain. Lähisynonyymivalinnat on havaittu toisen kielen oppimisen tutkimuksessa yhdeksi yleisimmistä leksikaalisen tason syistä sille, että vielä edistyneidenkin kielenoppijoiden kielenkäyttö voi vaikuttaa kontekstille epätyypilliseltä. Aiempi tutkimus on lisäksi osoittanut, että kaunokirjallisuuden lukeminen mahdollistaa sattumanvaraisen sanaston oppimisen niin ensikielisillä kuin toisen/vieraan kielen puhujilla.

Käyttöpohjaisen näkemyksen mukaan kieli ja sen oppiminen pohjautuvat luonnollisen kielen käyttöön. Kielenkäytön ytimessä ovat konstruktiot eli muoto-merkitys-käyttö-kombinaatiot, jotka muodostavat toisiinsa kytkeytyvän verkoston. Vertaamme S2- ja S1-puhujien adjektiivivalintoja niiden tyypilliseen käyttöön konstruktionäkökulmasta eli selvitämme, mitkä muuttujat selittävät sanavalintaa konstruktiossa, jossa adjektiivi toimii pääsanansa määritteenä. Varsinainen aineisto on kerätty kyselytutkimuksella, mutta käytämme valintaa selittävinä taustamuuttujina laajaa kirjoa erilaisia taustamuuttujia. Näihin kuuluu valinnanalaisen konstruktion ominaisuuksia (mm. sijamuoto), kielenkäyttäjäkohtaisia muuttujia (mm. kielitausta ja lukutottumukset) sekä useita korpuksista johdettuja, tarkastelunalaisen konstruktion tyypillistä käyttöä kuvaavia muuttujia (mm. adjektiivien ja niiden pääsanojen yhteisesiintymien suhteelliset vahvuudet kielen eri rekistereissä).

Tutkimuksen tuloksena havaittiin, että suomeksi lukevat S2-puhujat valitsevat adjektiiveja alkuperäisen kaunokirjallisen kontekstin mukaisemmin kuin muilla kielillä lukevat. S1-puhujilla ei vastaavaa vaikutusta ole. Konstruktioiden osalta tutkimus vahvistaa aiempaa tietoa siitä, että L1- ja edistyneet L2-puhujat hahmottavat konstruktioita huomattavan samantapaisesti. Tarkasteltujen kollokaatioiden tekstilajipreferenssillä ei ole suurta merkitystä korkeakoulutetuilla vastaajilla, mahdollisesti koska kyselyaineisto mittaa enemmän sanaston reseptiivistä kuin produktiivista hallintaa. Tulokset osoittavat, että tilastollisena menetelmänä käytetty monimuuttujainen sekamalli sopii tutkimuksessa tarkasteltavan ilmiön selittämiseen. Malli toimii molemmissa pesueissa hyvin, mutta kiinnostavasti se selittää selvästi paremmin *suuri*-pesueen adjektiivivalintojen vaihtelua. Synonymiaa käsittelevissä korpustutkimuksissa on usein tarkastelussa vain yksi lähisynonyymipari tai -pesue, joten tämän tutkimuksen pohjalta voi esittää, että synonyymiaa käsittelevässä tutkimuksessa yhden lähisynonyymipesueen tarkastelun pohjalta ei

VERTAISARVIOITU
KOLLEGIALT GRANSKAD
PEER-REVIEWED
www.tsv.fi/tunnus

Finnish Journal of Linguistics 37 (2024), 135–158
DOI: 10.61197/fjl.142204

voi kovin luotettavasti tehdä synonymiaa yleisesti koskevia tulkintoja. Lisäksi tutkimuksen tulokset kannustavat käyttämään kaunokirjallisuutta aikuisten kielenopetuksessa.

Avainsanat: kielenoppiminen, suomi, suomi toisena kielenä, synonymia, monimuuttujasekamalli, kaunokirjallisuus

1 Johdanto

Jos kirjoittaa internetin hakukoneeseen kysymyksen ”mitä eroa on sanoilla *iso* ja *suuri*”, saa hakutuloksina keskustelupalstaketjuja, joissa kielenkäyttäjät pohtivat näiden sanojen semantiikkaa. Vaikka ensikielinen puhuja osaa yleensä käyttää tällaisia lähisynonyymisia[1] sanoja intuitiivisesti, niiden välisen eron sanallistaminen voi olla vaikeaa. Kielenoppijalle puolestaan voi tuottaa päänvaivaa valita ilmaisu lähisynonyymien pesueesta, jossa ilmausten välillä on vain hyvin hienovaraisia merkityseroja. Ilmiön hankaluudesta huolimatta lähisynonyymien tunteminen mielletään yhdeksi sujuvan kielenkäytön tunnuspiirteeksi: ne osoittavat tekstilajien ja kielirekisterien tuntemusta, ja niiden avulla kieleen saa vaihtelua ja oikeanlaista sävyä (Edmonds & Hirst 2002; Jarvis 2013; Danglli & Abazaj 2014).

Edistyneiden kielenoppijoiden synonyymisten ilmausten käytön ja opettamisen ongelmat on tunnistettu jo varhain (esim. Martin 1984), ja lähisynonyymivalinnat on havaittu yhdeksi yleisimmistä leksikaalisen tason syistä sille, että vielä edistyneidenkin kielenoppijoiden kielenkäyttö voi vaikuttaa kontekstille epätyypilliseltä (Hasselgren 1994: 245). Vaikka kielenoppijoiden vaikeudet lähisynonyymien omaksumisessa ovat hyvin tiedossa, valtaosa synonymiatutkimuksesta on kuitenkin keskittynyt ensikielisten (englannin- tai muiden indoeurooppalaisten kielten puhujien) kielenkäyttöön (Liu & Zhong 2016: 242).

Epäodotuksenmukaisesti käytetyt lähisynonyymit saattavat vääristää merkityksiä tai tuottaa epätoivottuja sävyeroja (Danglli & Abazaj 2014: 632). Sanan osaamiseen ei riitä, että tietää sanan määritelmän tai käännöksen eri kielillä, vaan osaaminen vaatii myös tietoa siitä, millaisia ovat sanan eri muodot, minkä sanojen kanssa se tyypillisesti esiintyy ja missä konteksteissa sitä on sopivaa käyttää (Alderson ym. 2015: 101). Kaunokirjallisuuden lukemista pidetään yleisesti tehokkaana tapana oppia kieltä ja erityisesti uutta sanastoa (ks. esim. Krashen 1993; Kim 2004), oli kyse sitten ensikielestä tai myöhemmin opittavasta kielestä. Aiempi tutkimus on osoittanut, että kaunokirjallisuuden lukeminen mahdollistaa sattumanvaraisen sanaston oppimisen niin ensikielisillä kuin toisen/vieraan kielen puhujilla (esim. Reynolds 2015).

Tutkimme tässä artikkelissa määrällisin keinoin sitä, minkä lähisynonyymisen adjektiivin suomea ensikielenään (S1) ja edistyneellä tasolla[2] toisena kielenä (S2) puhuvat valitsevat kaunokirjallisessa tekstikontekstissa sekä millainen yhteys kaunokirjallisuuden lukemisella on synonyymisen adjektiivin valintaan aikuisilla S2- ja S1-puhujilla. Tarkastelun kohteena on kaksi lähisynonyymipesuetta: adjektiivi *suuri* ja sen lähisynonyymit *iso* ja *kookas* sekä adjektiivi *tärkeä* ja sen lähisynonyymit *keskeinen* ja *merkittävä*. Nämä ovat tutkimuksessa käytetyn kaunokirjallisuuskorpuksen yleisimmät synonyymiset adjektiivipesueet (ks. tarkemmin luku 3.1).

[1] Koska absoluuttista synonymiaa pidetään yleisesti kielenkäytölle epätodennäköisenä (esim. Cruse 1986; Lyons 1968), käytämme tutkimuksessa termiä *lähisynonyymi* kuvamaan sanoja, jotka ovat merkitykseltään ainakin osin niin samankaltaisia, että kielenkäyttäjät ymmärtävät niiden viittaavan samaan asiaan.

[2] Eurooppalaisen viitekehyksen taitotasoilla B–C.

Esimerkiksi monelle suomenoppijalle tuttu Kielitoimiston sanakirja (jatkossa KS, 2022) ei erottele näiden pesueiden variantteja toisistaan tyylillisten tai affektiivisten piirteiden perusteella. *tärkeä* mainitaan sekä *keskeisen* että *merkittävän* synonyyminä. *keskeinen* eroaa merkitykseltään kahdesta muusta pesueen adjektiivista lokatiivisissa yhteyksissä (*keskeinen sijainti*), ja on siten polyseemisempi kuin kaksi muuta pesueen tarkastelunalaista varianttia. KS (2022) esittelee *suuri*-adjektiivin variantit saman ensisijaisen esimerkin avulla: ”*suuri* ja vahva mies; *iso* mies; *kookas* mies”. Sekä *suurella* että *isolla* on sanakirjan mukaan polyseemista käyttöä vaikutukseltaan huomattavien tai merkittävien asioiden kuvaamisessa, mitä ei esiinny *kookkaan* määritelmissä. Suomenoppijoiden apuvälineeksi kehitetty korpuspohjainen ConLexis-sanakirja (2013) tarkentaa yleissanakirjasta löytyviä tietoja esittelemällä, mitkä ovat *suuren* ja *ison* tyypillisiä kollokaatteja. *suuren* kollokaatit se erittelee abstraktimmiksi ja *ison* konkreettisemmiksi. *kookas* on ConLexiksen mukaan huomattavasti harvinaisempi kuin *iso* ja *suuri*. Näistä tarkennuksista huolimatta suomenoppijoiden käyttämien sanakirjojen sana-artikkelit saattavat jättää mielikuvan, että osa tai kaikki lähisynonyymeistä ovat useimmissa tilanteissa keskenään vaihtoehtoisia.

Vertaamme S2- ja S1-puhujien adjektiivivalintoja niiden tyypilliseen käyttöön konstruktionäkökulmasta eli pyrimme selvittämään, mitkä muuttujat selittävät sanavalintaa konstruktiossa, jossa adjektiivi toimii pääsanansa määritteenä (esim. ***tärkeä/keskeinen/merkittävä*** *kokous*, ***suuri/iso/kookas*** *hankinta*). Varsinainen aineisto on kerätty kyselytutkimuksella, mutta käytämme valintaa mahdollisesti selittävinä taustamuuttujina laajaa kirjoa erilaisia taustamuuttujia, kuten valinnanalaisen konstruktion ja sen tyypillisen käytön ominaisuuksia sekä kielenkäyttäjäkohtaisia muuttujia. Näihin kuuluvat S1- ja S2-puhujien kaunokirjallisuuden lukeneisuutta tarkastelevat muuttujat. Tarkastelemme muuttujien vaikutusta monimuuttuja-sekamallin avulla, mikä mahdollistaa useiden keskenään erilaisten muuttujien vaikutuksen sekä niiden välisten suhteiden samanaikaisen tarkastelun.

Vertaamme tutkimuksessa S1- ja S2-puhujien kielenkäyttöä. Emme kuitenkaan arvota kielimuotoja suhteessa toisiinsa emmekä näe, että olisi olemassa jonkinlaista ”täydellistä”, ensikielistä kielimuotoa, jota kohti kielenoppija voisi pyrkiä (ks. esim. Deshors ym. 2018). Aiemmassa synonymiatutkimuksessa on huomattu, että ensikieliset puhujat eivät aina valitse samaa synonyymia samaan kontekstiin (Liu 2013). Enemmistön tekemä valinta on siitä huolimatta yleensä paras indikaattori ilmaisun kontekstisidonnaisesta merkityksestä (Liu & Zhong 2016: 244). Sovellamme siis tutkimuksessamme kontrastiiviselle oppijankielen tutkimukselle tyypillistä kuvailevaa lähestymistapaa, jossa S1- ja S2-puhujien erojen tarkastelun tarkoitus on kuvailla kielimuotoja ja niissä esiintyvää tyypillistä vaihtelua, ei arvioida niitä suhteessa toisiinsa tai ottaa kantaa yksittäisten ilmausten tai valintojen norminmukaisuuteen (Granger 2015).

2 Tutkimuksen teoreettiset lähtökohdat

2.1 Kaunokirjallisuuden lukemisen vaikutus sanaston oppimiseen

Lähestymme kieltä ja sen oppimista tässä artikkelissa käyttöpohjaisen näkemyksen mukaan, eli näemme, että sekä kieli itsessään että sen oppiminen pohjautuvat kielen käyttöön (esim. Tomasello 2003). Käyttöpohjaisessa näkökulmassa sosialisaatio on merkittävässä osassa kielenoppimista: osallistumalla vuorovaikutukseen kielen puhujien ja kieliympäristön kanssa oppija tulee osaksi kieliyhteisöä. Kielen oppiminen tarkoittaa siis pohjimmiltaan sitä, että oppija oppii tapoja tulla toimeen (kohdekielisen) maailman ja

sen merkitysten kanssa. (van Lier 2000: 246–247.) Kaunokirjallisuuden lukeminen voi monin tavoin toimia porttina kohdekielisen maailman ymmärtämiseen ja siitä osalliseksi tulemiseen (Kekki ym. 2023). Yleisesti uskotaan, että kaunokirjallisuuden ekstensiivinen eli omaehtoinen ja -tahtinen lukeminen kasvattaa niin ensikielisen kuin kielenoppijankin sanavarastoa tarjoamalla lukijalle mielekkään kontekstin suuremman sanamäärän kohtaamiseen, kuin ohjatussa opetustilanteessa olisi mahdollista (esim. Pietilä & Merikivi 2014; Reynolds 2015).

Sanojen omaksumista kaunokirjallisuutta lukemalla on tutkittu esimerkiksi mittaamalla, kuinka hyvin kielenoppijat oppivat kaunokirjallisessa tekstissä esiintyviä tai tekstiin lisättyjä tekaistuja (*nonce*) sanoja. Tulokset näistä tutkimuksista osoittavat, että sanavarasto voi kasvaa kaunokirjallisuutta lukemalla ainakin jonkin verran (esim. Waring & Nation 2004; Nation 2018; Reynolds & Ding 2022). Uusien sanojen omaksumisen todennäköisyyteen vaikuttaa muun muassa se, kuinka merkityksellisessä kontekstissa uusi sana tekstissä esiintyy (Webb 2008), kuinka usein ja tasapuolisesti sana esiintyy eri teksteissä sekä missä määrin uusi sana muistuttaa jotakin sanaa oppijan muissa osaamissa kielissä (Reynolds & Ding 2022).

Ensikieliset aikuiset näyttävät oppivan uusia sanoja kaunokirjallisuutta lukemalla tehokkaammin kuin kielenoppijat (Reynolds 2015: 121). Waringin ja Takakin (2003) tutkimukseen osallistuneet englanti vieraana kielenä -oppijat mielsivät tekaistun sanan olevan usein synonyyminen jollekin heidän jo osaamalleen (opittavan kielen) sanalle. Tämä antaa viitteitä siitä, että kaunokirjallisuuden lukeminen voi kasvattaa lukijan synonyymisten sanojen tuntemusta. Tekaisujen sanojen oppimista tarkastelevissa tutkimusasetelmissa luettavat tekstit ovat kuitenkin usein lyhyitä ja niiden lukemista mitataan laboratoriomaisissa olosuhteissa (Reynolds 2022). Kysymykseksi jää, millaisia pitkäaikaisvaikutuksia kokonaisten romaanien omaehtoisella lukemisella on sanavarastoon.

Monikielisten aikuislukijoiden kohdalla mielenkiintoinen kysymys on, millaisia ovat monella kielellä lukemisen vaikutukset, esimerkiksi millä tavoin vieraalla kielellä lukeminen vaikuttaa ensikielen hallintaan. Auttaako vieraalla kielellä lukeminen vain kohdekielen oppimisessa vai voiko se myös tukea (tai haitata) ensikielen taitoa vielä aikuisiälläkin? Yhä useammat suomea ensikielenään puhuvat nuoret lukevat vain englanniksi (Bono 2023), mikä on herättänyt opettajissa huolta ensikielenä puhuttavan suomen rapautumisesta. Esimerkiksi lukioikäisten unkarilaisnuorten vieraan kielen (englanti, ranska ja/tai venäjä) oppimisen vaikutusta ensikielen kirjoitustaidon kehittymiseen tutkineet Istvan Kecskes ja Tünde Papp (2000: 19, 30) esittävät pitkittäistutkimuksensa pohjalta, että vieraan kielen osaaminen edistyneellä tasolla vaikuttaa ensikielen käytön tapaan. Vaikutus näkyy ensisijaisesti ensikielen hienosyisemmässä käytössä, kuten monipuolisempien lauserakenteiden ja selektiivisemmän sanaston hallinnassa. Vieraan kielen oppiminen vaikuttaa ensikieleen kuitenkin vain, jos oppimisprosessi on tarpeeksi intensiivinen ja oppijan oma motivaatio korkea.

Prosessit, jotka mahdollistavat uusien sanojen oppimisen kaunokirjallisuutta lukemalla, ovat edelleen pääasiassa tuntemattomia (Pietilä & Merikivi 2014: 29). Tutkimus on lisäksi keskittynyt nuorempiin ja usein ensikielisiin lukijoihin, joten erityisesti monikielisten aikuisten osalta tutkimusta on vain vähän. Tutkimuksissa, joissa sanaston oppimista testataan viiveellä lukutapahtuman jälkeen, on huomattu, että heti lukemisen jälkeen muistetut uudet sanat unohtuvat usein nopeasti (esim. Waring & Takaki 2003). Monet sanaston satunnaista omaksumista käsittelevät tutkimukset käsittelevät ja testaavat opittavia sanoja yksittäisinä yksiköinä. Reynolds (2015: 124) huomauttaakin, että sanastonhallintaa käsittelevissä tutkimuksissa pitäisi kiinnittää enemmän huomiota kohdesa-

noja ympäröiviin, toistuviin malleihin. Omassa asetelmassamme pyrimme hahmottamaan konstruktiopohjaisen lähestymistavan avulla, onko kaunokirjallisuuden lukeneisuudella pitkäaikaista vaikutusta lähisynonyymisen sanaston käyttöön, ja kuinka muulla kuin omalla ensikielellä lukeminen vaikuttaa tilanteeseen.

2.2 Konstruktiopohjainen näkökulma lähisynonymiaan ja sen oppimiseen

Tämän artikkelin tutkimusasetelma perustuu ajatukseen kielen konstruktiopohjaisuudesta eli siitä, että kielen yksiköissä muoto, merkitys ja käyttö ovat erottamattomia (Goldberg 2003). Erilaajuiset rakenteet morfeemeista lausetyyppeihin ja sitäkin laajempiin yhdistelmiin voidaan mieltää konstruktioiksi. Esimerkiksi adjektiivimääritteen sisältävä substantiivilauseke on konstruktio, jonka käyttöön keskitymme tässä tutkimuksessa. Aiempi tutkimus antaa viitteitä siitä, että sekä ensikieliset puhujat että kielenoppijat hahmottavat tällaisia muoto-merkitys-käyttö-kombinaatioita eli konstruktioita huomattavan samantapaisesti (esim. Gries & Wulff 2005; Ivaska & Bernardini 2020). Ronald Langackerin juurtumisteorian (Langacker 1987: 59–60) mukaan mitä useammin konstruktiota käytetään, sitä automaattisemmaksi sen prosessointi muodostuu ja sitä hyväksyttävämpi se on puhujan intuitiolle. Vaikka pelkkä altistuminen kielelle ei riitä selittämään kielen oppimisen prosessia (Ellis 2008: 375), moniulotteinen tilastollinen analyysi kielenkäyttäjien synonyymisten konstruktioiden käytöstä tuottaa kielenoppimisen ymmärtämisen kannalta tärkeää tietoa siitä, millaisia valintamahdollisuuksia muoto-merkitys-käyttö-yhdistelmien käytössä on.

Lähisynonymian tutkimus lähtee yleensä liikkeelle synonyymisten ilmausten välisten erojen selvittämisestä. Oletus on, että jos kielessä on käytössä kaksi merkitykseltään samankaltaista muodoltaan poikkeavaa ilmausta, ne on otettu käyttöön erilaisten strategioiden seurauksena ja samankaltaisuudestaan huolimatta eroavat jollain lailla merkityksensä ja käyttönsä osalta (De Jonge 1993: 253). Erottavat piirteet voivat olla kielensisäisiä tai -ulkoisia, ja erojen suuruus vaikuttaa siihen, että jotkin lähisynonyymiset ilmaukset ovat synonyymisempiä kuin toiset (Cruse 1986: 267).

Kielensisäisistä piirteistä lähisynonyymivalintaa voi ohjata esimerkiksi tekstilaji: Jarmo Jantusen (2001: 172) tarkastelemassa akateemisessa tekstikorpuksessa käytetään melko tasaisesti *tärkeä*-pesueen monia variantteja, kun taas kaunokirjallisuuskorpuksessa *tärkeä* on huomattavasti kaikkia muita variantteja yleisempi. Bob De Jongen (1993: 534) italian lähisynonyymisiä *parere* ja *sembrare* (’näyttää’; ’vaikuttaa’) -verbien käyttöä tarkastelleen tutkimuksen mukaan verbit eivät ole mielivaltaisesti vaihtoehtoisia tai ainoastaan kielenkäyttäjän preferenssiin perustuvia, vaan niiden välillä on selkeitä funktionaalisia eroja. Tiheyttä ilmaisevien suomen kielen adjektiivien polysemiaa tarkastellut Anni Jääskeläinen (2023) havaitsi puolestaan, että *tiheän* ja *tiuhan* polyseemiset merkitykset ovat sidoksissa niiden käyttöön substantiivin määrite- tai adverbiaalikonstruktioissa. Kiinnostavasti siis ”jo yhden sanan eri polyseemisten merkitysten välillä voi olla suuria eroja esiintymiskonstruktioissa” (Jääskeläinen 2023: 399).

Korpuslingvistiikan menetelmät ovat mahdollistaneet viime vuosikymmeninä uudenlaisen tarkastelutavan lähisynonyymien tyypillisen käytön tunnistamiseksi. Erityisesti on tutkittu lähisynonyymisten verbien käyttöä (esim. italiassa De Jonge 1993; venäjässä Divjak & Gries 2006; suomessa Arppe 2008; englannissa Deshors & Gries 2016) ja kokoa kuvaavia adjektiiveja (esim. Biber ym. 1998; Gries & Otani 2009). Näiden tutkimusten tulokset antavat tukea teorialle, jonka mukaan leksikaalisten yksikköjen merkitys muodostuu suurelta osin kontekstin perusteella (Firth 1957). Tämä on erityisen selvää, kun tutkitaan

suhteellisia adjektiiveja, jollaisia tässä artikkelissa tarkasteltavat *tärkeä*- ja *suuri*-pesueiden adjektiivit ovat. Suhteellisten adjektiivien tulkinta riippuu vahvasti siitä pääsanasta, johon ne liitetään (Hakulinen ym. 2004: § 605): esimerkiksi *pieni koira* ja *iso rotta* voivat olla kokoluokaltaan samanlaiset, mutta pääsanoja on totuttu kuvaamaan erilaisilla skaaloilla.

Korpuspohjainen suomen kielen synonymiatutkimus on edelleen vähäistä ja tarvitaan enemmän tietoa siitä, kuinka erilaiset korpuslingvistiset menetelmät soveltuvat suomen kaltaiseen agglutinoivaan kieleen. Lisäksi suurin osa lähisynonyymeja koskevasta tutkimuksesta keskittyy yhteen synonyymipariin (esim. De Jonge 1993; Jantunen 2001; Kekki & Ivaska 2022) tai -pesueeseen (esim. Divjak & Gries 2006; Arppe 2008; Gries & Otani 2009). Ilmiö kaipaa lisävalotusta esimerkiksi sen suhteen, millä tavoin yhtä tietyssä konstruktiossa, kuten substantiivinmääritteenä, esiintyvää synonyymipesuetta koskevia tuloksia voidaan soveltaa muihin pesueisiin.

Toisen kielen tutkimuksessa on huomattu, että toisen kielen edistyneetkin käyttäjät voivat osata oppimaansa kieltä kieliopillisesti hyvin mutta tehdä käyttökontekstille epätyypillisiä valintoja (Ivaska 2015), kuten käyttää epätavanomaisia synonyymejä. Ensikieliset puhujat käyttävät tarkempia ilmaisuja, kun taas kielenoppijat turvautuvat usein esiintyviin sanoihin ja ilmaisuihin (Granger 2004: 132). Tämä voi tuottaa epätavanomaisia sanojen yhdistelmiä, kuten *keskeinen kokous*. Kielenoppijoiden ylikäyttämiin sanoihin on viitattu termillä leksikaaliset nallekarhut (Hasselgren 1994; Jantunen 2015): ne opitaan tunnistamaan neutraaleiksi jo kielenoppimisen alkutaipaleella ja tuntuvat siksi helpolta käyttää sillä seurauksella, että niiden käyttö laajenee myös konteksteihin, joissa ensikielinen puhuja valitsisi jonkin toisen ilmauksen (Ringbom 2007).

Siltä osin kun oppijankielen lähisynonyymien käyttöä on tutkittu, tutkimus on keskittynyt erityisesti leksikaaliseen kollokaationäkökulmaan eli siihen, minkä pääsanan yhteydessä mitäkin lähisynonyymia käytetään. Esimerkiksi Ching-Yin Lee ja Jyi-Shane Liu (2009: 252) esittävät kiinankielisten englanninoppijoiden lähisynonyymisia substantiiveja ja adverbeja koskevan tutkimuksensa pohjalta, että lähisynonyymien tyypillisten kollokaattien tunteminen on avainasemassa lähisynonyymien merkityseron ymmärtämisessä ja käytön oppimisessa. Konstruktioajattelun mukaan sanasto ei kuitenkaan ole irrallaan kieliopista tai syntaksista, vaan sanojen merkityksiä pitää lähestyä myös ne huomioiden *kollostrukturaalisesti* (Stefanowitsch & Gries 2003: 210). Konstruktionäkökulman huomioivien tutkimusten tulokset viittaavat siihen, että vähemmän käytettyjen lähisynonyymivarianttien omaksumiseen voi vaikuttaa negatiivisesti konstruktion kieliopillinen monimutkaisuus: oppijat siis turvautuvat ”oletusarvoiseen” varianttiin silloin, kun kognitiivinen prosessointikuorma on jo valmiiksi suuri (Deshors & Gries 2014; Deshors 2016; Kekki & Ivaska 2022). Jotta voimme ymmärtää kielenkäyttäjien tekemiä valintoja, tarvitsemme tietoa paitsi siitä sanasta, jota adjektiivi määrittää, myös määritekonstruktion kieliopillisista ominaisuuksista ja laajemmasta tekstikontekstista, jossa se esiintyy.

2.3 Tutkimuskysymykset

Lähestymme tutkimuksen tavoitetta seuraavien tutkimuskysymysten avulla:

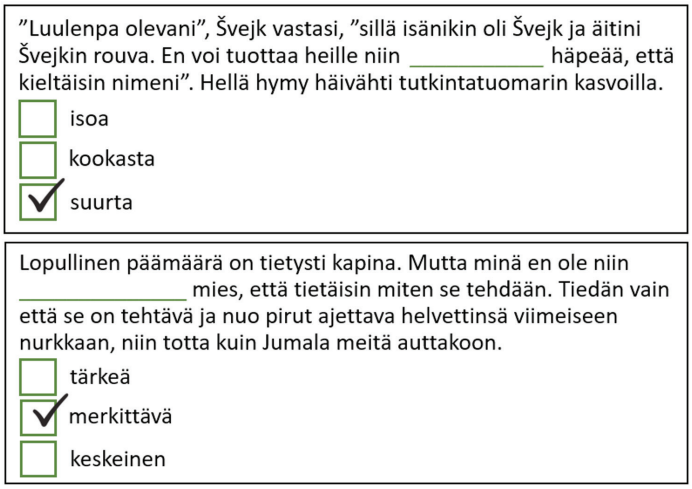
1) Millaisia adjektiivivalintoja S1- ja S2-puhujat tekevät alkuperäisee kaunokirjallisen tekstin kirjoittajaan verrattuna?
2) Millainen yhteys kaunokirjallisuuden lukeneisuudella on siihen, millaisia adjektiivivalintoja S1- ja S2-puhujat tekevät?
3) Mitkä tekstilajikäyttöön ja adjektiivimääritekonstruktioon liittyvät muuttujat vaikuttavat valintaan?

Kaunokirjalliset tekstit ovat kielellä luotua sanataidetta, joka voi kirjailijasta riippuen olla lähellä aikakautensa arkista kielenkäyttöä – tai hyvin kaukana siitä. Kaunokirjallisuuden tekstilajipiirteiden vuoksi oletamme, että sekä S1- että S2-puhujat tekevät osin erilaisia valintoja kuin kaunokirjallisen kontekstin alkuperäinen kirjoittaja, ja että vapaa-ajallaan kaunokirjallisuutta lukevat tekevät keskenään samansuuntaisempia valintoja.

3 Aineisto ja metodi

3.1 Aineistona korpuspohjainen synonymiakysely

Selvittääksemme kaunokirjallisuuden lukeneisuuden vaikutusta S1- ja S2-puhujien synonyymisten adjektiivien käyttöön, laadimme kaunokirjallisuuskorpusta hyödyntävän verkkokyselyn (ks. liite A; osin vastaavaa kokeellista menetelmää ovat hyödyntäneet esimerkiksi Antti Arppe ja Juhani Järvikivi (2007) sekä Dilin Liu ja Shouman Zhong (2016)). Kysely koostuu i) yleisistä taustatietokysymyksistä, ii) kaunokirjallisuuden lukemistottumuksia kartoittavista kysymyksistä, iii) monivalintakysymyksistä, jotka pakottavat valitsemaan yhden vaihtoehdon kolmen lähisynonyymisen adjektiivin pesueesta annettuun kaunokirjalliseen kontekstiin sekä iv) avoimista kysymyksistä, jotka selvittävät vastaajan syitä adjektiivien valinnalle. Tässä artikkelissa keskitymme määrällisen analyysimenetelmän avulla selvittämään, kuinka kaunokirjallisuuden lukemistottumukset liittyvät monivalintakysymyksissä tehtyyn adjektiivin valintaan. Monivalintakysymysten rakenne molempien lähisynonyymipesueiden osalta näkyy kuviossa 1.



Kuvio 1. Kaksi esimerkkiä aineistona käytettävän kyselyn monivalintakysymysten rakenteesta. Esimerkeissä on merkittynä kaunokirjallisen tekstin alkuperäinen adjektiivivalinta.

Tarkasteltavien lähisynonyymien osalta kysely on rakennettu korpusvetoisesti. Kyselyn pohjana olemme käyttäneet InterCorp Finnish v12 -korpuksen (Fárová & Vavřín 2019) suomenkielisiä kaunokirjallisuustekstejä. Valitsimme tarkasteltaviksi adjektiiveiksi *suuren* ja *tärkeän*, koska ne ovat korpuksen frekventeimmät adjektiivit, joilla on korpuksessa kolme vähintään 50 kertaa esiintyvää lähisynonyymia. *tärkeä/keskeinen*-lähisynonyymiparin käyttöä on tutkittu aiemmin S1-puhujien (Jantunen 2001; Vanhatalo 2003) sekä S2-puhujien predikatiivikonstruktion käytön (Kekki & Ivaska 2022) näkökul-

masta. Tässä artikkelissa keskitymme lähisynonyymien käyttöön substantiivin määritteinä, minkä lisäksi laajennamme synonyymiparista saatua tietoa ottamalla mukaan kolmannen synonyymivariantin ja vertailemalla tuloksia toiseen lähisynonyymipesueeseen.

Kyselyn pilotissa oli mukana neljä lähisynonyymia per synonyymipesue, mutta niiden määrä päätettiin karsia kolmeen, jotta vastausaika pysyisi kohtuullisena. Kyselyn monivalintakysymykset (ks. liite A) on laadittu tarkasteltavien lähisynonyymien kollokaattiesiintymien pohjalta: jokaiselle lähisynonyymille on valittu korpuksesta 10 esimerkkikatkelmaa, jossa adjektiivi määrittää sille tyypillistä pääsanaa eli substantiivikollokaattia. Kollokaatit etsittiin korpuksesta saatavien MI-arvojen (*mutual information score*) avulla, jotka mittaavat kahden leksikaalisen yhteisesiintymän todennäköisyyttä verrattuna sanojen esiintymiseen yksinään (Church & Hanks 1990). Mitä korkeampi MI-arvo, sitä todennäköisemmin pääsana esiintyy tutkitun adjektiivin kanssa kuin sattumanvaraisesti. Koska pesueiden vähiten frekventeillä adjektiiveilla (*kookas*, *keskeinen*) ei ollut riittävästi merkitseviä kollokaatteja, näiden adjektiivien tapauksessa esimerkkikatkelmiin on otettu mukaan myös konteksteja, joissa adjektiivi esiintyy minkä tahansa substantiivin määritteenä. Katkelmat on rajattu noin kolmen virkkeen mittaisiksi, ja niistä noin puolet on alun perin kirjoitettu suomeksi, noin puolet on suomeksi käännettyä kirjallisuutta. Osasta katkelmia on mahdollista tunnistaa alkuperäinen kaunokirjallinen teos esimerkiksi katkelmassa esiintyvien henkilöhahmojen nimien perusteella.

Vastausväsymyksen minimoimiseksi sekä adjektiivivaihtoehdot että kysymysten järjestys on satunnaistettu, kuitenkin niin, että kaikki vastaajat vastasivat ensin *suuri*-pesueen (30 katkelmaa) ja sitten *tärkeä*-pesueen (30 katkelmaa) monivalintakysymyksiin. Kyselyn keskimääräinen vastausaika oli 30 minuuttia. Vastaajia ohjattiin tekemään synonyymivalintansa seuraavanlaisella ohjeistuksella: ”Seuraavat esimerkit ovat kaunokirjallisista teksteistä. Valitse sana, joka mielestäsi sopii parhaiten kontekstiin. Johonkin kysymykseen voi olla mielestäsi enemmän kuin yksi oikea vastaus, mutta valitse se vaihtoehto, joka tuntuu sopivimmalta intuitiosi mukaan”.

3.2 Kyselyn osallistujat

Kyselyä[3] levitettiin suomalaisten korkeakoulujen postituslistojen kautta, sillä tavoitteena oli kerätä vastauksia mahdollisimman vertailukelpoiselta joukolta korkeakoulutettuja S1- ja S2-puhujia. Kyselyn vastaajina oli 234 ensikielistä suomenpuhujaa ja 28 suomenoppijaa. Molempien ryhmien kaikki vastaajat olivat korkeakoulutettuja. Vastaajamäärä on odotuksenmukainen ottaen huomioon, kuinka monta aikuisena suomea oppinutta, edistynyttä kielenkäyttäjää Suomen korkeakoulujen piirissä voi arvella olevan suhteessa suomea ensikielenään puhuviin. Vastaajien iän keskiarvo oli 31,4 vuotta (keskihajonta = 9,6), ja selvä enemmistö vastaajista (73,7 %) oli naisia. Naisten suuri osuus vastaajissa voi kertoa siitä, että heissä on ollut muita sukupuolia enemmän kyselyn aiheesta kiinnostuneita. Samoin kysely on saattanut kiinnostaa lukemista keskivertoa enemmän harrastavia ihmisiä. Molempien ryhmien vastaajista noin puolet vastaajista raportoi lukevansa kaunokirjallisuutta vapaa-ajallaan paljon: S1-puhujista 121 (51,7 %) ja S2-puhujista 14 (50 %) lukee vähintään muutamia kertoja viikossa. ”Ei koskaan” lukevia oli S1-vastaajista ainoastaan 1 (0,4 %) ja S2-vastaajista 2 (7,1 %).

[3] Kyselyn toteuttamisessa noudatettiin Tutkimuseettisen neuvottelukunnan (2019) ohjeistusta. Kyselyyn vastattiin anonyymisti. Aineiston käsittelyn yhteydessä on varmistettu, ettei vastaajien tunnistaminen ole mahdollista myöskään epäsuorasti esimerkiksi taustatietoja yhdistämällä. Kyselyyn vastaaminen perustui vapaaehtoisuuteen eikä vastaamisesta saanut palkkiota. Vastaajilta kerättiin informoitu kirjallinen suostumus vastausten käyttöön tutkimusaineistona.

Suomenoppijoiden osalta kysely suunnattiin saateviestissä edistyneen tason kielenoppijoille. Koska kaunokirjallisuudelle tyypillinen rikas kieli ja synonymia kielellisenä ilmiönä ylipäätään eivät vielä kuulu alkeistason kielenoppijoiden osaamisprofiiliin, mahdolliset alkeistason vastaajat olisivat luultavasti joutuneet turvautumaan vahvasti arvaamiseen vastauksissaan (Liu & Zhong 2016: 243). Tieto S2-vastaajien kielitaidosta perustuu heidän omaan arvioonsa: S2-vastaajista 14 (50 %) raportoi olevansa Eurooppalaisen viitekehyksen B- ja 14 (50 %) C-tasolla. Kaikki S2-vastaajat olivat oppineet suomea vasta aikuisiällä. S2-vastaajat raportoivat 14 eri ensikieltä (albania, arabia, englanti, espanja, italia, kiina, kurdi, persia/farsi, puola, saksa, serbokroatia, turkki, unkari, venäjä). Koska yksittäisellä kielellä on aineistossa niin vähän puhujia eikä kieliä saa mielekkäästi ryhmiteltyä, ensikieliä ei voitu käyttää analyysin muuttujina. Mahdollisen ensikielen siirtovaikutuksen tarkastelun sijaan tutkimus keskittyy kuvaamaan edistyneen oppijansuomen tyypillisyyksiä saatavilla olevan, ensikielitaustaltaan heterogeenisen aineiston avulla.

3.3 Monimuuttuja-sekamalli

Käytämme tässä tutkimuksessa pääasiallisena tilastollisena menetelmänä monimuuttujaista logistista regressio-sekamallia. Tällaisten mallien avulla pyritään mallintamaan jonkin vastemuuttujan arvo tarkastelemalla samanaikaisesti useita selittäviä muuttujia. Sekamalleissa osa selittävistä muuttujista on niin sanottuja kiinteitä muuttujia eli joko numeerisia tai sellaisia kategorisia muuttujia, joiden kaikki mahdolliset arvot tiedetään etukäteen. Osa puolestaan on niin sanottuja satunnaismuuttujia eli sellaisia kategorisia muuttujia, joiden kaikkia mahdollisia arvoja ei tiedetä etukäteen ja jotka voivat eri aineisto-otannassa saada myös muita arvoja. Monimuuttujamallin etuna verrattuna muuttujien tarkasteluun yksittäin on se, että se mahdollistaa yhden muuttujan tarkastelun siten, että muiden muuttujien vaikutus on vakioitu.

Tilastollisten regressiomallien soveltuvuutta arvioidaan usein kolmen eri suureen avulla: niin sanottu R^2-arvo kuvaa sitä, missä määrin malli kykenee tavoittamaan aineistossa havaittavan vaihtelun. Suureen arvot vaihtelevat välillä 0–1, jossa suuremmat arvot indikoivat vaihtelun paremmin tavoittavaa mallia. Sekamalleissa tämä suure jakautuu edelleen marginaaliseen ja ehdolliseen osaan, joista ensimmäinen kuvaa kiinteiden muuttujien osuutta ja jälkimmäinen kiinteiden ja satunnaismuuttujien kokonaisuutta. R^2-suureelle ei yleensä aseteta minkäänlaisia raja-arvoja, sillä nämä vaihtelevat tutkimusasetelmien ja aineistojen välillä suuresti. Tämän lisäksi arviointiin käytetään usein niin sanottua C-arvoa, joka kuvaa sitä, missä määrin malli soveltuu vastemuuttujansa mallintamiseen. Suureen arvot vaihtelevat välillä 0,5–1, ja 0,8 ylittävien arvojen katsotaan yleisesti indikoivan, että malli soveltuu kulloiseenkin tehtävään (esim. Gries 2021: 327). Mallin toimivuutta voidaan arvioida niin ikään sen perusteella, miten hyvin se kykenee ennustamaan vastemuuttujan arvot – tässä tutkimuksessa siis sitä, ovatko kielenkäyttäjät valinneet alkuperäisessä kontekstissa käytetyn adjektiivin. Tässä tapauksessa mahdolliset arvot ovat siis välillä 0–1, joissa suuremmat arvot kuvaavat paremmin ennustavaa mallia. Tässä vertailun minimiarvona pidetään usein vastemuuttujan yleisempää arvoa eli sitä arvoa, johon päästään arvaamalla aina yleisin vaihtoehto. Kuvaamme seuraavaksi kunkin tässä tutkimuksessa käytetyn muuttujan, perustellen samalla niiden roolin tässä tutkimuksessa. Tämän jälkeen esittelemme käyttämämme sekamallin yhtenä kokonaisuutena.[4]

[4] Tilastollinen analyysi on toteutettu R-ohjelmointiympäristössä (R Core Team 2022). Kyselylomake (Liite A), tilastollisessa analyysissa käytetty aineisto ja R-koodi ovat saatavilla osoitteessa: <https://osf.io/d2b9y/>

Tässä tutkimuksessa vastemuuttuja on luonteeltaan binäärinen: se kuvaa sitä, onko vastaaja valinnut lähisynonyymisten adjektiivien joukosta alkuperäisen käyttökontekstin mukaisen adjektiivin (alkuperäinen kirjailijan tai teoksen suomentajan **valinta**). Tutkimuksen keskeisenä tavoitteena on selvittää ensikielisyyden ja kaunokirjallisuuden lukemisen vaikutusta ja niiden välistä yhteyttä suhteessa lähisynonyymisten adjektiivien valintaan. Muuttuja **profiili** kuvaa, käyttääkö vastaaja suomea ensikielenään vai toisena kielenä.

Kaunokirjallisuuden lukeneisuuden määrää selvitetään kahdella muuttujalla: **kuinka usein** ja **kuinka monta kirjaa vuodessa keskimäärin vastaajat lukevat**. Pietilän ja Merikiven (2014) tutkimuksessa englantia vapaa-ajallaan lukevat nuoret pärjäsivät merkitsevästi paremmin vieraana kielenä opittavan englannin sanaston laajuutta mittaavissa testeissä kuin nuoret, jotka eivät lukeneet koulun ulkopuolella. Nuorten vieraan kielen sanavarasto oli myös sitä laajempi, mitä useammin he lukivat. (Pietilä & Merikivi 2014: 33–34.) Kummankin muuttujan osalta vastaajat valitsivat kyselyssämme seitsemän sanallisesti kuvatun vaihtoehdon väliltä parhaiten omia tottumuksiaan kuvaavan (ks. tarkemmin taulukko 2). Olemme muuntaneet muuttujat numeeriseksi siten, että sen pienin arvo (1) vastaa kyselyn määrällisesti pienintä vaihtoehtoa (esim. *En koskaan lue vapaa-ajallani kaunokirjallisuutta*) ja suurin arvo (7) vastaa kyselyn määrällisesti suurinta vaihtoehtoa (esim. *Luen vapaa-ajallani joka päivä*). Lukeneisuuden määrän lisäksi mallissa käytetään selittävänä muuttujana sitä, **millä kielellä** vastaajat pääasiassa lukevat. Tämän muuttujan avulla selvitetään, onko lähisynonyymien käytön kannalta väliä, lukeeko kaunokirjallisuutta nimenomaan sillä kielellä, jonka osaamista tutkitaan. Tämän kaksiarvoisen muuttujan arvoina ovat siis *suomi* ja *muut kielet*.

Useat tutkimukset nostavat esiin kollokaatioiden tärkeyden paitsi lähisynonyymisten sanojen (esim. Xiao & McEnery 2006), myös yleisesti toisen kielen kohdekielisen käytön (esim. Lu 2017) oppimisessa. Synonyymipesueiden kunkin adjektiivin ja sen pääsanan välistä yhteyttä mallinnetaan tässä tutkimuksessa kolmella muuttujalla: i) käytetyn adjektiivin ja sen kanssa esiintyvän pääsanan eli **kollokaation assosiaation vahvuudella** kaunokirjallisuuskorpuksessa, ii) **kollokaation kokonaisfrekvenssillä** kaunokirjallisuuskorpuksessa ja iii) **kollokaation tekstilajipreferenssillä**, eli sillä, missä tekstilajissa käytetyn kollokaation välinen assosiaatio on voimakkain. Sanojen välisen yhteisesiintymisen on osoitettu kiertyvän vahvasti kielitaitoon ja kielen idiomaattisuuteen – ja vapaan valinnan ja idiomaattisen valinnan välisen jatkumon kartoittamista (engl. *open choice principle* vs. *idiom principle*) pidetään usein yhtenä modernin korpuskielitieteen lähtökohtana (ks. esim. Sinclair 1991: 109–110). Sanojen välisen assosiaation mittaamiseen on käytetty monia eri mittareita, joista monissa ongelmana on se, että ne yhdistävät kaksi toisistaan erillistä ilmiötä – assosiaation vahvuuden ja yhteisesiintymisen käyttöfrekvenssin – niin, että näiden vaikutusta ei voida erotella toisistaan (Gries 2022). Olemmekin tässä tutkimuksessa erottaneet nämä toisistaan niin, että assosiaation mittarina käytetään alkuperäisessä tekstissä käytetyn adjektiivin ja pääsanan suhdetta kuvaavaa logaritmista vetosuhdetta (*log(arithmic) odds ratio*; laskukaavasta, ks. Gries 2022: 7) ja frekvenssin mittarina yhteisesiintymien määrää kaunokirjallisuuskorpuksessa.

Lähisynonyymien käytön erot ovat osin tekstilajisidonnaisia, sillä sekä lähisynonyymisten adjektiivien että niiden kanssa esiintyvien pääsanojen jakauma on erilainen eri tekstilajeissa (Biber ym. 1998; Biber & Conrad 2005). Tekstilajipreferenssin laskemiseksi haimme kyselyssä esiintyvät kollokaatiot neljästä kirjoitetun kielen korpuksesta, jotka näkyvät taulukossa 1. Tämän jälkeen laskimme vastaajan valitseman adjektiivin ja kunkin

pääsanan muodostaman kollokaation assosiaation voimakkuuden kussakin korpuksessa edellä kuvatulla tavalla. Muuttujan arvona on se tekstilaji, jossa assosiaatio oli kaikkein voimakkain eli jossa yhteisesiintyminen on kaikkein tiiveintä.

Taulukko 1. Tekstilajipreferenssin laskemiseen käytetyt korpuslähteet

kaunokirjallisuustekstit	kaannossuomi-korp (*akateemiset tekstit, tietotekstit, biografiat* jätetty haun ulkopuolelle)
sanomalehtitekstit	lehdet90ff-v2 (valittuna *muut lehdet; tiedelehdet* jätetty haun ulkopuolelle)
akateemiset tekstit	e-thesis-fi (valittuna *väitöskirjat*)
internetkeskustelut	suomi24-2001-2020-korp

Koska tutkimuksen keskiössä on ensikielisyyden vaikutus lähisynonyymivalintaan, kielenkäyttäjäprofiili on sisällytetty malliin ns. vuorovaikutusmuuttujana kaikkiin muihin edellä kuvattuihin kiinteisiin muuttujiin. Näin mallin avulla voidaan tarkastella kunkin muuttujan osalta sitä, eroavatko S1- ja S2-kielenkäyttäjät siltä osin toisistaan.

Yksittäiset adjektiivivalinnat eivät ole kyselyssä riippumattomia toisistaan, sillä yhtäältä yhden vastaajan vastaukset voivat heijastaa idiosynkraattisia piirteitä ja toisaalta vastaukset tiettyyn kysymykseen ovat väistämättä sidoksissa ko. käyttökontekstiin myös muilta kuin edellä kuvattujen muuttujien osalta. Otamme nämä riippuvuussuhteet huomioon sisällyttämällä **vastaajan** ja **kysymyksen** yksilöllisen tunnisteen malliin satunnaismuuttujina. Kielen rakenteellisena selittävänä tekijänä käytämme **sijamuotoa**. Aiemman tutkimuksen (Jantunen 2001; Kekki & Ivaska 2022) perusteella sijamuoto vaikuttaa merkitsevästi siihen, käyttääkö puhuja akateemisessa tekstissä *tärkeää* vai *keskeistä*. Kyselyyn poimitut tekstikatkelmat on valittu adjektiivin ja pääsanan perusteella, ja aineistossa ei esiinny kummassakaan synonyymipesueessa kaikkia sijamuotoja. Tästä syystä sijamuoto on sisällytetty malliin satunnaismuuttujana, jolloin laskennallinen malli ei käsittele aineistossa havaittujen vaihtoehtojen joukkoa sulkeisena. Muuttujien tyypit ja tasot on koottu taulukkoon 2.

Taulukko 2. Analyysissä käytetyt muuttujat ja niiden tasot

Tyyppi	Muuttuja	Tasot
Vastemuuttuja	vastaako vastaajan VALINTA alkuperäistä kontekstia	kyllä, ei
Kielenkäyttäjä-kohtaiset	PROFIILI	S1 (n=234), S2 (n=28)
	KUINKA MONTA KIRJAA LUKEE VUODESSA	0−1 (1), 2−5 (2), 6−10 (3), 11−15 (4), 6−20 (5), 21−30 (6), >30 (7)
	KUINKA USEIN LUKEE	en koskaan (1), harvemmin (2), noin kerran kuukaudessa (3), muutamia kertoja kuukaudessa (4), noin kerran viikossa (5), muutamia kertoja viikossa (6), joka päivä (7)
	MILLÄ KIELELLÄ LUKEE	suomeksi, muulla kielellä kuin suomeksi
Muista korpuksista johdetut	ALKUPERÄISEN KOLLOKAATION FREKVENSSI kaunokirjallisuus-korpuksessa	*suuri*-pesue: 0…115 *tärkeä*-pesue: 0…127
	KOLLOKAATION TEKSTILAJI-PREFERENSSI	kaunokirjallisuus internetkeskustelut akateemiset tekstit sanomalehtitekstit
Konstruktio-kohtaiset	ALKUPERÄINEN kirjailijan/suomentajan VALINTA	*suuri, iso, kookas* *tärkeä, keskeinen, merkittävä*
	adjektiivin ja sen pääsanan ASSOSIAATION VAHVUUS	*suuri*-pesue: 0,935…9,704 *tärkeä*-pesue: 1,133…8,392
Satunnais-muuttujat	määritettävän sanan SIJAMUOTO	*suuri*-pesue: ablatiivi, adessiivi, allatiivi, genetiivi, illatiivi, inessiivi, instruktiivi, nominatiivi, partitiivi *tärkeä*-pesue: adessiivi, elatiivi, genetiivi, inessiivi, nominatiivi, partitiivi, translatiivi
	KYSYMYS	Q1−Q60
	VASTAAJA	N=262

4 Tulokset

4.1 Mallien kokonaiskuva

***suuri*-pesue.** Ensimmäisen mallin avulla pyrimme mallintamaan edellä (ks. luku 3.3) kuvattujen muuttujien avulla sitä, mitkä seikat korreloivat sen kanssa, valitseeko kielenkäyttäjä *suuri*-pesueesta sen adjektiivin, jota kulloinkin tarkasteltavassa kaunokirjallisessa teoksessa on vastaavassa kontekstissa käytetty. *suuri*-pesuetta koskeva malli toimii kaikkiaan erittäin hyvin: kiinteiden muuttujien selitysastetta kuvaava marginaalinen R^2-arvo on 0,50, kun myös satunnaismuuttujat sisältävä ehdollinen R^2-arvo on peräti 0,91. Mallin C-arvo on 0,957 ja sen kyky ennustaa vastemuuttujan oikea arvo on 0,90 (vertailuna käytettävä vastemuuttujan yleisemmän arvon todennäköisyys – se, että vastaaja on valinnut alkuperäistä kontekstia vastaavan adjektiivin – on 0,62). Kaikkiaan diagnostiset tulokset osoittavat mallin soveltuvan hyvin tarkastelunalaisen ilmiön kuvaamiseen, mikä osaltaan myös vahvistaa sen avulla saatavien tulosten luotettavuutta.

Taulukko 3 kuvaa mallin kunkin muuttujan ja vuorovaikutuksen osalta sitä, onko niiden vaikutus[5] kokonaisuuteen tilastollisesti merkitsevä. Mallin yhdeksästä kiinteästä muuttujasta neljän vaikutus on tilastollisesti merkitsevä. Lisäksi viiden muuttujan vuorovaikutus vastaajan ensikielisyyden kanssa on tilastollisesti merkitsevä, mikä tarkoittaa kyseisen muuttujan vaikutuksen olevan erilainen ensikielenään suomea käyttävien vastaajien ja suomea toisena kielenä käyttävien vastaajien kohdalla. *suuri*-pesueen kohdalla esimerkiksi vastaajan ensikielisyys ($p = 0{,}021$) ja vastaajan vuodessa lukemien kirjojen määrä vuorovaikutuksessa ensikielisyyden kanssa ($p = 0{,}026$) osoittautuivat tilastollisesti merkitseviksi muuttujiksi. Tarkastelemme osioissa 4.2–4.4 tarkemmin yksittäisten muuttujien osuutta kokonaisuudessa keskittyen erityisesti niihin muuttujiin, joiden vaikutus on tilastollisesti merkitsevä.

***tärkeä*-pesue.** Toisessa mallissa käytimme vastaavia muuttujia mallintamaan sitä, valitseeko vastaaja *tärkeä*-pesueesta saman adjektiivin, jota on käytetty kaunokirjallisuudessa tarkastelunalaisessa kontekstissa. Tämä malli ei yllä aivan *suuri*-pesuetta kuvaavan mallin lukemiin, mutta sen voidaan silti katsoa kykenevään aineistossa havaitun vaihtelun kohtalaisen hyvin: marginaalinen R^2-arvo on 0,16 ja satunnaismuuttujat sisältävä ehdollinen R^2-arvo on 0,57. C-arvo on 0,819, mikä ylittää hyvän sovellettavuuden rajana pidetyn arvon 0,8. Se kykenee ennustamaan vastemuuttujan arvon todennäköisyydellä 0,79, kun vertailuna käytettävä vastemuuttujan yleisemmän arvon todennäköisyys on 0,62. Taulukko 3 kuvaa mallin kunkin muuttujan ja vuorovaikutuksen osalta sitä, onko niiden vaikutus kokonaisuuteen tilastollisesti merkitsevä. Mallissa neljän kiinteän muuttujan vaikutus on tilastollisesti merkitsevä, minkä lisäksi neljän muuttujan vuorovaikutus vastaajan ensikielisyyden kanssa on tilastollisesti merkitsevä. Myös tämän pesueen osalta tarkastelemme yksittäisiä muuttujia tarkemmin osioissa 4.2–4.4.

[5] Käytämme regressioanalyysien vakiintuneen tavan mukaan termiä *vaikutus* (engl. *effect*) kuvaamaan muuttujien välistä suhdetta. Termi ei ota kantaa muuttujien mahdollisiin kausaalisuhteisiin.

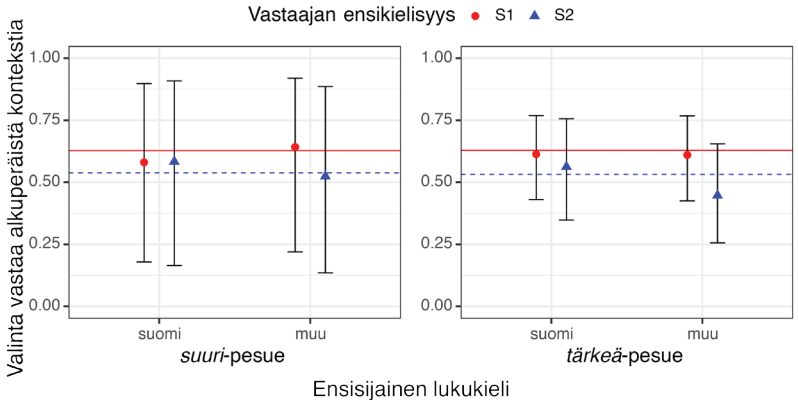
Taulukko 3. Uskottavuusosamäärä-testien (LRT) tulokset tutkittujen pesueiden osalta. Kynnysarvoa 0,05 alemmat arvot (merkitty asteriskilla *) tulkitaan tilastollisesti merkitseviksi alkuperäisen kontekstin adjektiivivalinnan kannalta. Df = vapausaste[6]

Muuttuja	Df	*suuri*-pesue		*tärkeä*-pesue	
		LRT	P-arvo	LRT	P-arvo
Vastaajan ensikielisyys	1	5,337	0,021*	15,016	0,0001*
Vastaajan ensisijainen lukukieli	1	3,457	0,063	1,039	0,308
Vuorovaikutus: Vastaajan ensikielisyys		4,076	0,043*	4,649	0,031
Vastaajan vuodessa lukemien kirjojen määrä	numeerinen	0,035	0,851	1,672	0,196
Vuorovaikutus: Vastaajan ensikielisyys		4,962	0,026*	0,908	0,341
Vastaajan lukemisen frekvenssi	numeerinen	0,148	0,700	0,097	0,756
Vuorovaikutus: Vastaajan ensikielisyys		5,404	0,020*	0,715	0,398
Valitun adjektiivin ja sen pääsanan välinen assosiaatio	numeerinen	1093,542	< 0,0001*	415,055	< 0,0001*
Vuorovaikutus: Vastaajan ensikielisyys		0,110	0,740	5,154	0,023*
Valitun adjektiivin ja sen pääsanan tekstilajipreferenssi	2	405,977	< 0,0001*	99,627	< 0,0001*
Vuorovaikutus: Vastaajan ensikielisyys		3,864	0,145	8,150	0,017*
Pääsanan esiintymisfrekvenssi kaunokirjallisuudessa	numeerinen	78,539	< 0,0001*	0,522	0,470
Vuorovaikutus: Vastaajan ensikielisyys		30,090	< 0,0001*	1,063	0,302
Alkuperäisessä kontekstissa käytetty adjektiivi	2	5,463	0,065	12,866	0,002*
Vuorovaikutus: Vastaajan ensikielisyys		55,164	< 0,0001*	3,242	0,198
Sijamuoto (satunnaismuuttuja)		Estimaatin keski-hajonta S1: 0,744 S2: 1,076	0,0001*	Estimaatin keski-hajonta S1: 0,389 S2: 0,555	0,915
Korrelaatio ensikielisyyden ja sijamerkinnän välillä		0,10	< 0,0001*	0,94	0,812
Kysymys (satunnaismuuttuja)		Estimaatin keski-hajonta: 3,637	< 0,0001*	Estimaatin keski-hajonta: 1,722	< 0,0001*
Vastaaja (satunnaismuuttuja)		Estimaatin keski-hajonta: 0,094	0,809	Estimaatin keski-hajonta: 0,216	0,013*

[6] Vapausaste tarkoittaa niiden muuttujien määrää, jotka voivat vaihdella. Esimerkiksi vastaajan ensikielisyys -muuttujan Df = 1, koska vaihtoehtoja on kaksi (S1 tai S2) ja alkuperäisessä kontekstissa käytetyn adjektiivin Df = 2, koska vaihtoehtoja on 3 (*suuri/iso/kookas* tai *tärkeä/keskeinen/merkittävä).*

Tarkastelemme seuraavaksi yksittäisten muuttujien vaikutusta siihen, onko vastaaja valinnut saman adjektiivin kuin alkuperäisen tekstin kirjoittaja. Esittelemme rinnakkain tulokset *suuri*- ja *tärkeä*-pesueista. Käsittelemme ensin lukemistottumuksiin liittyvät muuttujat ja sen jälkeen käyttöä, merkitystä ja muotoa kuvaavat konstruktiokohtaiset muuttujat. Kuviot 2–9 kuvaavat tarkasteltujen muuttujien vaikutusta lähisynoyymisen adjektiivin valintaan. Selittävien muuttujien arvot ovat kuvioissa vaaka-akselilla ja vastemuuttujan (eli vastaajan adjektiivivalinnan suhteen alkuperäisen kontekstin adjektiiviin) arvot pystyakselilla. Pystyakselin arvot vaihtelevat välillä 0–1 niin, että suuremmat arvot tarkoittavat suurempaa todennäköisyyttä valita sama adjektiivi, jota on käytetty alkuperäisessä kaunokirjallisessa kontekstissa. Merkintä S1 tarkoittaa kuvioissa suomea ensikielenään puhuvia ja S2 suomea toisena kielenä puhuvia vastaajia.

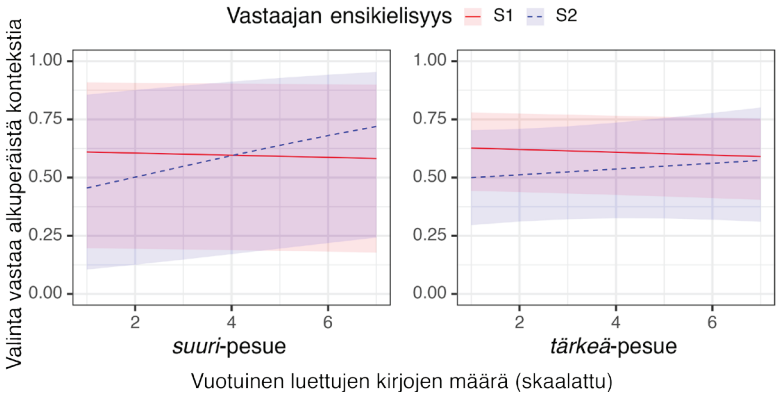
4.2 Lukemistottumukset



Kuvio 2. Vastaajan ensisijaisen lukukielen vaikutus adjektiivivalintaan[7]

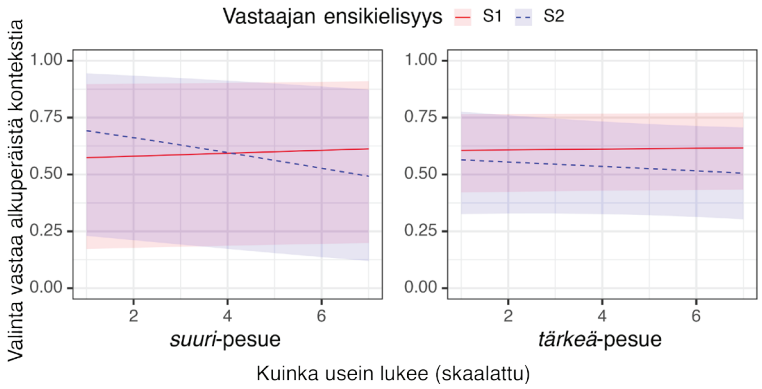
Vastaajan ensisijainen lukukieli näyttää kiinnostavan tendenssin monikielisen lukemisen hyödyllisyydestä kuviossa 2. Molempien pesueiden kohdalla ne S2-puhujat, jotka raportoivat lukevansa tällä hetkellä ensisijaisesti suomeksi, ovat valinneet adjektiiveja samansuuntaisemmin alkuperäisen kaunokirjallisen kontekstin kanssa. S1-puhujien osalta päinvastoin muulla kielellä kuin suomeksi lukeminen näyttää tuottavan kontekstinmukaisempia valintoja. Ero on *suuri*-pesueen osalta lähellä tilastollisesti merkitsevää. Tulos voi viitata siihen, että vieraan kielen osaaminen edistyneellä tasolla näkyy oman ensikielen hienosyisemmässä käytössä, kuten Kecskes ja Papp (2000) esittävät. Vieraalla kielellä lukeminen voisi siis parantaa herkkyyttä oman ensikielen käytössä. S1- ja S2-puhujien tendenssit ensisijaisen lukukielen vaikutuksesta kertovat hieman eri asiasta, koska mallissa käytetyllä muuttujalla ei saada S2-puhujien osalta selville, tapahtuuko muulla kielellä kuin suomeksi lukeminen vastaajien ensikielellä vai jollakin muulla heidän osaamallaan kielellä, kuten englanniksi. Erilaisten lukukieliasetelmien tarkasteluun tarvittaisiin tätä tutkimusta laajempi aineisto.

[7] Punainen pallo ja sininen kolmio kuvaavat kyseisen muuttujan estimaattia eli keskimääräistä vaikutusta monimuuttujamallissa. Pystysuuntainen virhepalkki kuvaa 95 % luottamusväliä eli sitä väliä, jolla arvojen oletetaan korkeintaan vaihtelevan eri aineistoa käytettäessä. Vaakasuuntaiset viivat kuvaavat alkuperäisen kontekstin mukaisten vastausten mallista riippumatonta keskiarvoa. Punainen yhtenäinen vaakasuuntainen viiva kuvaa S1-vastaajien keskiarvoa ja sininen katkoviiva S2-vastaajien keskiarvoa. Sama merkintätapa on käytössä myös kuvioissa 7, 8 ja 9.



Kuvio 3. Vastaajan vuodessa lukemien kirjojen määrän vaikutus adjektiivivalintaan

Myös vuodessa luettujen kirjojen määrä kuviossa 3 näyttää samansuuntaisen tendenssin molempien pesueiden osalta, vaikka *tärkeä*-pesueessa ero ei olekaan tilastollisesti merkitsevä. S2-puhujien osalta mitä enemmän vastaaja lukee kirjoja, sitä yhdenmukaisempia hänen adjektiivivalintansa ovat alkuperäisen kontekstin kanssa. Ensikielisten aikuisten on havaittu voivan oppia kaunokirjallisuutta lukemalla keskimäärin enemmän sanoja kuin kielenoppijoiden (Reynolds 2015: 121). S1-puhujilla kaunokirjallisuuden lukemisen määrä ei tässä tutkimuksessa kuitenkaan näytä vaikuttavan adjektiivivalintojen kontekstinmukaisuuteen. S1-puhujilla on aikuisiällä suomea oppineita pidempi altistumisaika suomenkieliselle kaunokirjallisuudelle. Täten heidän aikuisiällä lukemiensa kirjojen määrä ei välttämättä enää vaikuta tässä kyselyssä tarkasteltuun sanaston osaan, joka kuuluu sekä tarkasteltujen synonyymisten adjektiivien että niiden kanssa esiintyvien pääsanojen osalta melko arkiseen perussanastoon.

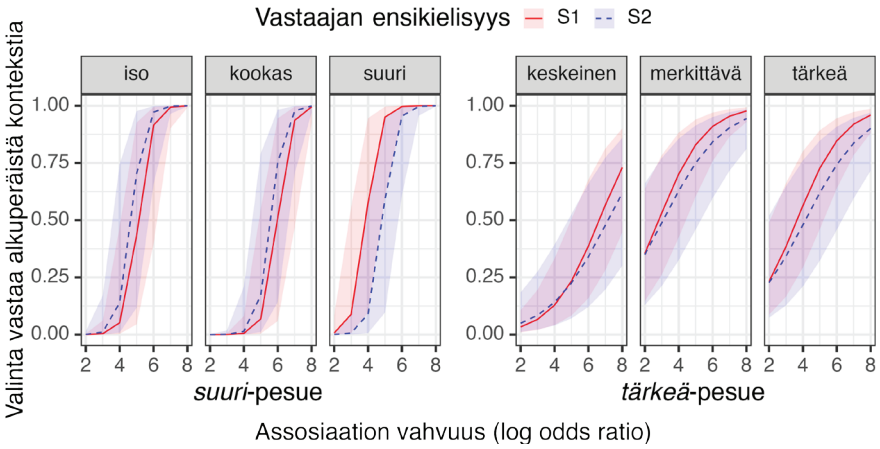


Kuvio 4. Vastaajan lukemiseen käyttämän ajan vaikutus adjektiivivalintaan

Kuvio 4:n kuvaama tendenssi on jossain määrin epäodotuksenmukainen erityisesti suhteessa edelliseen kuvioon 3. Siinä missä vuodessa luettujen kirjojen suurempi määrä nostaa S2-puhujilla adjektiivivalintojen yhdenmukaisuutta alkuperäisen kontekstin kanssa, yksittäisten lukutuokioiden suurempi määrä näyttää puolestaan laskevan sitä. Luettujen kirjojen ja lukutuokioiden kasvavat määrät ovat aiemmissa tutkimuksissa (esim. Pietilä & Merikivi 2014) molemmat korreloineet positiivisesti sanaston laajuuden kanssa nuorilla vieraan kielen oppijoilla. Vastaajat, jotka raportoivat kyselyssämme lukevansa kaunokirjallisuutta ”joka päivä” saavat luettua hyvin eri määrän kirjoja: monet heistä

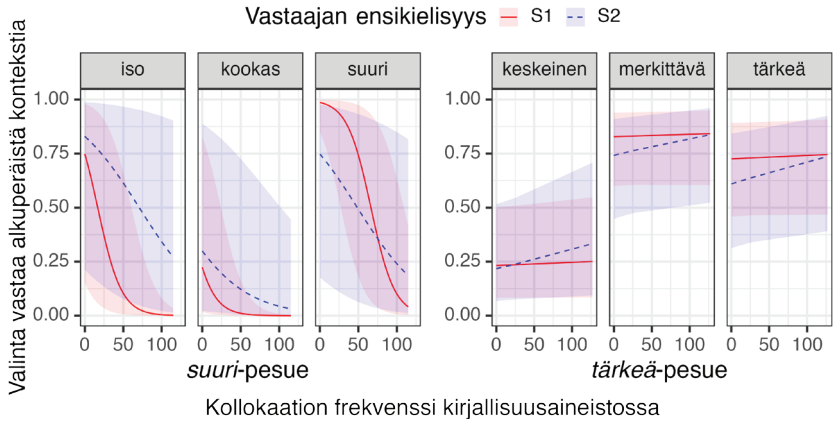
lukevat yli 30 kirjaa vuodessa, mutta jotkut vain 2–5. Synonyymisen sanaston oppimisen näkökulmasta tärkeää näyttäisi siis olevan nimenomaan se, kuinka paljon kaunokirjallista kielen mallia eli syötöstä saa, eli kuinka monta kokonaisteosta lukee.

4.3 Tyypillinen käyttökonteksti



Kuvio 5. Adjektiivin ja sen pääsanan eli kollokaation assosiaation vahvuus S1- ja S2-vastaajien adjektiivivalintojen erottimena

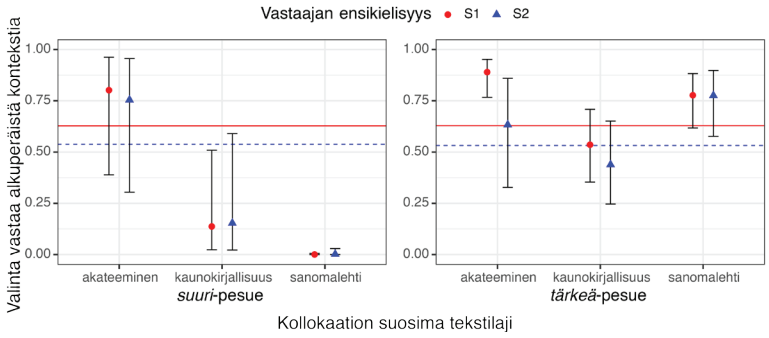
Kuviossa 5 näkyvä tendenssi vahvistaa aiemmasta tutkimuksesta (esim. Gries & Wulff 2005) saatua tietoa siitä, että ensikieliset puhujat ja edistyneet kielenoppijat hahmottavat konstruktioita hyvin samantapaisesti: mitä voimakkaammin alkuperäisessä kontekstissa käytetty adjektiivi ja sen kanssa esiintyvä pääsana ovat assosioituneet toisiinsa kaunokirjallisuuskontekstissa ylipäätään, sitä todennäköisempää on, että sekä S1- että S2-vastaajat valitsevat alkuperäisen kontekstin mukaisia adjektiiveja. Kuviosta näkyy niin ikään *suuri*-pesueen *suuri*-variantin kohdalla ja *tärkeä*-pesueen kaikkien varianttien osalta se, että S1-puhujat ovat tälle assosiaatiovihjeelle hieman edistyneitä S2-oppijoita herkempiä.



Kuvio 6. Kollokaation yleisyys S1- ja S2-vastaajien adjektiivivalintojen erottimena

tärkeä-pesueen kohdalla kollokaation absoluuttisen määrän yleisyys vaikuttaa odotuksenmukaisesti kuviossa 6: mitä yleisempi alkuperäisessä kontekstissa käytetty kollokaatio on kaunokirjallisuuskontekstissa ylipäätään, sitä samankaltaisempia valintoja sekä

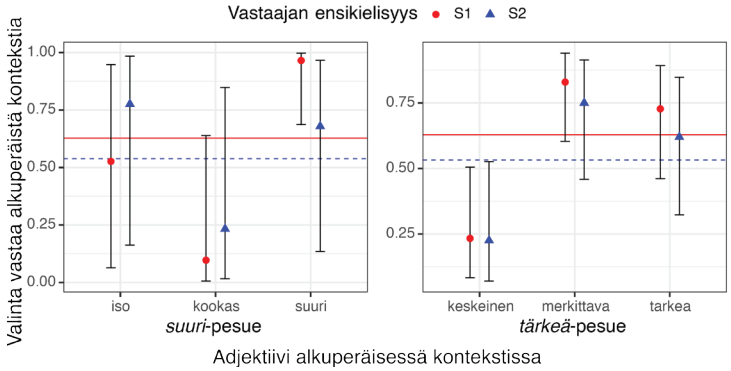
S1- että S2-vastaajat tekevät alkuperäisen kontekstin kanssa. *suuri*-pesueen kohdalla tilanne näyttää ensi alkuun päinvastaiselta, ja adjektiivi-substantiivikombinaation absoluuttinen frekvenssi näyttää aiheuttavan molemmilla vastaajaryhmillä erisuuntaisia vastauksia kuin alkuperäisessä kontekstissa. Ero selittyy kuitenkin sillä, että kyselyyn valikoituneet *suuri*-pesueen pääsanat ovat ylipäätään hyvin yleisiä ja monikäyttöisiä, eli kollokaatioilla ei ole suuresta frekvenssistään huolimatta kovinkaan vahvaa assosiaatiosuhdetta. Todella frekventtejä, yli sata kertaa esiintyviä kollokaatioita on molemmilla pesueilla vain yksi, joten kuvio on melko hypoteettinen sen suhteen, millainen vaikutus todellisella frekvenssillä oikeasti on valintaan. Tämä vahvistaa aiempaa käsitystä (esim. Gries & Deshors 2014: 111) siitä, että pelkkä esiintymien frekvenssi ei riitä tilastolliseen mallintamiseen vaan ilmiön ymmärtämiseksi tarvitaan esimerkiksi assosiaatiolaskuja (ks. kuvio 5 edellä).



Kuvio 7. Kollokaation tekstilajipreferenssi S1- ja S2-vastaajien adjektiivivalintojen erottimena

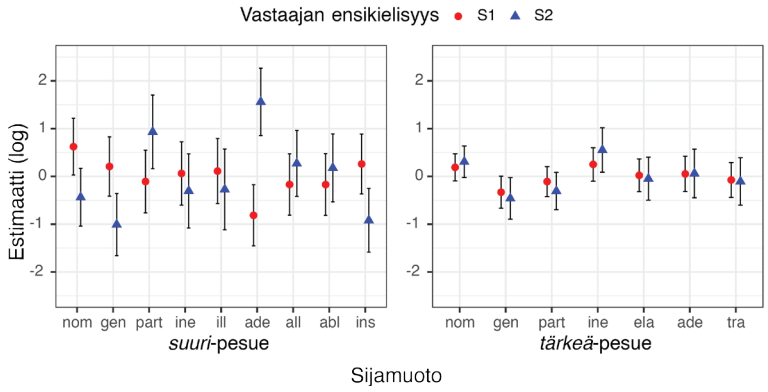
Kollokaation preferoimaa tekstilajia tarkasteltiin neljän kirjoitetun tekstilajin avulla: kaunokirjallisuus, akateemiset tekstit, sanomalehtitekstit ja internetkeskustelut. Koska yksikään kollokaatio ei preferoinut voimakkaimmin internetkeskusteluja, se ei ole näkyvissä kuviossa 7. Kuviossa näkyy kiinnostavasti, että tekstilajipreferenssi on ohjannut S1- ja S2-vastaajien valintoja hyvin samansuuntaisesti molempien adjektiivipesueiden osalta. Pesueiden väliset erot kuvaavat mahdollisesti jossain määrin tekstilajien välisiä eroja sen suhteen, mitä eri tekstilajeissa puhutaan: koosta puhuminen saattaa olla tyypillistä nimenomaan akateemisille teksteille, kun taas merkityksestä puhutaan tasaisemmin kaikissa tekstilajeissa. Aiemmassa tutkimuksessa (esim. Ivaska 2015) edistyneiden suomenoppijoiden kirjoittamien akateemisten tekstien rakenne-eroja ensikielisten teksteihin on perusteltu nimenomaan eriasteisella akateemisen kirjoittamisen kielenkäytön tuntemuksella. Tämän tutkimuksen aineiston perusteella tekstilajilla ei kuitenkaan näytä olevan kovin suurta eroa korkeakoulutetuilla vastaajilla, mikä johtunee siitä, että aineisto kuvaa ennemmin synonyymisen sanaston reseptiivistä kuin produktiivista hallintaa.

4.4 Alkuperäisen sanan ominaisuudet



Kuvio 8. Yleiskatsaus adjektiivivalintojen vastaavuudesta alkuperäiseen kontekstiin

Yleiskatsaus vastaajien adjektiivivalinnoista suhteessa alkuperäisen kaunokirjallisen tekstin kirjoittajan tekemään valintaan kuviossa 8 osoittaa selvän eron *suuri*- ja *tärkeä*-pesueiden välillä. *tärkeä*-pesueessa vähiten frekventti *keskeinen* noudattaa molemmilla vastaajaryhmillä samaa linjaa, mutta kiinnostavasti yhdenmukaisimmin alkuperäisen kontekstin kanssa on käytetty varianttia *merkittävä*, ei siis frekventeintä *tärkeää*. *suuri*-pesueen osalta muuttuja ei ollut tilastollisesti merkitsevä, mutta sen eroa toiseen pesueeseen voi silti pitää kiinnostavana. S1-vastaajat ovat käyttäneet pesueen frekventeintä varianttia eli *suurta* kontekstinmukaisimmin. S2-vastaajilla *iso* nousee useimmin kontekstinmukaisesti valituksi, mikä saattaisi viitata siihen, että *iso* mielletään pesueen monikäyttöisimmäksi variantiksi. Molemmissa vastaajaryhmissä epäfrekventeintä *kookasta* on valittu vähiten samankaltaisesti alkuperäisen kontekstin kanssa. Osasyynä *kookkaan* ja *keskeisen* vähemmän kontekstinmukaiseen valintaan lienee se aineistonkeruuseen liittyvä seikka, että pesueiden yleisemmistä adjektiiveista poiketen osa näiden sanojen yhteydessä esiintyvistä pääsanoista ei ollut yhtä vahvasti kyseiseen adjektiiviin assosioituneita. Selvää on, että eri varianttien kokonaisfrekvenssin vaikutus valintaan ei ole yksiselittei-nen eivätkä S2-vastaajat eroa S1-vastaajista sen suhteen yhdenmukaisesti, vaan vaikutus näyttää vaihtelevan synonyymipesueittain.



Kuvio 9. Konstruktion sijamuoto S1- ja S2-vastaajien adjektiivivalintojen erottimena

Kuvio 9 visualisoi sijamuodon vaikutusta siihen, ovatko vastaajat valinneet alkuperäiskontekstia vastaavan adjektiivin. Koska sijamuoto on käytetyssä mallissa luonteeltaan satunnaismuuttuja, ei sen vaikutusta tule vertailla suoraan edellä kuvattujen

muuttujien kanssa. Muuttujan vaikutusta ja kuviota 9 voi tarkastella siitä näkökulmasta, miten kukin muuttujan arvo suhteellisesti vaikuttaa adjektiivin valintaan: nollan ylittävät arvot kuvaavat sijamuotoja, joissa esiintyvät adjektiivit valittiin alkuperäiskontekstin mukaisesti keskimääräistä yleisemmin ja nollan alittavat arvot puolestaan sijamuotoja, joissa esiintyvät adjektiivit poikkesivat keskimääräistä yleisemmin alkuperäiskontekstista. Huomionarvoista on erityisesti se, että *suuri*-pesueen kohdalla sijamuotojen välistä vaihtelua on paljon enemmän kuin *tärkeä*-pesueessa. Lisäksi on mielenkiintoista, että *suuri*-pesueessa S1-ja S2-vastaajat poikkeavat monissa kohdin toisistaan selvästi, kun taas *tärkeä*-pesueessa tällaista eroa ei ole. Tämä tukee ja tarkentaa taulukon 3 perusteella tehtyä päätelmää, jonka mukaan *suuri*-pesueessa S1- ja S2-vastaajien välillä on tilastollisesti merkitsevä ero mutta *tärkeä*-pesueessa ei. Vastaajaryhmien välinen ero on selvä yhtäältä suuri-pesueen nominatiivissa ja instruktiivissa esiintyneiden tapausten kohdalla, joissa S1-vastaajat ovat valinneet selvästi S2-vastaajia useammin alkuperäisen kontekstin mukaisen adjektiivin. Toisaalta partitiivissa ja erityisesti adessiivissa esiintyneiden adjektiivien kohdalla S2-vastaajat ovat valinneet selvästi S1-vastaajia useammin alkuperäisen kontekstin mukaisen adjektiivin.

5 Johtopäätökset

Olemme tässä artikkelissa käsitelleet kokeellisen kyselyaineiston pohjalta tehdyn tilastollisen regressiomallin avulla sitä, minkä lähisynonyymisen adjektiivin ensikieliset suomenpuhujat ja edistyneet suomenoppijat valitsevat kaunokirjalliseen tekstikontekstiin substantiivin määriteasemassa. Tutkittavina synonyymipesueina olivat *suuri/iso/kookas* ja *tärkeä/merkittävä/keskeinen*, jotka ovat kaunokirjallisuuskorpuksen yleisimmät adjektiivipesueet. Tarkastelimme valintaa suhteessa alkuperäisen kirjoittajan tekemään valintaan, vastaajien kaunokirjallisuuden lukemistottumuksiin sekä adjektiivimääritekonstruktion ominaisuuksin ja tekstilajikäyttöön.

Vastauksena ensimmäiseen tutkimuskysymykseen huomattiin sekä S1- että S2-vastaajien adjektiivivalintojen eroavan jossain määrin kaunokirjallisten tekstien alkuperäisten kirjoittajien valinnoista. Lisäksi kielitaustalla on tilastollisesti merkitsevä ero valintoihin. Useat lukemistottumuksia ja adjektiivien konstruktiokäyttäytymistä mittaavat muuttujat osoittautuivat tilastollisesti merkitseviksi. Valintaan pakottavan kyselytestin avulla voidaan tarkastella lähisynonyymisten adjektiivien reseptiivistä hallintaa, mikä tuottaa osin erinäköisiä tuloksia kuin aiemmat L2-puhujien synonyymisten ilmausten produktiiviseen käyttöön keskittyneet korpustutkimukset.

Toisena tutkimuskysymyksenä kysyimme, millainen yhteys kaunokirjallisuuden lukeneisuudella on S1- ja S2-puhujien adjektiivivalintoihin. Lukemistottumusten osalta voidaan todeta, että kaunokirjallisuuden lukeminen vaikuttaa samansuuntaisesti tarkasteltavaan ilmiöön eli adjektiivin valintaan määritekonstruktiossa. Edistyneillä S2-puhujilla valinnat ovat sitä kontekstinmukaisempia, mitä enemmän he lukevat kohdekielellä eli suomeksi. S1-puhujilla vastaavaa yhteyttä ei mallinnuksen perusteella voi osoittaa. Sen sijaan ensikielisillä aikuisilla vieraalla kielellä lukeminen saattaa tehdä herkemmäksi kaunokirjallisuuskontekstin mukaisille valinnoille.

Myös kolmannessa tutkimuskysymyksessä esittämämme tekstilajipreferenssiä ja kollokaation assosiaatiota kuvaavat muuttujat vaikuttivat samansuuntaisesti molemmissa pesueissa. Alkuperäisen sanan ominaisuudet, eli kaunokirjallisessa kontekstissa alun perin käytetty adjektiivi ja konstruktion sijamuoto, puolestaan vaikuttivat eri tavoin

tarkastelujen pesueiden kohdalla. Tämä on erityisen tärkeä havainto jatkotutkimuksen kannalta, kun mietitään, kuinka synonyymisten konstruktioiden rakennetta tarkastelevien tutkimusten tuloksia voidaan yleistää koskemaan synonymiaa laajemmin.

Tulokset osoittavat, että tilastollisena menetelmänä käytetty monimuuttujainen logistinen regressio-sekamalli sopii tutkimuksessa tarkasteltavan ilmiön selittämiseen. Malli toimii molemmissa pesueissa hyvin, mutta kiinnostavasti se selittää selvästi paremmin *suuri*-pesueen adjektiivivalintojen vaihtelua. Ero selittyy ennen kaikkea sillä, että aiemman tutkimuksen perusteella identifioiduilla, lukutottumuksiin ja konstruktiokohtaisiin ilmiöihin kohdistuneilla kiinteillä muuttujilla on selvästi pienempi rooli *tärkeä*-pesueen valintoja mallinnettaessa. Kysymys- ja vastaajakohtaiset satunnaismuuttujat sen sijaan selittävät karkeasti ottaen yhtä suuren osan vaihtelusta kummassakin pesueessa, eli kyselyn rakenne tuskin selittää pesueiden välistä eroa. Synonymiaa käsittelevissä korpustutkimuksissa on usein tarkastelussa vain yksi synonyymipari tai -pesue. Tämän tutkimuksen pohjalta voi esittää, että synonyymiutta käsittelevässä tutkimuksessa tulisi huomioida yksittäisten lähisynonyymisten pesueiden sijaan useita eri synonyymipesueita, jotta sanakohtaisen vaihtelun sijaan olisi ylipäätään mahdollista päästä tarkastelemaan synonymiaa laajempana ilmiönä.

Tutkimuksessa käytettyyn kyselyasetelmaan liittyy joitakin rajoituksia, jotka on hyvä ottaa huomioon tulosten tulkinnassa. Vastausväsymyksen ja arvaamiseen turvautumisen minimoimiseksi adjektiivivalintaan pakottavat monivalintakysymykset ja niiden vastausvaihtoehdot on satunnaistettu, mutta kyselyalustasta johtuvista teknisistä syistä pesueet esitettiin kuitenkin kaikille vastaajille samassa järjestyksessä: *suuri*-pesue ennen *tärkeä*-pesuetta. Vastausväsymys on siis saattanut olla suurempi jälkimmäisen pesueen kohdalla, mikä voi heijastua huonompina arvoina pesueen mallintamisen toimivuudessa. Kysymys- ja vastaajakohtaiset satunnaismuuttujat kattavat kuitenkin karkeasti ottaen yhtä paljon vaihtelusta pesueesta riippumatta, joten kyselyn rakenne tuskin on tässä kovinkaan merkittävässä osassa. S1- ja S2-vastaajaryhmät olivat hyvin erikokoisia (N = 234 vs. N = 28), joten S1-vastaajaryhmää voi pitää edustavampana kuin S2-ryhmää. Monimuuttujamalli ottaa ryhmien kokoeron kuitenkin lähtökohtaisesti huomioon, joten erikokoisten vastaajaryhmien ei pitäisi merkittävästi vaikuttaa tulosten luotettavuuteen. Yhteen vastausvaihtoehtoon pakottavassa kyselyssä ei lisäksi saada huomioitua synonyymisten ilmausten välimaastoa, siis tapauksia, joissa ensikielinen vastaaja voisi todennäköisesti käyttää jompaakumpaa tai mitä tahansa annetuista vaihtoehdoista (Gries & Deshors 2014; Deshors & Gries 2016; Kekki & Ivaska 2022). Välimaasto saattaa selittää niitä kyselyn kohtia, joiden valinnoissa on erityisen paljon vaihtelua vastaajien välillä.

Lähisynonyymisten ilmausten avulla kielenkäyttäjä voi esimerkiksi osoittaa tuntevansa tekstilajien välisiä eroja ja tuottaa hienosyisiä sävyeroja. Tutkimuksen tulokset kannustavat käyttämään kaunokirjallisuutta aikuisten kielenopetuksessa, jotta näiden sävyerojen tunnistaminen mahdollistuisi. Tutkimuksessa käytetty kysely (ks. liite A) voi lisäksi toimia opetusmateriaalina sellaisenaan ja auttaa opettajaa tarkastelemaan yhdessä suomenoppijoiden kanssa, mitkä tekijät vaikuttavat adjektiivin valintaan kaunokirjallisuuskontekstissa.

Korpuslähteet

InterCorp - Finnish, version 12: Fárová, Lenka & Vavřín, Martin. 2019. Institute of the Czech National Corpus, Faculty of Arts, Charles University.
kaannossuomi-korp: Mauranen, Anna. 1999. Käännössuomen korpus, Korp [tekstikorpus]. Kielipankki.
suomi24-2001-2020-korp: City Digital Group. 2021. Suomi24 virkkeet -korpus 2001-2020, Korp-versio [tekstikorpus]. Kielipankki.
e-thesis-fi: Helsingin yliopisto. 1999. Helsingin yliopiston suomenkielisen E-thesiksen Korp-versio [tekstikorpus]. Kielipankki.
lehdet90ff-v2: Helsingin yliopisto. 2017. 1990- ja 2000-luvun suomalaisia aikakaus- ja sanomalehtiä -korpus, versio 2 [tekstikorpus]. Kielipankki.

Lähteet

Alderson, Charles & Haapakangas, Eeva-Leena & Huhta, Ari & Nieminen, Lea & Ullakonoja, Riikka. 2015. *The diagnosis of reading in a second or foreign language*. New York: Routledge, Taylor & Francis Group.
Arppe, Antti. 2008. *Univariate, bivariate and multivariate methods in corpus-based lexicography: A study of synonymy*. Helsingin yliopisto. (Väitöskirja).
Arppe, Antti & Järvikivi, Juhani. 2007. Every method counts: Combining corpus-based and experimental evidence in the study of synonymy. *Corpus Linguistics and Linguistic Theory* 3(2). 131–159. https://doi.org/10.1515/CLLT.2007.009
Biber, Douglas & Conrad, Susan. 2005. Register Variation: A Corpus Approach. Teoksessa Schiffrin, Deborah & Tannen, Deborah & Hamilton, Heidi (toim.), *The Handbook of Discourse Analysis*, 175–196. Wiley. https://doi.org/10.1002/9780470753460.ch10
Biber, Douglas & Conrad, Susan & Reppen, Randi. 1998. *Corpus Linguistics: Investigating Language Structure and Use.* Cambridge University Press. https://doi.org/10.1017/CBO9780511804489
Bono, Susanna. 2023. 14-vuotias Alma Hurme lukee vain englanniksi – ”Suomeksi samat kirjat saattavat kuulostaa vähän hassuilta”. YLE. https://yle.fi/a/74-20043198 (Viitattu 16.8.2023.)
Church, Kenneth & Hanks, Patrick. 1990. Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics* 16(1). 22–29.
Conlexis. 2013. http://wiki.osao.fi/conlexis/ (Viitattu 1.8.2023.)
Cruse, David. 1986. *Lexical Semantics*. Cambridge University Press.
Danglli, Leonard & Abazaj, Griselda. 2014. Lexical Cohesion, Word Choice and Synonymy in Academic Writing. *Mediterranean Journal of Social Sciences* 5(14). 628–632. https://doi.org/10.5901/mjss.2014.v5n14p628
De Jonge, Bob. 1993. The existence of synonyms in a language: Two forms but one, or rather two, meanings? *Linguistics* 31(3). 521–538. https://doi.org/10.1515/ling.1993.31.3.521
Deshors, Sandra. 2016. Inside phrasal verb constructions: A co-varying collexeme analysis of verb-particle combinations in EFL and their semantic associations. *International Journal of Learner Corpus Research* 2(1). 1–30. https://doi.org/10.1075/ijlcr.2.1.01des
Deshors, Sandra, & Gries, Stefan. 2016. Profiling verb complementation constructions across New Englishes: A two-step random forests analysis of *ing* vs. *to* complements. *International Journal of Corpus Linguistics* 21(2). 192–218. https://doi.org/10.1075/ijcl.21.2.03des
Deshors, Sandra & Götz, Sandra & Laporte, Samantha. 2018. Linguistic innovations in EFL and ESL: Rethinking the linguistic creativity of non-native English speakers. Teoksessa Deshors, Sandra & Götz, Sandra & Laporte, Samantha (toim.), *Benjamins Current Topics*, 1–20. John Benjamins Publishing Company. https://doi.org/10.1075/bct.98.01des
Deshors, Sandra & Gries, Stefan. 2014. A case for the multifactorial assessment of learner language: The uses of *may* and *can* in French-English interlanguage. Teoksessa Glynn, Dylan & Robinson, Justyna (toim.), *Human Cognitive Processing*, 179–204. John Benjamins Publishing Company. https://doi.org/10.1075/hcp.43.07des
Divjak, Dagmar & Gries, Stefan. 2006. Ways of trying in Russian: Clustering behavioral profiles. *Corpus Linguistics and Linguistic Theory* 2(1). 23–60. https://doi.org/10.1515/CLLT.2006.002
Edmonds, Philip & Hirst, Graeme. 2002. Near-Synonymy and Lexical Choice. *Computational Linguistics* 28(2). 105–144. https://doi.org/10.1162/089120102760173625

Ellis, Nick. 2008. Usage-based and form-focused language acquisition: The associative learning of constructions, learned attention, and the limited L2 endstate. Teoksessa Robinson, Peter & Ellis, Nick (toim.), *Handbook of Cognitive Linguistics and Second Language Acquisition*, 382–415. Routledge. https://doi.org/10.4324/9780203938560-24

Firth, John. 1957. A Synopsis of Linguistic Theory, 1930–55. *Studies in Linguistic Analysis, Special Volume of the Philological Society*. 1–31.

Goldberg, Adele. 2003. Constructions: A new theoretical approach to language. *Trends in Cognitive Sciences* 7(5). 219–224. https://doi.org/10.1016/S1364-6613(03)00080-9

Granger, Sylviane. 2004. Computer Learner Corpus Research: Current Status and Future Prospects. Teoksessa Connor, U. & Upton, T. (toim.), *Applied Corpus Linguistics: A Multidimensional Perspective*, 123–145. https://doi.org/10.1163/9789004333772

Granger, Sylviane. 2015. Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research* 1(1). 7–24. https://doi.org/10.1075/ijlcr.1.1.01gra

Gries, Stefan & Otani, Naoki. 2009. Behavioral profiles: A corpus-based perspective on synonymy and antonymy. *ICAME Journal* 2009(34). 121–150.

Gries, Stefan. 2021. *Statistics for Linguistics with R* (3rd edition). De Gruyter Mouton.

Gries, Stefan. 2022. What do (some of) our association measures measure (most)? Association? *Journal of Second Language Studies* 5(1). 1–33. https://doi.org/10.1075/jsls.21028.gri

Gries, Stefan & Deshors, Sandra. 2014. Using regressions to explore deviations between corpus data and a standard/target: Two suggestions. *Corpora* 9(1). 109–136. https://doi.org/10.3366/cor.2014.0053

Gries, Stefan & Wulff, Stefanie. 2005. Do foreign language learners also have constructions? *Annual Review of Cognitive Linguistics* 3. 182–200. https://doi.org/10.1075/arcl.3.10gri

Hakulinen, Auli & Vilkuna, Maria & Korhonen, Riitta & Koivisto, Vesa & Heinonen, Tarja-Riitta & Alho, Irja (toim.). 2004. *Ison suomen kieliopin verkkoversio*. Kotimaisten kielten tutkimuskeskus. https://kaino.kotus.fi/visk/etusivu.php (Viitattu 1.6.2023)

Hasselgren, Angela. 1994. Lexical teddy bears and advanced learners: A study into the ways Norwegian students cope with English vocabulary. *International Journal of Applied Linguistics* 4(2). 237–258. https://doi.org/10.1111/j.1473-4192.1994.tb00065.x

Ivaska, Ilmari. 2015. *Edistyneen oppijansuomen konstruktiopiirteitä korpusvetoisesti: Avainrakenneanalyysi.* (Annales Universitatis Turkuensis C 409). Turun yliopisto. (Väitöskirja).

Ivaska, Ilmari & Bernardini, Silvia. 2020. Constrained language use in Finnish: A corpus-driven approach. *Nordic Journal of Linguistics* 43(1). 33–57. https://doi.org/10.1017/S0332586520000013

Jantunen, Jarmo. 2001. ”Tärkeä seikka” ja ”keskeinen kysymys”: Mitä korpuslingvistinen analyysi paljastaa lähisynonyymeista? *Virittäjä* 105(2). 170–192.

Jantunen, Jarmo. 2015. Oppimiskontekstin vaikutus oppijanpragmatiikkaan: Astemääritteet leksikaalisina nallekarhuina. *Lähivõrdlusi. Lähivertailuja* 25. 105–136. https://doi.org/10.5128/LV25.05

Jarvis, Scott. 2013. Capturing the Diversity in Lexical Diversity: Lexical Diversity. *Language Learning* 63. 87–106. https://doi.org/10.1111/j.1467-9922.2012.00739.x

Jääskeläinen, Anni. 2023. Adjektiivien polysemiaa konstruktioissa. *Virittäjä* 127(3). 316–344. https://doi.org/10.23982/vir.84893

Kecskes, Istvan & Papp, Tünde. 2000. *Foreign language and mother tongue*. Lawrence Erlbaum Associates.

Kekki, Niina & Ivaska, Ilmari. 2022. The use of synonymous adjectives of Finnish as a second language learners: Applying the MuPDAR(F) approach. *International journal of learner corpus research* 8(1). 67–96. https://doi.org/10.1075/ijlcr.21006.kek

Kekki, Niina & Jytilä, Riitta & Parente-Čapková, Viola. 2023. Kulttuurisen kielenoppimisen jäljillä: Aikuisten suomenoppijoiden kokemuksia lukupiirissä. *Sananjalka* 65(65). 196–214. https://doi.org/10.30673/sja.115631

Kielitoimiston sanakirja. 2022. https://www.kielitoimistonsanakirja.fi (Viitattu 1.8.2023)

Kim, Myonghee. 2004. Literature Discussions in Adult L2 Learning. *Language and Education* 18(2). 145–166. https://doi.org/10.1080/09500780408666872

Krashen, Stephen. 1993. *The power of reading: Insights from the research*. Libraries Unlimited.

Langacker, Ronald. 1987. *Foundations of cognitive grammar. Vol. 1: Theoretical prerequisites*. Stanford University Press.

Lee, Ching-Yin & Liu, Jyi-Shane. 2009. Effects of Collocation Information on Learning Lexical Semantics for Near Synonym Distinction. *International Journal of Computational Linguistics & Chinese Language Processing* 14(2). 205–220. https://aclanthology.org/O09-4004

van Lier, Leo. 2000. From input to affordance: Social-interactive learning from an ecological perspective. Teoksessa Lantolf, James (toim.), *Sociocultral Theory and Second Language Learning*, 245–260. Oxford University Press.

Liu, Dilin. 2013. Salience and construal in the use of synonymy: A study of two sets of near-synonymous nouns. *Cognitive Linguistics* 24(1). 67–113. https://doi.org/10.1515/cog-2013-0003
Liu, Dilin & Zhong, Shouman. 2016. L2 vs. L1 Use of Synonymy: An Empirical Study of Synonym Use/Acquisition. *Applied Linguistics* 37(2). 239–261. https://doi.org/10.1093/applin/amu022
Lu, Yuanwen. 2017. *A corpus study of collocation in Chinese learner English*. Routledge.
Lyons, John. 1968. *Introduction to theoretical linguistics*. Cambridge University Press. https://doi.org/10.1017/CBO9781139165570
Martin, Marilyn. 1984. Advanced Vocabulary Teaching: The Problem of Synonyms. *The Modern Language Journal* 68(2). 130–137.
Nation, Paul. 2018. Reading a whole book to learn vocabulary. *ITL - International Journal of Applied Linguistics* 169(1). 30–43. https://doi.org/10.1075/itl.00005.nat
Pietilä, Päivi & Merikivi, Riikka. 2014. The Impact of Free-time Reading on Foreign Language Vocabulary Development. *Journal of Language Teaching and Researc*h 5(1). 28–36. https://doi.org/10.4304/jltr.5.1.28-36
R Core Team. 2022. *R: A Language and Environment for Statistical Computing* [Software]. R Foundation for Statistical Computing. https://www.R-project.org/
Reynolds, Barry. 2015. A Mixed-Methods Approach to Investigating First- and Second-Language Incidental Vocabulary Acquisition Through the Reading of Fiction. *Reading Research Quarterly* 50(1). 111–127. https://doi.org/10.1002/rrq.88
Reynolds, Barry. 2022. Situated incidental vocabulary acquisition: The effects of in-class and out-of-class novel reading. *Applied Linguistics Review* 13(5). 705–733. https://doi.org/10.1515/applirev-2019-0059
Reynolds, Barry & Ding, Chen. 2022. Effects of word-related factors on first and second language English readers' incidental acquisition of vocabulary through reading an authentic novel. *English Teaching: Practice & Critique* 21(2). 171–191. https://doi.org/10.1108/ETPC-05-2021-0049
Ringbom, Håkan. 2007. *Cross-linguistic similarity in foreign language learning*. Multilingual Matters.
Sinclair, John. 1991. *Corpus, concordance, collocation*. Oxford U.P.
Stefanowitsch, Anatol & Gries, Stefan. 2003. Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics* 8(2). 209–243. https://doi.org/10.1075/ijcl.8.2.03ste
Tomasello, Michael. 2003. *Constructing a language: A usage-based theory of language acquisition.* Harvard University Press.
Vanhatalo, Ulla. 2003. Kyselytestit vs. Korpuslingvistiikka lähisynonyymien semanttisten sisältöjen arvioinnissa: Mitä vielä keskeisestä ja tärkeästä? *Virittäjä* 107(3). 351–369.
Waring, Rob & Nation, Paul. 2004. Second language learning and incidental vocabulary learning. *Angles on the English Speaking World* 4. 97–110.
Waring, Rob & Takaki, Misako. 2003. At what rate do learners learn and retain new vocabulary from reading a graded reader? *Reading in a Foreign Language* 15(2). 130–163.
Webb, Stuart. 2008. The effects of context on incidental vocabulary learning. *Reading in a Foreign Language* 20(2). 232–245.
Xiao, Richard & McEnery, Tony. 2006. Collocation, Semantic Prosody, and Near Synonymy: A Cross-Linguistic Perspective. *Applied Linguistics* 27(1). 103–129. https://doi.org/10.1093/applin/ami045

Liite A Tutkimuskysely

Kyselylomake (Liite A) on saatavilla osoitteessa: https://osf.io/d2b9y/

Yhteystiedot:

Niina Kekki
Turun yliopisto
niina.kekki@utu.fi

Ilmari Ivaska
Turun yliopisto
ilmari.ivaska@utu.fi

« Derrière OGM il y a *modifié* ! » : Étude sémantique sur la plurivocité du sigle *OGM*

Kim Lehtonen
Université de Turku

Résumé

Dans sa forme longue, *OGM* peut être écrit *organisme génétiquement modifié* ou *organisme génétiquement manipulé*. Cet article a pour but d'offrir une description sémantique de la notion d'OGM et d'en connaître les conceptions des locuteurs natifs du français. Pour ce faire, des francophones ont été interrogés dans le cadre d'un sondage sur la forme longue qu'ils préfèrent et les raisons de choisir cette forme. Leurs justifications sont abordées en prenant en compte des aspects sémantiques, normatifs et idéologiques afin de donner un regard détaillé sur la signification du sigle *OGM*. Les résultats de l'article montrent qu'une majorité des répondants du sondage ont préféré l'adjectif *modifié*. Plusieurs répondants ont trouvé que *manipulé* véhicule difficilement l'idée de changement qui leur était essentielle dans la signification d'*OGM*. Les réponses retenues esquissent une vision complexe de l'OGM qui pourrait servir pour des analyses ultérieures.

Mots-clés : OGM, organisme génétiquement modifié, organisme génétiquement manipulé, sémantique des possibles argumentatifs, sigle

Abstract

The French acronym for GMO (*OGM*) can be read in its long form as "genetically modified organism" (*organisme génétiquement modifié*) or "genetically manipulated organism" (*organisme génétiquement manipulé*). This article aims to describe the French GMO concept and see how the French speakers conceive of it. In order to do this, a group of French speakers were asked which of the long forms they prefer and why they chose that form. The justifications are discussed under semantic, normative, and ideological aspects in order to produce a detailed view of the concept. The results of the article show that the majority of the participants preferred the adjective *modifié*. Many participants found that *manipulé* does not convey well the idea of change that they found essential to the signification of *OGM*. The answers to the survey sketch out a complex vision of the OGM concept that could serve for further analysis.

Keywords: GMO, genetically modified organism, genetically manipulated organism, semantics of argumentative possibilities, acronym

1 Introduction

L'utilisation de l'expression *organisme génétiquement* ***manipulé*** est peut-être dépassée : la plupart des dictionnaires français, sauf *Le Grand Robert*, donnent seulement la forme longue *organisme génétiquement* ***modifié*** pour OGM. Quant à l'expression *organisme*

VERTAISARVIOITU
KOLLEGIALT GRANSKAD
PEER-REVIEWED
www.tsv.fi/tunnus

Finnish Journal of Linguistics 37 (2024), 159–180
DOI: 10.61197/fjl.142765

génétiquement manipulé, elle semble aller contre la norme discursive. D'une part, il s'agit d'un effet sémantique, et d'autre part, la volonté d'utiliser *manipuler* semble relever d'un choix idéologique[1]. Dans cet article, nous étudions les conceptions possibles de la notion d'*OGM* (*organisme génétiquement modifié* et/ou *manipulé*).

La notion d'OGM a déjà été prise en compte dans des recherches en sémantique. Par exemple, le sigle OGM a été le point de départ d'une étude sur la sémantique des lexèmes *manipuler*, *manipulation* et *modifier*, *modification* dans le discours bioéthique, lorsqu'ils désignent des actions qui portent sur du vivant (Lehtonen 2022a et b). Tout récemment, les sémantismes de *manipuler* et *modifier* ont été explorés afin de savoir si le choix du verbe est guidé par une préférence combinatoire avec des mots comme *embryon*, *cerveau*, *ADN*, etc. (Cozma & Lehtonen 2023).

Le but du présent article est 1) d'offrir une description de la notion d'OGM ; 2) de connaître les conceptions possibles des locuteurs natifs du français à travers un sondage où ils sont interrogés pourquoi ils trouvent l'une des formes plus adaptée que l'autre en parlant d'*OGM* et 3) de comprendre leurs préférences pour les adjectifs *manipulé/modifié* quand on parle d'OGM. Ce qui nous intéresse est de voir quels types de différences les répondants conçoivent entre ces deux notions. Pour ce faire, nous construirons d'abord une description sémantique pour le sigle *OGM* et nous étudierons ensuite comment il peut être compris de manières différentes. Comme point de départ, nous nous servons de la description des lexèmes *manipuler* et *modifier* proposée dans Cozma et Lehtonen (2023) pour savoir ensuite quelles orientations sont supposées être mobilisées par les expressions *organisme génétiquement modifié* et *organisme génétiquement manipulé*. Cette information sera comparée avec les réponses obtenues à l'aide d'un sondage.

Les répondants de notre sondage doivent dire s'ils préfèrent la forme *organisme génétiquement modifié* ou *organisme génétiquement manipulé*, et pourquoi. Pour répondre à ce pourquoi, les participants doivent justifier leur choix en s'appuyant sur leurs connaissances métalinguistiques. Nous supposons que les répondants ont à considérer trois perspectives quand ils sont demandés à choisir l'un des adjectifs :

1. Les raisons sémantico-pragmatiques, telles que la compatibilité des mots et l'orientation axiologiquement négative de *manipulé*, offrent la base qui rend possibles les différentes combinaisons et qui conduit aux différentes conceptions d'OGM.
2. La forme la plus répandue, à savoir *modifié*, façonne la perception sur ce qui est l'usage « correct »[2].
3. Des raisons culturellement motivées (ou idéologiques) peuvent faire préférer l'adjectif moins fréquent, *manipulé*. Dans ce cas, nous supposons qu'il y a surtout une volonté de présenter les OGM (et par conséquent, le génie génétique) sous une lumière négative.

---

[1] L'idéologie comme ensemble des opinions influence la conduite d'une personne, y compris ses choix linguistiques. La connexion entre l'idéologie et la langue utilisée est visible par exemple dans le fait que la forme longue *organisme génétiquement manipulé* a été décrite comme une dénomination employée par les « détracteurs de la culture des OGM » (Depecker 2013 : 16). L'utilisation du verbe *manipuler* – et de l'adjectif *manipulé* – peut orienter à considérer que l'acte est critiquable, ce qui est noté également dans Lehtonen (2022a).

[2] Cet usage se voit en France surtout sur les emballages étiquetés selon leur contenu. Dans l'Union européenne, les produits doivent être étiquetés lorsqu'ils contiennent des organismes génétiquement modifiés (Règlement (CE) n° 1829/2003). De plus, en France, les produits dont il existe des espèces génétiquement modifiées mais qui n'en contiennent pas, peuvent être étiquetés comme des produits « sans OGM » (Décret n° 2012-128).

Les grandes lignes ci-dessus correspondent à une vision de la normativité où les normes sont divisées en règles et en principes (cf. Mäkilähde, Leppänen & Itkonen 2019 : 4−7). Les règles déterminent comment la langue est utilisée correctement : dans la première perspective, la question porte sur la compatibilité sémantique[3]. En revanche, les principes opèrent sur le niveau de la force illocutionnaire où le choix de l'expression peut signaler la visée de l'énonciateur, ce qui est manifesté dans la deuxième et la troisième perspective. En étudiant ce qui conduit à préférer l'une ou l'autre des formes longues, nous faisons l'hypothèse que les conceptions des répondants sur *OGM* varient selon l'adjectif (sur la base de ce qui a été fait dans Cozma & Lehtonen 2023).

Cette étude sera divisée en trois parties. Dans la section 2, nous présentons notre point de vue théorique, la sémantique des possibles argumentatifs et les mécanismes sémantico-discursifs qui touchent le sémantisme d'*OGM*. À l'intérieur de la partie théorique, nous discutons également les descriptions sémantiques de *manipuler* et *modifier*, et proposons une description d'*OGM* à partir des dictionnaires non spécialisés. Dans la section 3, nous présentons notre sondage, et dans 4, nous analysons les réponses en les comparant à la description formulée dans la section 2.4.

2 Partie théorique

Dans cette section, nous abordons la sémantique des possibles argumentatifs (développée par Olga Galatanu 1999 ; 2022) et certains mécanismes sémantico-discursifs décrits dans ce cadre. Ensuite, nous offrons un bref regard sur les descriptions sémantiques des lexèmes *manipuler* et *modifier* (d'après Cozma & Lehtonen 2023) qui nous servira plus tard pour comprendre les raisonnements des répondants au sondage. Pour finir, nous proposons une description sémantique d'*OGM*, formulée à partir des entrées dictionnairiques.

2.1 La sémantique des possibles argumentatifs

Nous adoptons la vision de la semantique des possibles argumentatifs (SPA) (Galatanu 1999 ; 2018 ; 2022) sur la construction de la signification lexicale et du sens discursif. Développée par Olga Galatanu, la SPA est un modèle sémantique qui se situe dans la filiation des sémantiques argumentatives, se sert de l'idée de stéréotype linguistique et se veut un outil pour l'analyse linguistique du discours.

La signification et le sens[4] sont stratifiés d'après la SPA – d'après l'idée de stéréotype de Putnam (1975 : 249–250 ; cf. aussi Galatanu 2018 : 60–61). La signification est divisée en éléments stables constituant le noyau et en éléments culturellement établis[5] nommés stereotypes. Les stéréotypes génèrent des possibles argumentatifs – à un niveau prédiscursif – qui se manifestent finalement dans le discours comme des deploiements

[3] Dans la vision de la théorie à laquelle nous nous attachons, les effets pragmatiques qu'un élément de la langue peut évoquer sont intégrés dans la signification. C'est pourquoi nous parlons de raisons sémantico-pragmatiques.

[4] La distinction entre langue et parole, entre signification et sens utilisée notamment par Ducrot (cf. 1972) a été reprise par Galatanu pour la SPA.

[5] Par éléments culturellement établis, nous entendons tout ce qui est distinctif dans un groupe donné de locuteurs et qui, de ce fait, est visible dans la signification d'un mot : dans le cas d'OGM, il s'agit d'abord de l'entité dans sa matérialité et des moyens de sa création (qui exigent des connaissances et des techniques) ainsi que des savoirs intellectuels et affectifs qui influencent la manière dont l'entité est perçue.

ARGUMENTATIFS (id. : 163–167). Notre étude s'intéresse surtout à la signification et pour cela, nous décrirons la notion d'OGM à partir des entrées des dictionnaires. Utiliser les dictionnaires fait partie de la méthode proposée dans la SPA (id. : 264–267) : ceux-ci devraient contenir les informations essentielles de la signification recherchée (les éléments du noyau) et quelques stéréotypes culturellement établis, comme il est souvent impossible de décrire un lexème dans sa totalité dans une seule entrée de dictionnaire.

Cette vision stratifiée du sémantisme intègre certaines pratiques de la sémantique argumentative (Anscombre & Ducrot 1983 ; Carel & Ducrot 1999). La SPA ne recourt pas à une métalangue complexe mais s'appuie sur les mots de la langue et présente leurs interconnexions, leurs orientations de l'un vers l'autre, qui sont mises en évidence à l'aide des connecteurs logiques abstraits DONC et POURTANT, ainsi qu'à l'aide de la négation, notée « nég- » (Galatanu 2018 : 282–286). Le connecteur DONC montre le caractère normatif et admissible de l'orientation argumentative, tandis que POURTANT indique une orientation transgressive et nécessite un raisonnement plus complexe que pour les orientations en DONC.

Selon la vision adoptée par la SPA, la modalisation est inscrite dans la signification des mots (Galatanu 2018 : 87). Les valeurs modales typiquement reconnues dans l'étude de la modalisation (aléthiques, déontiques, épistémiques, etc.) sont incluses dans ce modèle (Galatanu 2005 : § 11). De plus, le sémantisme d'un mot peut véhiculer une attitude polaire, une modalité axiologique qui concerne par exemple le bien et le mal, le beau et le laid, etc. En donnant un statut central aux modalités axiologiques, il est possible d'analyser le lexique en observant les attitudes négatives ou positives. Ces attitudes déjà inscrites dans la signification du mot servent surtout dans l'analyse linguistique du discours. Par rapport à la modalisation qui figure dans les sens de *manipuler* et *modifier*, la modalisation axiologique négative (exprimant le mépris) et déontique (pour interdire ces actions) semblent les plus saillantes (Lehtonen 2022b : 186–189).

2.2 Mécanismes sémantico-discursifs

Dans cette section, nous abordons deux mécanismes sémantico-discursifs qui influencent les significations d'*OGM*. D'abord, nous décrivons les modificateurs sémantiques qui font baisser ou augmenter la force des stéréotypes d'un mot, un mécanisme qui peut expliquer certaines accentuations dans le discours. Ensuite, nous discutons la contamination discursive, un effet non négligeable dans le discours bioéthique, qui ajoute à la modalisation axiologique négative du mot *OGM*.

Dans la notion d'OGM, *manipulé* et *modifié* fonctionnent comme des adjectifs qui qualifient le nom *organisme*. Ils peuvent donc être considérés comme des modificateurs sémantiques (dans les termes de Ducrot 1995) qui affaiblissent ou renforcent la force argumentative d'un mot. Par exemple, pour le mot *parent*, le modificateur réalisant renforce les stéréotypes de *parent*, comme dans le syntagme *un parent proche*, et un modificateur déréalisant les affaiblit, comme dans le syntagme *un parent éloigné*. Or, dans le cas des modificateurs sémantiques, il est intéressant de remarquer qu'ils fonctionnent en même temps pour renforcer certains stéréotypes et pour en affaiblir d'autres comme cela a été mis en évidence par Cozma & Galatanu (2019 : 263–264) dans une étude sémantique de la notion de *démocratie*. En étudiant des syntagmes tels que *démocratie représentative*, *directe*, *participative* et *du contrôle*, les deux auteures font appel à la notion de modificateur sémantique pour expliquer ce qui se passe au niveau sémantique quand un caractérisant est utilisé pour modifier le nom *démocratie*.

La modification sémantique n'opère pas par hasard. Le modificateur touche certains stéréotypes en fonction de son propre sémantisme. Par exemple, dans l'expression *démocratie représentative*, *représentative* fonctionne comme modificateur réalisant pour des stéréotypes de *démocratie* comme « élections » et « vote » ; en même temps, il déréalise le stéréotype « participation active des citoyens » (Cozma & Galatanu 2019 : 264).

Par contamination discursive, on comprend une synergie des significations, où les significations des mots fonctionnent ensemble pour s'influencer les unes les autres (Cozma & Galatanu 2019 : 260–262). Ce mécanisme fait qu'une orientation argumentative absente de la signification du mot étudié est ajoutée au sens discursif de ce mot – un effet contraire à la modification sémantique qui influence les stéréotypes déjà présents dans la signification d'un mot. Dans leur étude de la notion de démocratie, Cozma & Galatanu (2019) notent que les mots axiologiquement négatifs qui l'accompagnent ajoutent leurs orientations à celles de *démocratie*. Elles montrent que cet effet est réalisé dans le discours avec des mots ou des syntagmes axiologiquement monovalents comme *détester* et *souffrir d'une maladie*, voire en utilisant des mots axiologiquement bivalents comme *réveiller* et *moderniser*.

Sans faire appel à la notion de contamination discursive, Lehtonen (2022b) étudie *manipuler* et *modifier* dans le contexte de la bioéthique. Il distingue les collocations *embryon*, *cerveau* et *ADN* pour identifier des orientations argumentatives du type « action de manipuler l'embryon DONC risque ». Dans toutes les catégories identifiées, il constate que *manipuler* et *modifier* sont influencés par des mots axiologiquement négatifs utilisés dans les commentaires laissés dans le cadre d'une consultation citoyenne. D'après les significations de *manipuler* et *modifier* dans le contexte de la bioéthique telles que présentées dans Cozma & Lehtonen (2023 : § 99), les orientations axiologiques négatives ne sont pas complètement inattendues mais exigent une activation de ce potentiel.

2.3 Les descriptions sémantiques de *manipuler* et *modifier*

Nous avons déjà évoqué à plusieurs reprises les significations de *manipuler* et *modifier* ; nous les présentons brièvement dans cette section, car elles nous serviront pour comprendre la notion d'OGM. Cozma & Lehtonen (2023) proposent une description sémantique des deux verbes, qu'ils utilisent ensuite comme une base pour étudier la combinatoire des mots et les représentations dans la bioéthique (cf. aussi Cozma & Lehtonen 2024). Pour eux, les éléments stables de la signification (les éléments du noyau) de *manipuler* sont organisés de la manière suivante :

main/instrument DONC action de tenir DONC série d'actions organisées

Dans cette description, ils mettent l'accent sur le fait d'utiliser la main et par la suite, la volonté de l'agent d'agir sur un patient (souvent sans que le patient s'en rende compte)[6]. La totalité de cette action est une opération qui n'engendre pas nécessairement un changement. Ils constatent également que le patient, qui fait l'objet de l'acte de *manipuler*, est plus saillant que l'agent : beaucoup d'importance est donnée aux différents objets dans les dictionnaires.

[6] Les notions d'agent, de patient et d'instrument relèvent de la grammaire des cas introduite par Fillmore (1968).

Selon la description de Cozma & Lehtonen (2023), seule l'orientation axiologiquement négative est inscrite dans la signification de *manipuler*, ce qui veut dire que le mot a typiquement une orientation négative. Ces orientations sont visibles à partir des stéréotypes, attachés notamment à l'élément du noyau « série d'actions organisées » :

série d'actions organisées DONC magouille, fraude, vision faussée de la réalité

Cette vision est confirmée par leurs répondants, qui complètent la liste des stéréotypes – la bioéthique toujours à l'esprit – par exemple dans les mêmes lignes que « magouille », etc. : « malhonnêteté, mensonge, trucage ».

Quant à *modifier*, la description proposée est la suivante :

action de changer POURTANT essence non altérée DONC traits/parties changées

Le premier élément du noyau, « action de changer », donne déjà une des distinctions centrales par rapport à *manipuler* : un changement doit être visé et, ainsi, le mot décrit un procès transitionnel. Ensuite, la description prend en compte l'essence qui ne change pas malgré le premier élément, « action de changer ». Toutefois, les auteurs font la remarque que le mot oriente fortement vers « les traits changés », ce qui surpasse dans l'usage l'essence qui reste la même. Étroitement lié à l'essence de l'objet modifié, Cozma & Lehtonen (2023 : § 21) expliquent que celui-ci fonctionne à la fois comme l'objet des modifications et comme le lieu où la modification est effectuée. Ils illustrent ce phénomène avec des exemples de dictionnaires comme *modifier un passage dans un écrit*, où le patient est conçu à la fois comme une partie et comme un tout.

De par ses stéréotypes, *modifier* est un mot axiologiquement bivalent : il peut décrire des situations où les éléments changés sont orientés en bien ou en mal (p. ex. « amélioration » et « dégradation »). De plus, les données analysées ont fourni beaucoup de stéréotypes axiologiquement positifs, liés à la guérison, par exemple « soigner », « sauver », « traitement de maladie génétique », etc. (Cozma & Lehtonen 2023 : § 55).

2.4 La description sémantique du sigle *OGM*

Les sigles participent à la construction du sens comme les autres entités linguistiques[7] (Courbon, Lambert & Dion-Girardeau 2016 : 200–201) et, par conséquent, ils peuvent être soumis à une étude sémantique. Nous considérons que ce traitement est validé par le fait que les sigles se combinent avec d'autres mots et se mettent en relation avec des sigles similaires, comme test *ADN* et *ADN–ARN* (id. : 205). *OGM* fonctionne similairement, pour former, par exemple, *produit OGM* et *OGM–PGM* (*plante génétiquement modifiée*).

OGM est un sigle plurivoque (dans les termes de Courbon et al. 2016), qui peut provoquer des formes longues différentes, similairement que *CB – carte bancaire* ou *carte bleue*. Comme mentionné plus haut, les dictionnaires ne font pas de distinction entre un organisme génétiquement modifié et un organisme génétiquement manipulé, étant donné

[7] Excepté toutefois les sigles de circonstance qui doivent être fortement liés à leurs formes longues. Les sigles de circonstance sont des abréviations de syntagmes récurrents dans un texte qui ne sont pourtant pas en usage régulier : par exemple, ADN a été utilisé comme sigle d'un domaine spécifique avant de devenir courant (Courbon et al. 2016 : 186).

que les deux syntagmes font appel au même référent (bien que de manières distinctes). L'utilisation des sigles est justifiée par un accès plus rapide à la réalité (Courbon et al. 2016 : 205) ; une réalité remise en question dans le cas de l'OGM.

Des techniques agricoles ont été utilisées depuis longtemps : déjà 10 000 ans avant notre ère, des plantes et des animaux ont été domestiqués et soumis à la sélection artificielle (FAO 2004 : 11). L'Organisation des Nations unies pour l'alimentation et l'agriculture, FAO (idem.), définit deux ères selon l'apparition des techniques liées à l'hérédité : l'ère conventionnelle (qui commence vers la fin du 19e siècle avec les découverts de Gregor Mendel) et l'ère moderne (commençant dans les années 1970 avec les premières techniques de recombinaison d'ADN). Cette division est visible dans la législation de l'Union européenne[8], où l'OGM est défini comme un organisme dont le matériel génétique est combiné d'une manière qui ne s'effectue pas par recombinaison naturelle. Cette définition précise aussi les techniques du génie génétique qui sont utilisées pour produire des OGM ainsi que les techniques exclues, comme la fécondation in vitro et la conjugaison des bactéries.

Les OGM provoquent régulièrement des controverses en France et ailleurs, comme les avancées scientifiques font évoluer les techniques de la modification génétique (Demortain 2015 : 122–123). La notion d'OGM recouvre donc des organismes issus de techniques multiples, et le manque de notions a été critiqué pour la difficulté à communiquer autour des développements en biotechnologie (Tournay & Pagès 2015 : 236) : surtout la définition juridique, qui range toutes ces techniques sous une seule notion, est utilisée comme illustration de cette pratique problématique. De plus, comme la naturalité est une qualité déterminante dans les définitions officielles, la volonté de l'Homme de toucher au génome des vivants est vue par certains comme un continuum qui inclut également les croisements d'espèces (id. : 237–238). Cette différence entre les manières multiples de produire des OGM rend difficile l'évaluation de leur innocuité ou leurs effets bénéfiques (id. : 231).

Dans ce qui suit, nous proposons une description sémantique d'*OGM* à partir des dictionnaires *Le Grand Robert*, *Dictionnaire de l'Académie française*, *Larousse* et *Wikipédia*, dictionnaires où le mot *OGM* a une entrée propre. Wikipédia, en tant qu'encyclopédie, a été sélectionné à côté des dictionnaires afin que nous puissions inclure également des stéréotypes plus précis issus des sciences mais qui peuvent être inscrits dans une signification « naïve ». Quant au *Trésor de la Langue Française informatisé* (*TLFi*), il ne contient pas d'entrée pour *OGM*. Dans la Figure 1, les éléments du noyau sont organisés à gauche. Il s'agit des éléments stables de la signification d'*OGM : organisme, génome* et *acte de modifier/manipuler* qui sont à lire de haut en bas. Ces éléments sont vectoriels (Galatanu 2018 : 164) : ils sont organisés, ils ont une direction, et leur relation est marquée à l'aide des connecteurs abstraits DONC ou POURTANT. Ces éléments sont prolongés par des stéréotypes, présentés dans la colonne de droite. La liste des stéréotypes n'est pas exhaustive, comme le modèle de la SPA conçoit ces éléments comme prédisposés à des changements graduels par l'ajout de nouveaux éléments.

[8] Directive 2001/18/CE, article 2.

Les éléments du noyau	Les stéréotypes
organisme	→ vivant → être vivant → animal → plante, végétal → bactérie → aliment(s) → propriété, caractère → naturel, état naturel
DONC	
génome	→ gène, ADN → héritage → patrimoine génétique
POURTANT	
acte de modifier/manipuler	→ intervention humaine → nouvelle propriété → résistance aux parasites, au gel, aux herbicides → nég-l'état naturel → nég-naturel → ajout, insertion, transgénèse, transfert → donneur → induction artificielle de mutations, mutagénèse → gène étranger, gène d'une autre espèce → génie génétique → sélection artificielle, croisement

Figure 1. Le noyau et les stéréotypes d'*OGM* selon les dictionnaires

Dans cette description, les éléments nucléaires « organisme DONC génome » décrivent un fait biologique, à savoir le fait que tous les organismes vivants comportent un génome. Ce fait est modifié avec l'orientation transgressive « génome POURTANT acte de modifier/manipuler », une orientation qui montre qu'à l'état naturel, le génome d'un organisme n'est pas modifié par l'Homme. Cet aspect transgressif, qui signale qu'il s'agit d'un cas particulier, est présent par exemple dans la définition du *Grand Robert* : « Organisme dont le génome a été modifié […] afin de lui conférer une propriété qu'il ne possède pas naturellement. »

Au niveau des stéréotypes, nous trouvons dans les dictionnaires la mention des différents organismes touchés par le génie génétique et une remarque sur le fait que les organismes portent des propriétés. L'élément du noyau « acte de modifier/manipuler » est lié à des stéréotypes qui portent sur les techniques et les nouvelles propriétés obtenues. Ici, nous trouvons également un indice sur comment *modifié* et *manipulé* fonctionnent comme modificateurs sémantiques : le stéréotype « naturel » lié à *organisme* est renversé par l'acte de modifier ou manipuler pour donner « nég-naturel ». Que la signification d'*OGM* consiste en des stéréotypes opposés, est en accord avec la vision de la SPA selon

laquelle les éléments des stéréotypes peuvent être non concordants. Cette complexité est possible grâce à une dynamique culturelle et subjective de la construction du sens (Galatanu 2018 : 176–178).

3 Données et méthode

Pour effectuer cette recherche, nous avons lancé un sondage, où nous avons demandé l'avis des locuteurs natifs du français sur l'utilisation de la notion d'OGM. Ce sondage a été réalisé sous format numérique sur Webropol, une plateforme dédiée aux questionnaires, et il a été ouvert pour des réponses de mars 2023 à mai 2023. Le sondage a été distribué sur des listes de diffusion des universités françaises et sur des réseaux sociaux. Ainsi, il se peut qu'une partie des répondants aient une formation en linguistique. Aucune connaissance préalable sur la sémantique n'était demandée aux répondants, mais le sondage suppose que les répondants connaissent le sigle OGM. Pour cette étude, nous prenons en compte toutes les réponses, sans considérer les variables sous-jacentes, comme l'âge des répondants (qui varie de 22 à 85 ans, avec une moyenne de 43 ans) ou leur genre. Dans le sondage, nous avons posé la question suivante :

> Derrière le sigle OGM, on peut avoir : a) « organisme génétiquement modifié » ou b) « organisme génétiquement manipulé ». À votre avis, quelle expression est la plus adaptée en parlant des OGM : a) ou b) ? Et pourquoi ?

Nous avons obtenu la réponse de 51 personnes, nombre qui nous semble suffisant pour cette enquête qualitative sur la sémantique du sigle OGM. Sur ces 51 personnes, 44 (86,3 %) ont indiqué que a) *organisme génétiquement modifié* est plus adapté en parlant des OGM, et 7 (13,7 %) ont choisi b) *organisme génétiquement manipulé*. Sur l'ensemble des répondants, 44 ont donné une justification pour leur choix. La longueur de leurs justifications varie entre 1 mot et 91 mots. Dans les justifications, les répondants évoquent des éléments sémantiques, normatifs et idéologiques, et certains pensent que les deux alternatives sont possibles selon le contexte d'occurrence.

Dans l'analyse, nous classifierons les justifications selon la nature des raisons évoquées, en prenant en compte leurs aspects sémantiques, normatifs et idéologiques (voir la section 1). Cette division est bien sûr artificielle, et nous sommes conscient que les justifications idéologiques sont étroitement liées aux différences sémantiques que les répondants ont conçues entre ces deux variantes. Nous présumons que la compatibilité sémantique est une étape importante dans la justification du choix, qui suscite des réponses sur le sémantisme d'*OGM*, de *modifier* et *manipuler*. Les réponses basées sur des considérations sémantiques serviront pour vérifier la description d'*OGM* présentée dans 2.4 et pour connaître les conceptions possibles des formes différentes d'*OGM*.

Le sondage nous offre également un regard sur les principes d'usage. Nous avons demandé quelle est la préférence des répondants entre *organisme génétiquement modifié* et *organisme génétiquement manipulé*. Nous considérons que cette préférence est étroitement liée à la normativité, que nous discuterons dans 4.2. Ensuite, toujours lié aux principes d'usage, nous avons anticipé qu'*organisme génétiquement manipulé* sera chargé d'une orientation axiologique négative que les francophones veulent soit éviter soit mettre en évidence. Cette visée idéologique qui apparaît dans les justifications sera discutée dans 4.3.

Dans ce qui suit, nous analyserons les justifications données par les participants au sondage. Une réponse peut faire appel à plusieurs manières de justifier le choix ; dans ces cas, nous présenterons les extraits de ces justifications à l'intérieur des sections auxquelles ils s'appliquent.

4 Le sigle OGM vu par les répondants au sondage

Pour organiser cette section, nous nous servons de l'hypothèse sur les trois raisons de choisir l'adjectif *manipulé* ou *modifié* que nous avons formulée dans l'introduction. Dans un premier temps (sous-section 4.1), nous discutons le sémantisme du sigle *OGM* en comparant les réponses obtenues avec les descriptions présentées en 2.3 et 2.4. Dans un deuxième temps (sous-sections 4.2 et 4.3), nous analysons les réponses qui relèvent de la normativité et celles qui abordent les principes culturels gouvernant le choix de *manipulé* ou *modifié*.

4.1 Les raisons sémantiques pour le choix de l'adjectif

Dans cette sous-section, nous discutons les raisons sémantiques liées au choix de *modifié* ou *manipulé* et le rejet de l'autre. La diversité des réponses qui s'inscrivent dans cette catégorie peut être appréhendée à l'aide de quelques regroupements, autour des idées suivantes : 1) le changement inscrit dans la signification de *modifier*, 2) l'opposition entre deux conceptions de la notion d'OGM correspondant à des techniques différentes, 3) l'orientation axiologique d'*OGM* et 4) les stéréotypes mobilisés dans les justifications.

4.1.1 Le changement effectué et sa permanence

Le changement produit sur l'organisme est un aspect sur lequel les répondants s'appuient souvent pour justifier leur choix. Les justifications présentées ici mentionnent le changement, la permanence du changement, l'objet des changements et le moyen de produire ce changement. Certains, comme dans (1)–(4), font appel uniquement au changement – toujours en lien avec la réponse *modifié* –, ce qui montre que *manipuler* ne véhicule pas le stéréotype « changement ».

(1) Car on a opéré un changement concret

(2) Je dirais la a). Modifié me semble plus connoté du caractère final de la manipulation. La modification connote à mon avis davantage le changement.

(3) Il y a effectivement modifications

(4) Réponse a. Les ogm sont censés être transformés afin d'acquérir de nouvelles propriétés.

D'autres répondants considèrent que la *modification* effectuée sur l'organisme est permanente – ce qui donne également lieu à la conception selon laquelle une *manipulation* est réversible, un point que nous discutons plus bas.

(5) A - Car l'organisme touché n'est plus sous forme initiale de façon permanente.

(6) […] Ces organismes sont modifiés dans le sens où ce n'est pas réversible […]

(7) Modifié car il ne pourra plus revenir à son état initial (sauf erreur de ma part)

(8) Modifié, la finalité est irrémédiable. S'il pouvait être à nouveau modifié alors il serait manipulable…

Certains font appel à la combinatoire de *modifié* et *manipulé* et mentionnent les « gènes », comme en (9)–(12). Ces exemples soulignent le fait que les gènes sont modifiés ; avec *OGM*, ce n'est pas l'action de manipuler les gènes qui compte, mais le résultat obtenu, d'avoir des gènes modifiés. Ainsi, la signification essentielle d'*OGM* est le changement effectué, idée qui ne peut pas être véhiculée avec le verbe *manipuler*. Selon cette logique, la « manipulation génétique » n'engendre pas de changement.

(9) A). La manipulation n'est pas « génétique », c'est la modification qu'elle provoque qui l'est.

(10) parce qu'il y a une intervention sur le contenu même du génome, ce qui a impliqué un changement de ses caractéristiques en termes d'aspect physique, comportement par rapport à l'environnement, immunité etc.

(11) a - modifié car une action de modification génétique est réalisée

(12) a) parce qu'il y a eu une modification des gènes, on les a changés

Finalement, le moyen d'obtenir un OGM est discuté dans (13), qui évoque le stéréotype « main » de *manipuler*, et (14), qui fait appel aux types de procès selon lesquels *manipuler* ne pourrait pas exprimer un résultat obtenu.

(13) […] car on n'a pas seulement touché au génome, mais on l'a activement altéré

(14) […] peut-être une question de conceptualisation du procès : modifier me paraît désigner le résultat de tout le processus de modification, alors que manipuler pourrait être 1 étape dans ce processus, mais pas le résultat final

Nous pouvons également classer au sein de cette sous-catégorie portant sur le changement quelques justifications qui apportent des précisions sur le changement effectué : *modifier* désigne fortement un changement effectué ; ce changement est permanent ; *manipuler* et *génome* sont incompatibles dans le cas d'*organisme génétiquement manipulé* ; et association « manipuler DONC main » (qui représente un possible argumentatif d'*OGM*) bloque le stéréotype d'OGM « nouvelle propriété ».

4.1.2 Deux conceptions différentes d'*OGM* : une question d'époque et de technique

Certains répondants font appel à la distinction des différentes techniques liées à la modification du génome : *organisme génétiquement modifié* correspond aux techniques modernes de la recombinaison d'ADN, tandis qu'*organisme génétiquement manipulé* fait écho aux techniques conventionnelles de l'agriculture, telles que la sélection artificielle. Les trois justifications suivantes expriment directement le lien entre *manipulation* et les techniques « conventionnelles » utilisées depuis longtemps déjà, faisant ainsi intervenir un ordre chronologique : la manipulation génétique est alors perçue comme étant antérieure et liée au passé.

(15) Derrière OGM il y a modifier ! car avant cela toutes les recherches des agriculteurs par exemple étaient de la sélection de graines ou d'espèces donc manipulation si on peut dire

(16) […] on n'a pas juste influencé le changement comme on aurait pu le faire avec les croisements de semences comme par le passé, on va directement 'modifier' le génome pour être certain du résultat.

(17) a) modifié pour dire transformé radicalement. L'humanité manipule les végétaux, les organismes vivants, depuis des siècles.

Dans l'exemple (15), les deux conceptions sont clairement juxtaposées : pour le répondant, *organisme génétiquement manipulé* est issu de la sélection de graines ou d'espèces, une opération effectuée à la main et perceptible à l'œil. La réponse en (16) fait la même distinction mais va plus loin, contrastant influence indéterminée et résultats certains. La justification en (17) ajoute à cette distinction que *modifier* exprime une transformation dans ce contexte. L'idée de changement véhiculée par l'adjectif *modifié* vient d'être discutée en 4.1.1, mais en (17), ce changement est radical et, de plus, il est mis en opposition avec l'activité de *manipuler*.

La justification en (18) est entièrement basée sur la distinction entre modification et manipulation de l'ADN du point de vue de la technique utilisée. Le répondant mentionne, d'une part, la transgénèse (l'introduction d'un gène étranger dans une autre espèce) comme un moyen de modifier l'ADN et, d'autre part, la modification épigénétique qui ne change pas l'ADN de l'organisme. De plus, la remarque sur la gradualité sous-entend que les modifications et les manipulations de l'ADN ont un degré de naturalité différent.

(18) On introduit parfois un gène d'une autre espèce, on modifie donc l'ADN d'origine. La manipulation pour moi correspondrait à utiliser uniquement l'ADN du sujet mais choisir l'expression des gènes ou les combinaisons des allèles. [ce n'est pas exactement ce que je veux montrer ; pourtant, cette remarque montre que la naturalité est graduelle]

La réponse en (18) est également intéressante d'un point de vue sémantique, car elle précise les actants impliqués dans le processus : pour la modification de l'ADN, une autre entité biologique que celle visée par la modification est également impliquée.

Dans le Tableau 1, nous présentons les éléments d'analyse obtenus dans cette sous-section, en les mettant en rapport avec les éléments du noyau d'*OGM* (qui figurent dans la colonne de gauche : « organisme DONC génome POURTANT acte de modifier/manipuler » ; cf. 2.4). Nous les organisons dans deux colonnes parallèles, en leur donnant le statut de stéréotypes, conformément à la représentation sémantique en SPA.

Tableau 1. Les deux conceptions d'*OGM* (selon la technique et l'époque)

Noyau	Stéréotypes liés à la conception moderne d'*OGM* (basée sur l'action de modifier)	Stéréotypes liés à la conception conventionnelle d'*OGM* (basée sur l'action de manipuler)
organisme	connaît un changement bactérie, virus, plante	traité à la main plante, animal
génome	gène, héritage	gène, héritage
acte de modifier/ manipuler	transgénèse, gène étranger, mutagénèse nég-naturel nouvelle propriété résultat permanent résultat certain transformation	sélection artificielle, croisement, hybride naturel nouvelle propriété changement résultat incertain

Dans ce tableau, les stéréotypes qui sont attachés à l'élément du noyau « organisme » sont différents selon les deux visions : la conception moderne met en avant « bactérie » et « virus », par contraste avec « plante » et « animal » de la conception conventionnelle qui sont touchés par le croisement artificiel. De plus, le tableau souligne la différence quant à ce que subit l'organisme quand il est modifié ou manipulé. Au niveau du « génome », nous ne trouvons pas de différence de stéréotypes entre les deux conceptions.

La plus grande différence du Tableau 1 se trouve au niveau de l'acte effectué, *modification* ou *manipulation*. Nous y trouvons les techniques spécifiques comme « transgénèse » et « mutagénèse » liées aux organismes génétiquement modifiés, en opposition avec « sélection artificielle », « croisement », des opérations basées sur l'hérédité mais qui ne nécessitent pas un laboratoire. Comme l'a remarqué l'un de nos répondants, les techniques s'inscrivent sur un continuum de la naturalité ; toutefois, nous nous focalisons sur l'opposition entre les techniques modernes non naturelles et les techniques conventionnelles naturelles. Cette opposition montre également comment les adjectifs *manipulé* et *modifié* fonctionnent en tant que modificateurs pour réorienter le mot *OGM*. Finalement, d'après ce qu'ont indiqué plusieurs répondants, nous observons une différence entre la certitude des opérations : l'acte de modifier est plus concret et exact, par contraste avec l'acte de manipuler, qui est inexact et difficile à concevoir dans le cadre de la bioéthique.

Bien que ces différences entre *organisme génétiquement modifié* et *organisme génétiquement manipulé* ne soient pas établies sur la base de textes définitoires, nous considérons que cette catégorisation exprime une conceptualisation éclairante d'*OGM* en lien avec *modifier/manipuler* : cette distinction semble correspondre également à l'idée

des modificateurs sémantiques qui renforcent ou affaiblissent certaines potentialités de sens. De plus, la comparaison réunit surtout des éléments discutés en lien avec le changement effectué : pour les répondants, *modifier* exprime fortement un changement permanent et *manipuler* prend difficilement l'ADN comme objet dès qu'il évoque l'idée de « main ».

4.1.3 L'orientation axiologique

L'orientation axiologique est fortement liée aux raisons idéologiques que nous discutons dans 4.3. Pourtant, nous avons voulu la prendre en compte également dans les raisons sémantiques. C'est pourquoi nous faisons une distinction entre ces deux catégories : ici, nous présentons les justifications qui mentionnent une orientation négative ou positive mais n'explicitent pas le raisonnement derrière le sentiment des répondants. Dans 4.3, nous discuterons des justifications plus complexes, par exemple, qui indiquent le mécontentement face aux conséquences sociales des OGM.

Dans les trois grandes lignes que nous avons présentées au début (§ 1), nous avons suggéré que le choix de l'adjectif se fait selon des considérations sémantiques, normatives et/ou idéologiques. Dans cette approche, il arrive que des répondants utilisent une même raison pour justifier des choix opposés. C'est par exemple le cas de l'orientation axiologiquement négative de *manipuler*, qui amène les répondants tantôt à éviter son usage, tantôt à le mettre en avant, en fonction de la visée adoptée. Cet effet est visible dans la justification suivante (19), où le répondant remarque que la question est idéologique.

(19) […] Pour choisir l'une ou l'autre des expressions, il faut savoir si l'on soutient cette façon de faire ou pas. […]

Nous discutons cet exemple plus en détail dans la sous-section 4.3, mais il contraste avec les justifications suivantes (20)–(23) où les répondants ont choisi *organisme génétiquement modifié* précisément pour éviter l'orientation négative de *manipuler*.

(20) a) – « manipulé » laisse entendre qu'il y a un résultat néfaste ou au moins questionnable ; c'est donc moins neutre

(21) […] Je pense que modifié souligne l'approche des scientifiques et la deuxième est plus critique

(22) a) parce que 'manipulé' a une connotation négative, alors que 'modifié' est plus neutre, voire positif dans ce contexte

(23) Organisme génétiquement modifié: on retrouve l'idée de changer pour le meilleur quelque chose, donc de le modifier (là où "manipulé" implique simplement la possibilité d'observer l'élément).

Toutes ces justifications ont en commun le fait que *modifier* est perçu comme étant neutre, voire positif, ce qui est tout à fait en accord avec les significations de *modifier* et *manipuler*. Avec *modifier*, les répondants mentionnent les scientifiques (21) et l'amélioration (23) – *amélioration* étant un stéréotype applicable uniquement à *modifier*. Pour *manipuler*, (20)

évoque le stéréotype *résultat* (*néfaste*). L'orientation négative de manipuler est utilisée surtout pour justifier son évitement. Ainsi, choisir l'option *organisme génétiquement manipulé* semble nécessiter une prise de position idéologique claire.

4.1.4 Les stéréotypes mobilisés par *modifié* et *manipulé*

Les justifications contiennent de nombreux stéréotypes de *modifier* et *manipuler*. Dans cette sous-section, nous discutons de ceux qui n'entrent pas dans les autres catégories de l'analyse. Par exemple, les justifications (24)–(27) évoquent le possible argumentatif « manipuler DONC modifier » qui indique que l'acte de manipuler est concomitant avec l'acte de modifier, ou qu'il le précède et y conduit.

(24) je dirais a parce que c'est plus franc, plus honnête et c'est plus simple à comprendre. […] En effet, il me semble que a recouvre b : intuitivement, je dirais qu'il ne peut pas y avoir de manipulation anodine. Tel que je le comprends, toute manipulation est une intervention humaine : celle-ci ne peut faire autrement que de laisser une trace, donc une forme de modification, aussi légère soit-elle.

(25) a) modifié, c'est le résultat de manipulé

(26) A) parce que le résultat de la manipulation dans ce cas est la modification

(27) a) parce que la modification "sur le terrain" est le résultat obtenu par suite de manipulations réussies en laboratoire

En fait, le stéréotype « manipuler DONC modifier » rend possible l'utilisation de l'expression *organisme génétiquement manipulé*, puisque l'intervention humaine qui vise un changement est essentielle dans la signification d'*OGM*. *Modifié* est aussi perçu comme étant plus concret que *manipulé*, ce qui va de pair avec le stéréotype « manipuler DONC opération » évoqué en (24).

Du côté opposé, l'extrait (28) évoque l'association inverse « modifier DONC manipuler », qui montre que les sens de *modifier* et *manipuler* se sont rapprochés fortement dans le contexte de la bioéthique (ce qui a été remarqué également dans Cozma & Lehtonen 2023 : § 98).

(28) […] à force de modifier, on en arrive à manipuler : croisement, sélection, mutagénèse, où va-t-on s'arrêter ?

Cependant, l'enchaînement n'est pas complètement inattendu, car il a comme effet de donner à *modifier* la même orientation axiologique négative que celle contenue dans *manipuler*. De plus, la progression « croisement, sélection, mutagénèse » suivie de la question « où va-t-on s'arrêter ? » va dans le même sens, puisqu'elle sous-entend une évaluation négative de la part de l'énonciateur.

Dans l'extrait (29), le répondant justifie sa préférence pour *organisme génétiquement modifié*.

(29) […] parce que c'est plus précis que manipulé, car avec manipuler on ne sait pas ni quel est le but ni ce qui a été effectué, alors que avec modifié on a l'organisme (avant) et l'organisme génétiquement modifié (après)

L'idée selon laquelle l'opération effectuée semble moins exacte quand le verbe *manipuler* est utilisé, tandis que *modifié* est plus précis, a déjà été discutée en 4.1.2. En (29) nous pouvons noter le stéréotype « but » – but imprécis dans le cas de *manipuler* et but précis et clair dans le cas de *modifier* – ainsi que le changement (*avant–après*) inscrit dans le noyau *modifier*.

Dans la justification (30), le répondant discute surtout l'incompatibilité de *manipulé* dans la notion d'OGM.

(30) Organisme génétiquement modifié. Manipulé semble artificiel et forcé pour introduire une connotation.
Manipulation me fait aussi toujours penser à une action a la main par un individuel qui intervient personnellement avec un objectif, ce qui ne colle pas vraiment

Plusieurs stéréotypes liés à *manipuler* sont évoqués dans (30), *main*, *intention de l'agent* et *but*. Ces orientations argumentatives sont incompatibles quant à la signification d'OGM : *manipuler* a une orientation axiologique négative et cette orientation conduit à considérer que la visée énonciative d'*organisme génétiquement manipulé* serait excessivement idéologique.

Une fois discutées toutes ces raisons d'ordre sémantique, nous pouvons nous tourner vers la deuxième raison annoncée dans l'introduction, celle de la normativité.

4.2 Une norme qui guide le choix de l'adjectif

Nous avons commencé cet article en remarquant qu'*organisme génétiquement modifié* est le plus fréquent en usage. Les réponses au sondage vont dans ce sens : 86 % des répondants ont préféré la forme avec *modifié*. Nous avons supposé également que la fréquence d'usage devrait apparaître dans les justifications pour le choix. En effet, l'usage et la norme sont des justifications récurrentes dans le sondage, avec 14 répondants qui mentionnent l'usage (sur le total de 44 qui justifient leur choix). Un bon nombre des répondants mentionnent tout simplement qu'*organisme génétiquement modifié* est la forme qu'ils connaissent le mieux, comme dans (31)–(34).

(31) La a) je suppose, c'est la seule que j'ai entendue.

(32) A, car j'ai entendu la réponse A et non la réponse B

(33) c'est celle qu'on entend le plus, notamment dans le domaine agricole

(34) J'ai plutôt entendu la première version. […]

Cette habitude de voir la forme en *modifié* (35)–(37) donne l'impression que *manipulé* serait inacceptable dans cette position (36).

(35) a) parce que j'ai l'habitude de l'entendre et de le lire comme ça […]

(36) A. Mais parce que je suis habitué à la formule "organisme génétiquement modifié", choisir B me semble alors décalé.

(37) "organisme génétiquement modifié" sont les mots habituellement associés au sigle.

Le fait qu'une formule soit utilisée depuis longtemps rentre dans la catégorie de la normativité. En réalité, les deux formes ont été utilisées depuis le développement rapide du génie génétique et des OGM au tournant de la décennie 1980–1990[9]. Pourtant, certains répondants ont le sentiment qu'*organisme génétiquement modifié* a été utilisé pendant plus longtemps que sa contrepartie.

(38) A. Expression employée depuis longtemps. […]

(39) a) Il me semble que c'est le terme le plus usité et premier. Est-ce que manipulé ne serait pas venu plus tard ?

(40) le terme OGM a été tout de suite inscrit comme modifié […]

Les exemples (41)–(42) indiquent aussi qu'*organisme génétiquement modifié* est plus courant, avec la précision (en 41) que cette forme est utilisée sur les emballages pour indiquer si le contenu est sans ou avec OGM.

(41) a), c'est la transcription la plus courante du sigle, celle qu'on retrouve dans les produits de consommation (par exemple les boîtes de maïs dit "sans OGM").

(42) C'est la réponse a) qui est utilisée couramment […]

Enfin, en (43)–(44), l'utilisation d'*organisme génétiquement modifié* est issue d'une convention officielle – ce qui est d'ailleurs conforme à l'utilisation dans la législation.

(43) officiellement c'est "modifié" […]

(44) Réponse a tout simplement car c'est le véritable acronyme

Pour conclure sur les justifications liées à la normativité, comme nous avons supposé, la plupart des répondants ont préféré la forme *organisme génétiquement modifié* et plusieurs ont justifié leur choix en évoquant directement ou indirectement la fréquence d'utilisation : s'ils n'ont jamais vu écrit *organisme génétiquement manipulé*, c'est que cette version est plus rare. Plusieurs répondants ont indiqué qu'*organisme génétiquement modifié* est

[9] Voir par exemple la presse : « "Mutants en cavale". Des chercheurs californiens pulvérisent sur les cultures des organismes vivants manipulés. Les écologistes s'inquiètent », publié dans *Le Monde*, le 6 mai 1987 ; « Marchands de vie. Micro-organismes, plantes, animaux : depuis dix ans, la brevetabilité du vivant semble poursuivre une marche irréversible. Jusqu'à l'espèce humaine ? », publié dans *Le Monde*, le 26 juin 1991.

la forme qu'ils ont lue ou entendue. D'autres ont cherché à argumenter qu'*organisme génétiquement modifié* a été utilisé plus longtemps, ce qui n'est pas tout à fait vrai mais dans l'étendue de notre travail, nous ne pouvons pas prendre en compte le possible changement d'usage dans le temps. Deux répondants ont noté qu'il existe une norme officielle concernant ces notions. Certes, la forme utilisée dans la législation et dans les dictionnaires est un indice conduisant à considérer *organisme génétiquement modifié* comme standardisé.

4.3 Les raisons culturellement motivées pour le choix de l'adjectif

Par raisons culturellement motivées, nous entendons les raisons qui ne sont pas directement issues du sémantisme d'*OGM*. Cette séparation entre sémantisme et idéologie est artificielle, surtout du point de vue de la SPA, théorie selon laquelle la culture est inscrite dans les significations des mots. Néanmoins, puisque nous avons demandé aux répondants de choisir leur forme longue préférée, ils ont probablement parcouru les trois grandes lignes que nous avons définies au départ de cette étude. Ces éléments sont visibles dans la justification suivante (45).

(45) Je pense que a) est plus couramment utilisé mais […] pour moi, b) correspond mieux car j'y vois des tests en laboratoire, des recherches, une volonté de contrôle absolu sur cet organisme.

Selon (45), *organisme génétiquement modifié* est la version la plus fréquente mais il choisit *organisme génétiquement manipulé* pour viser des stéréotypes spécifiques. De plus, il exemplifie les stéréotypes visés : « OGM DONC laboratoire, recherche, volonté de celui qui manipule » qui font tous partie des stéréotypes de *manipuler* (identifiés par Cozma & Lehtonen 2023 : § 16). *Laboratoire* et *recherche* n'ont guère de charge axiologique mais *volonté de contrôle absolu* détient une orientation négative.

Certains répondants remarquent que le choix entre *modifié* et *manipulé* indique l'attitude de l'énonciateur, c'est-à-dire ils perçoivent qu'une modalisation axiologique négative fait partie du sens d'*organisme génétiquement manipulé*. Les répondants des passages suivants ont choisi *modifié*.

(46) […] Pour choisir l'une ou l'autre des expressions, il faut savoir si l'on soutient cette façon de faire ou pas. Je ne suis pas suffisamment familiarisée à cette problématique pour avoir un avis tranché.

(47) […] polémiquement ça pourrait être "manipulé". C'est plus politiquement correct de dire modifié

Ces deux répondants montrent que le choix de l'adjectif a une base idéologique. Celui de (46) dit clairement qu'il fallait avoir une prise de position pour savoir choisir entre les deux et garde une distance par rapport à ce sujet. Celui de (47) souligne que choisir *manipulé* dans ce contexte montre la volonté de l'énonciateur de susciter une polémique alors que choisir *modifié* montre une attitude impartiale.

D'autres vont plus loin pour dire que *manipulé* serait adéquat pour évoquer exprès des orientations négatives liées à *OGM*.

(48) […] si ce changement implique la raréfaction de certaines espèces végétales, on peut parler d'organisme génétiquement manipulé. Je précise toutefois que je n'ai pas les connaissances scientifiques nécessaires pour valider (ou non) le lien entre cette raréfaction et l'usage des OGM.

(49) […] le terme est aussi plus passable au point de vue du grand public sachant que les OGMs sont déjà controversés.

(50) Tout dépend de la nature des modifications mais disons la b) car le sujet me paraît déjà plus vraiment sous contrôle

En (48), le répondant exprime qu'il est acceptable d'utiliser la version à l'orientation axiologique négative, *organisme génétiquement manipulé*, quand les conséquences négatives sont présentes. Pourtant, ce répondant ne prend pas en charge cette association. L'extrait (49) propose un stéréotype « OGM DONC controverse » et indique que l'utilisation de *manipulé* est un choix plus adapté par rapport aux OGM si l'on prend en compte l'opinion publique. En (50), le répondant exprime qu'il y existe différentes manières de modifier un organisme mais arrive à choisir *manipulé* pour indiquer un mécontentement sur l'utilisation actuelle des OGM.

Comme nous l'avons vu à partir de la description sémantique, *OGM* en soi n'est pas orienté vers une modalisation axiologique négative. Les répondants en (51) et (52) suivent le même raisonnement que les précédents : ils choisissent *manipulé* pour indiquer une attitude négative envers les OGM. Il semble qu'ils aient eu des attendes vis-à-vis des OGM mais ont été déçus.

(51) b) j'aimerais que les OGM servent à résoudre des situations difficiles, mais les conséquences pour les autres organismes vivants sont à mon sens problématiques et incontrôlables

(52) Aujourd'hui je dirai b) car nous constatons que les modifications au départ censées être au profit de l'humanité se retrouvent finalement être responsables de misère dans bien des pays.

Ces justifications contiennent les déploiements argumentatifs optimistes « OGM DONC résolution des difficultés, profit de l'humanité », qui ne sont pourtant pas pris en charge. Au lieu de cela, ces répondants proposent les déploiements argumentatifs « OGM DONC conséquences problématiques/ incontrôlables ; misère », et ils trouvent que ces orientations sont plus facilement obtenues en utilisant la forme *organisme génétiquement manipulé*.

La réponse en (53) semble combiner les différentes significations de *manipuler*.

(53) Réponse b. J'ai choisi le terme manipulation il semble adapté car après la modification des gènes, il y a une volonté de la faire admettre par le plus grand nombre donc : manipulation des opinions dans un but mercantile.

Le répondant choisit *organisme génétiquement manipulé*. Pourtant, il traite les opérations de génie génétique comme une modification et explique que tout ce qui vient après (la volonté d'influencer les opinions pour vendre des produits OGM) est de la manipulation.

Dans l'extrait (54), *organisme génétiquement manipulé* est considéré plus vague et moins fort que sa contrepartie.

(54) […] Je vois b comme une sorte d'euphémisme servant à dédramatiser les choses en passant sous silence la conséquence concrète de la manipulation. […]

Le répondant de cet extrait a préféré *organisme génétiquement modifié*. Pour lui, *modifié* laisse moins à deviner, comme cet adjectif exprime directement le changement effectué. Cet extrait montre qu'*organisme génétiquement modifié* peut être conçu de manière qui vise des orientations axiologiquement négatives et que la force de cette orientation peut dépasser celle d'*organisme génétiquement manipulé*, conçu uniquement comme ayant une orientation négative.

Dans cette sous-section nous avons discuté les raisons culturellement motivées pour le choix entre *organisme génétiquement modifié* et *organisme génétiquement manipulé*. Nous avons constaté que les répondants ont considéré la normativité, même lorsqu'ils ont choisi la version moins fréquente, *manipulé*. D'après les réponses, il s'agit effectivement d'un choix idéologique pour les répondants et il semble que le fardeau de la prise en charge de l'orientation négative du verbe *manipuler* peut conduire à l'éviter. Pour d'autres, les raisons idéologiques liées aux difficultés produites par les OGM ont pesé plus et ces répondants ont choisi *organisme génétiquement manipulé* pour indiquer leur attitude critique envers les OGM.

De plus, les répondants ont proposé des déploiements argumentatifs qui aident à comprendre l'utilisation d'*OGM* en discours. Les déploiements argumentatifs « OGM DONC laboratoire, recherche, volonté de celui qui manipule » proposés par un des répondants nous indiquent qu'*OGM* incorpore les stéréotypes de *manipuler* et *modifier* (identifiés par Cozma & Lehtonen 2023). Les autres déploiements argumentatifs étendent la description sémantique d'*OGM* – « OGM DONC controverse » n'est pas décrit dans les dictionnaires mais fait partie des savoirs culturellement établis et s'inscrit ainsi dans les stéréotypes ; quant aux « OGM DONC conséquences problématiques/ incontrôlables ; misère », surtout les conséquences sont discutées dans la bioéthique (cf. la remarque de Lehtonen 2022 : 195–196), et ces éléments s'ajoutent aux stéréotypes.

5 Conclusion

Dans cet article, nous avons abordé l'utilisation de la notion d'*OGM* en nous basant sur une distinction entre *organisme génétiquement modifié* et *organisme génétiquement manipulé*. Notre but était d'offrir une description d'*OGM*, de la préciser à partir des conceptions des locuteurs natifs du français et de voir s'ils préfèrent une des formes. Cet article a cherché à s'interroger sur l'utilisation de la notion d'OGM – une notion qui recouvre plusieurs techniques et cache des nuances pertinentes pour comprendre le débat qui concerne l'environnement et l'alimentation. En même temps, il s'agit d'un travail sur une notion scientifique devenue quasiment quotidienne, et se rattache à l'effort de comprendre la propagation des mots d'un contexte à un autre.

Nous avons d'abord proposé, à l'aide des entrées de dictionnaires, une description de la notion d'OGM contenant les éléments stables de signification « organisme DONC génome POURTANT action de modifier/manipuler ». Le dernier élément, « action de modifier/manipuler », est prolongé par des stéréotypes qui orientent, entre autres, vers

« intervention humaine », « nouvelle propriété », « transgénèse » et « génie génétique ». Étant donné que la liste des stéréotypes (qui sont culturellement établis) d'un mot est toujours incomplète, nous avons eu recours à un sondage dont les réponses nous ont servi à enrichir cette liste. Ce sondage a attiré les réponses de 51 personnes, un nombre assez limité, mais bien suffisant pour cette recherche – dans le cas d'un nombre plus grand, les justifications utilisées seraient probablement répétitives. Les répondants ont évoqué des stéréotypes tels que « OGM donc controverse, conséquences problématiques/incontrôlables, misère », qui illustrent une attitude négative à l'égard des OGM.

À l'aide du sondage, nous avons trouvé que notre échantillon des répondants ont préféré la notion *organisme génétiquement modifié* à celle d'*organisme génétiquement manipulé*. Il semble que pour justifier l'usage d'une notion ou de l'autre, sont considérées les raisons sémantiques permettant de vérifier si les mots (dans ce cas *organisme*, *génome* et *manipuler/modifier*) sont compatibles. Lié à cette considération, ils tiennent compte de la forme qui suit un principe d'usage et du positionnement idéologique qui vont avec le choix de la notion.

Les justifications montrent que le changement effectué sur un organisme est une orientation essentielle qui est difficilement véhiculée par *manipulé*. Le changement et sa permanence font apparaître également une distinction entre *organisme génétiquement modifié* qui serait issu des techniques modernes et *organisme génétiquement manipulé* qui serait, par exemple, le résultat d'un croisement. Finalement, au niveau de la modalisation, *organisme génétiquement modifié* est perçu comme neutre, voire positif, et *organisme génétiquement manipulé* est perçu comme étant négatif. Vérifier ces résultats sur des données authentiques et spontanées, et continuer par une analyse discursive de ces notions serait, bien entendu, une piste intéressante pour des recherches ultérieures.

Les dictionnaires utilisés

Le Grand Robert électronique. (https://grandrobert.lerobert.com/robert.asp). (Consulté 2023-12-28).
Dictionnaire de l'Académie française. (https://www.dictionnaire-academie.fr). (Consulté 2024-1-4).
Larousse : dictionnaire de français. (https://www.larousse.fr/dictionnaires/francais-monolingue). (Consulté 2024-1-4).
Wikipédia. (https://fr.wikipedia.org). (Consulté 2024-1-4).

Bibliographie

Anscombre, Jean-Claude & Ducrot, Oswald. 1983. *L'argumentation dans la langue*. Liège : Mardaga.

Carel, Marion & Ducrot, Oswald. 1999. Le problème du paradoxe dans une sémantique argumentative. *Langue française* 123. 6–26. https://doi.org/10.3406/lfr.1999.6293

Courbon, Bruno & Lambert, Maxime & Dion-Girardeau, Samuel. 2016. La fabrique du sigle : entre focalisation référentielle et (re)dénomination. *Neologica* 10. 171–216. https://doi.org/10.15122/isbn.978-2-406-06279-0.p.0171

Cozma, Ana-Maria & Galatanu, Olga. 2019. La construction discursive dévalorisante du concept de démocratie. *Neuphilologische Mitteilungen* 119(2). 249–272.

Cozma, Ana-Maria & Lehtonen, Kim. 2024. Combinatoire lexicale et profilage du sens du lexème embryon dans le discours de la bioéthique. *Neuphilologische Mitteilungen* 125(2). 222–252. https://doi.org/1051814/nm.145708

Demortain, David. 2015. Comment faire preuve en régime de controverse ? Retour sur l'histoire de l'évaluation des OGM. *Hermès* 73. 122–128. https://doi.org/10.3917/herm.073.0122

Depecker, Loïc. 2013. Pour une ethnoterminologie. In Quirion, Jean & Depecker, Loïc & Rousseau, Louis-Jean (éds.), *Dans tous les sens du terme*, 13–29. Ottawa : Presses de l'Université d'Ottawa

Ducrot, Oswald. 1972. Langue et parole. In Ducrot, Oswald & Todorov, Tzvetan (éds.), *Dictionnaire encyclopédique des sciences du langage*, 155–161. Paris : Éditions du Seuil.

Ducrot, Oswald. 1995. Les modificateurs déréalisants. *Journal of Pragmatics* 24(1). 145–165. https://doi.org/10.1016/0378-2166(94)00112-R

FAO – Food and Agriculture Organization of the United Nations. 2004. *La situation mondiale de l'alimentation et de l'agriculture 2003-2004*. Rome : Organisation des Nations Unies pour l'alimentation et l'agriculture.

Fillmore, Charles. 1968. The case for case. In Bach, Emmon & Harms, Robert T. (éds.), *Universals in linguistic theory,* 1–88. New York : Holt, Rinehart and Winston.

Galatanu, Olga. 1999. Argumentation et analyse du discours. In Gambier, Yves & Suomela-Salmi, Eija (éds.), *Jalons 2*, 41–55. Turku : Université de Turku.

Galatanu, Olga. 2005. La sémantique des modalités et ses enjeux théoriques et épistémologiques dans l'analyse des textes. In Gouvard, Jean-Michel (éd.), *De la langue au style*. Lyon : Presses universitaires de Lyon. https://doi.org/10.4000/books.pul.20785

Galatanu, Olga. 2018. *Sémantique des possibles argumentatifs : génération du sens discursif et (re) construction des significations linguistiques.* Bruxelles : Peter Lang.

Galatanu, Olga. 2022. Sémantique des possibles argumentatifs. In Biglari, Amir & Ducard Dominique (éds.), *La sémantique au pluriel*, 99–119. Rennes : Presses Universitaires de Rennes.

Lehtonen, Kim. 2022a. *Manipuler/modifier le vivant : entre bioconservatisme et bioprogressisme. Étude des orientations argumentatives des lexèmes manipuler et modifier dans un débat de bioéthique.* Université de Turku. (Mémoire de master). (https://www.utupub.fi/handle/10024/153635). (Publié 2022-2-21).

Lehtonen, Kim. 2022b. Manipulation et modification scientifique : un regard sur les orientations argumentatives. *Synergies pays riverains de la Baltique* 16. 179−198.

Mäkilähde, Aleksi & Leppänen, Ville & Itkonen, Esa. 2019. Norms and normativity in language and linguistics: Basic concepts and contextualisation. In Mäkilähde, Aleksi & Leppänen, Ville & Itkonen, Esa (éds.), *Normativity in language and linguistics*, 1–28. Amsterdam : John Benjamins. https://doi.org/10.1075/slcs.209.01mak

Putnam, Hilary. 1975. The meaning of 'meaning'. *Philosophical papers* vol. 2, 215–271. Cambridge: Cambridge University Press.

Tournay, Virginie & Pagès, Jean-Christophe. 2015. OGM : un terme polysémique à l'épreuve de la communication et de l'évaluation. *Hermès* 73. 233−243. https://doi.org/10.3917/herm.073.0233

Coordonnées :

Kim Lehtonen
Université de Turku
Courriel : khjleh@utu.fi

Dalmi, Gréte & Witkoś, Jacek & Cegłowski, Piotr (toim.). 2020. *Approaches to Predicative Possession: The View from Slavic and Finno-Ugric*. (Bloomsbury Studies in Theoretical Linguistics). London: Bloomsbury Academic. 228 s.

Kirjoittanut Maria Kok

1 Johdanto

Approaches to Predicative Possession. The View from Slavic and Finno-Ugric on tiivis ja jäntevä artikkelikokoelma omistuslauseiden tai lausemaisten omistusrakenteiden typologiasta slaavilaisissa ja uralilaisissa kielissä. Vaikka teoksen nimi viittaa suomalais-ugrilaisiin kieliin, yksi teoksen artikkeleista (ks. luku 10) käsittelee predikoivia omistusrakenteita selkupin murteissa, ja kuten artikkelin kirjoittaja aivan oikein huomauttaa, samojedikieliin kuuluva selkuppi ei ole suomalais-ugrilainen vaan uralilainen kieli. Kaikki teoksen artikkelit on kirjoitettu englanniksi.

Teos koostuu johdantoartikkelista sekä yhdeksästä tutkimusartikkelista, joista neljä käsittelee predikoivia omistusrakenteita slaavilaiskielissä ja viisi uralilaisissa kielissä. Teoksen päättää lyhyt loppukatsaus, jossa vedetään yhteen tutkimusartikkeleiden tärkeimpiä lopputuloksia sekä pohditaan jatkonäkymiä predikoivien omistusrakenteiden tutkimukselle erityisesti niissä kielissä, joissa ilmiötä on toistaiseksi kuvattu vähän.

1.1 Mitä ovat predikoivat omistusrakenteet?

Predikoivilla omistusrakenteilla tarkoitetaan lauseen laajuisia tai lausemaisia omistusilmauksia, joista ainakin osaa voidaan nimittää – tai myös nimitetään – omistuslauseiksi. Termiä predikoivat omistusrakenteet (*Predicative Possession*) käytetään erottamaan puheena olevat ilmaukset adnominaalisista tai lausekkeen sisäisistä omistusrakenteista, kuten omistusta ilmaisevista attribuuteista tai possessiivisuffikseista.

Kielitypologisessa tutkimuksessa predikoivat omistusrakenteet on tapana ryhmitellä kahteen päätyyppiin: transitiivisen *habeo*-verbin ympärille rakentuviin HAVE-possessiiveihin (1) sekä eksistentiaalisiin BE-possessiiveihin (2), jotka sisältävät olemassaoloa ilmaisevan *esse*-verbin (’olla’) merkitysvastineen. Näiden lisäksi tunnetaan kopulatiivisia possessiiveja (3, 4), joista verbi joko puuttuu tai se on kopulatiivisen *olla*-verbin merkitysvastine.[1]

1) I have a dog.
2) Minulla on koira.
3) The dog is mine.
4) Koira on minun.

[1] Johdantoluvussa ei ole käyttöesimerkkejä lainkaan, joten olen konstruoinut tähän itse muutaman perustapauksen.

DOI: 10.61197/fjl.149385

Niissä indoeurooppalaisissa kielissä, joissa transitiivinen *habeo*-verbi esiintyy, suositaan HAVE-possessiiveja, kun uralilaisissa kielissä BE-possessiivit ovat tavallisempia. On myös kieliä, kuten valkovenäjä, joissa molempaa päätyyppiä esiintyy. Kopulatiivisia omistusrakenteita esiintyy sekä HAVE- että BE-possessiiveja suosivissa kielissä. Onkin esitetty (esim. Benveniste 1966; Freeze 1992; Kayne 1993), että transitiiviset HAVE-possessiivit ja eksistentiaaliset BE-possessiivit olisivat kumpikin johdettavissa kopulatiivisista BE-possessiiveista ja kehittyneet niistä erilaisten kieliopillisten transformaatioiden kautta. Käsillä olevassa teoksessa ei sitouduta tähän ns. derivationaaliseen lähestymistapaan vaan tuodaan pikemminkin esille sen ongelmakohtia ja pyritään etsimään sille vaihtoehtoja.

Arvioitavan artikkelikokoelman eri luvuissa tarkastellaan, millaisia omistusrakenteita tutkimuksen kohteena olevissa kielissä esiintyy ja mitä päätyyppiä ne edustavat. Mikäli tarkasteltavassa kielessä esiintyy useita erityyppisiä omistusrakenteita, niiden välistä vaihtelua ja keskinäistä suhdetta pohditaan aina siitä tutkimuksellisesta kehyksestä käsin, jota artikkelin kirjoittaja on soveltanut. Lisäksi artikkeleissa tarkastellaan kolmea kieliopillista ilmiötä, jotka usein esiintyvät predikoivien omistusrakenteiden yhteydessä: slaavilaisissa kielissä on tavallista ns. kiellon genetiivi -sääntö eli *Genitive of Negation* (GoN), jonka mukaan transitiiviverbin objektin sija vaihtuu akkusatiivista genetiiviin kiellon vaikutuksesta. Myös eksistentiaalilauseen läsnäolevan jäsenen tai omistuslauseen omistettavan jäsenen sija vaihtuu kieltolauseessa nominatiivista genetiiviin. Uralilaisten kielten omistusrakenteisiin liittyy taas usein possessiivisuffiksien käyttö, ja kussakin kielessä on omat käytänteensä liittyen mm. possessiivisuffiksien pakollisuuteen tai valinnaisuuteen sekä niiden taipumukseen painottaa pääsanaansa tai ilmaista määräisyyttä. Yleinen, joskus lähes kieliuniversaaliksi katsottu ilmiö (ks. Milsark 1979) on puolestaan määräisyysvaikutus eli *Definiteness Effect* (DE), joka säätelee mm. englannin eksistentiaalilauseiden rakennetta. Niissä kielissä, joissa DE vaikuttaa, eksistentiaalilauseiden teeman on oltava merkitykseltään epämääräinen, eikä se voi saada määritteikseen esim. kvanttoreita, joilla on rajaava ja definiittinen merkitys, kuten ’kaikki’, ’molemmat’ tai ’jokainen’. Koska BE-possessiivit ovat usein olemukseltaan eksistentiaalilauseita, voi DE vaikuttaa myös omistuslauseiden rakenteeseen sekä määritteisiin, jotka omistettava voi saada.

1.2 Johdantoartikkeli

Teoksen avaa Gréte Dalmin kirjoittama johdantoartikkeli *Introduction*, jossa esitellään aluksi tutkimuskysymys ja teoksen rakenne sekä predikoivien omistusrakenteiden pääryhmät Stassenin (2009) typologian mukaan. Johdannon luvussa 1.2 esitellään aikaisempaa tutkimusta sekä mainitaan lyhyesti myös kiellon genetiivi sekä DE, josta käytetään tässä luvussa kuitenkin nimitystä *Definiteness Restriction*. Termien ja lyhenteiden yhdenmukaistamista olisi ehkä voinut harkita, sillä muihinkin ilmiöihin viitataan useilla eri nimityksillä eri kirjoittajien artikkeleissa. Esimerkiksi kiellon genetiivistä käytetään johdantoluvussa merkintätapaa (GEN NEG), kun osassa artikkeleita suositaan merkintää GoN. Toisaalta lukijan on hyödyllistä tietää, millaista termistöä tai vaihtoehtoisia merkintätapoja eri tutkimustraditioissa käytetään.

Johdantoartikkelin teorialuku 1.2. on lyhyt ja tiivis, vaikkakin selkeä. Olisin suonut, että esimerkiksi Stassenin (2009) tutkimusta olisi avattu hieman yksityiskohtaisemmin, vaikka predikoiva omistusrakenne onkin käsitteenä varsin ymmärrettävä. Myös jokunen käyttöesimerkki tai lainaus esim. tärkeimmistä lähdeteoksista olisi ollut, jos ei nyt

suorastaan välttämätön, niin ainakin havainnollinen ja kiinnostava. Voi kuitenkin olla, ettei kirjoittajalla ole ollut valinnan varaa. Teosta leimaa nimittäin kauttaaltaan tiivis ja virtaviivainen esitystapa, joka mahdollisesti on julkaisijan vaatimusten sanelema.

Alaluvussa 1.3. Dalmi siirtyy esittelemään kokoelman artikkelit. Myös tämä tapahtuu tiiviisti mutta esimerkillisen hyvässä järjestyksessä. Tästä alaluvusta lukijan on helppo löytää tarvitsemansa tieto kuten artikkelien aiheet sekä päätyypit (BE tai HAVE), joita tarkasteltavien kielten omistusrakenteet edustavat. Lisäksi luvussa 1.3. tehdään selkoa niistä teoreettisista ja menetelmällisistä viitekehyksistä, joista käsin kukin kirjoittaja on työskennellyt. Vaikka teoksessa tarkasteltava kielen ilmiö on yhteinen, jokainen kirjoittaja lähestyy predikoivia omistusrakenteita omalla tavallaan ja hieman eri perspektiivistä.

Teosta ei ole ryhmitelty laajemmiksi osioiksi, kuten artikkelikokoelmissa usein on tapana tehdä, vaan tutkimusartikkelit seuraavat toisiaan juoksevasti. Kokoelman neljä ensimmäistä artikkelia on kuitenkin omistettu slaavilaiskielille ja loput viisi uralilaisille kielille. Lopuksi seuraa vielä kaikkien toimittajien yhdessä kirjoittama loppukoonti *Conclusions*.

2 Artikkelit 2–5: predikoivat omistusrakenteet slaavilaisissa kielissä

Jacek Witkoś käsittelee artikkelissaan kiellon genetiiviä (GoN) puolan lokatiivisissa ja possessiivisissa eksistentiaalilauseissa. Hän hyödyntää artikkelissaan generatiivista lähestymistapaa ja käyttää kiellon genetiiviä välineenä, jolla testataan sijan korvausta (case overwriting), sijaprojektiota sekä derivationaalisia vaiheita. Artikkeli ei ole tämän kokoelman selkeimmästä päästä. Käyttöesimerkkien perusteella lukija pystyy kuitenkin päättelemään, että puolan predikoiva omistusrakenne on tyypiltään HAVE-possessiivi, vaikka tätä ei eksplikoida eikä kirjoittaja itse hyödynnä termejä HAVE-possessiivit ja BE-possessiivit.

Artikkeli alkaa johdantoluvulla 1, *Core data on the Genetive of Negation (GoN)*, jossa havainnollistetaan kiellon genetiivin (GoN) toimintaa aluksi esimerkkitapausten avulla. Esimerkkejä on runsaasti – pelkästään johdantoluvussa kaikkiaan 27 lausetta – ja ne edustavat transitiivilauseita, eksistentiaalilauseita ja omistuslauseita. Niiden avulla lukijan on mahdollisuus saada alustava ja jokseenkin selkeä kuva puolan predikoivista omistusrakenteista sekä siitä, kuinka kiellon genetiivin sääntö niihin liittyy. Tämä onkin hyvä asia, sillä kuva ei tästä ainakaan kirkastu, kun seuraavissa luvuissa 2–4 näitä rakenteita ryhdytään analysoimaan generatiivisen kieliopin menetelmin. Analyysiluvut sulkevat ulkopuolelle sellaiset lukijat, jotka eivät ole perehtyneet generatiiviseen lähestymistapaan, termistöön ja merkintäkonventioihin. Olisinkin toivonut pientä kädenojennusta esimerkiksi historiallis-vertailevia menetelmiä käyttäville tutkijoille tai kognitiivisen kielitieteen harjoittajille, joiden joukossa on myös paljon kielikontakteista ja kielitypologisesta tutkimuksesta kiinnostuneita.

Seuraavakin artikkeli käsittelee puolan kieltä. Siinä Piotr Cegłowski pureutuu omistusilmausten nominaalisiin täydennyksiin ja niiden piirteisiin sekä puolan nomini-ilmausten typologiaan yleisesti. Lähtökohtana artikkelissa on maailman kielten jako typologisesti kahteen ryhmään niiden nomini-ilmausten rakenteiden perusteella, eli ns. NP-kieliin ja DP-kieliin. Ilmauksen NP (< Noun Phrase) oletan olevan kaikille tuttu. Lyhenteellä DP viitataan puolestaan termiin *Determiner Phrase*, mikä mainitaan artikkelin lopusta löytyvässä lyhenneluettelossa. Luvussa 2 kirjoittaja esittelee aluksi NP- ja DP-jaotteluun liittyvät tärkeimmät hypoteesit, joita ovat Universal DP-Hypothesis

(DPH) ja Small Nominal Hypothesis. Esittely on hyvin lyhyt eikä DP:n käsitettä avata juuri lainkaan, kuten ei myöskään sitä, mikä on NP:n ja DP:n periaatteellinen ero. Saan lukemani perusteella kuitenkin sen käsityksen, että NP- ja DP-kielet eroavat toisistaan mm. sen suhteen, esiintyykö niissä artikkeleita vai ei, sallitaanko vasemmalle lohkeaminen (*Left Branch Extractions*) tai muunlainen nominaali-ilmauksen hajoaminen (*scrambling*) vai ei, tai esiintyykö kielessä adnominaalisia genetiivejä vai ei (ks. Bošković 2008, 2009). Asiaa ei ole ilmaistu selkeimmällä mahdollisella tavalla. Tulkitsen selonteon nyt kuitenkin niin, että niitä kieliä, joissa mainittuja rakenteita esiintyy, pidettäisiin DP-kielinä, kun taas niitä, joista ne puuttuvat, pidettäisiin NP-kielinä. Cegłowski pitääkin puolaa Boškovićin tapaan NP-kielenä, vaikka ei täysin ongelmitta. Epäselväksi itselleni kuitenkin jää, miten DP- ja NP-jaottelu liittyy puolan predikoiviin omistusrakenteisiin, joista kirjassa pitäisi kuitenkin olla kyse. Artikkelissa käytetyt esimerkkilauseet edustavat nimittäin muitakin lausetyyppejä kuin predikoivia omistusrakenteita. Samoin kyselytutkimuksessa, jonka avulla erilaisten nominaalisten lausekkeiden hyväksyttävyyttä testattiin, hyödynnettiin paitsi omistus- ja eksistentiaalilauseita myös transitiivilauseita, eikä omistusrakenteiden asema vaikuta millään tavoin korosteiselta.

Molemmat puolaa käsittelevät artikkelit ovat vaativaa luettavaa. Kummastakin näkyy kirjoittajan syvällinen asiantuntemus sekä teorian ja menetelmien hallinta. Niille lukijoille, jotka eivät ole perehtyneet juuri näissä artikkeleissa käytettyihin menetelmiin ja esitystapoihin, on kuitenkin haasteellista päästä sisälle kirjoittajan ajatusmaailmaan ja ottaa kantaa artikkeleissa esitettyihin väitteisiin tai päätelmiin. Käyttöesimerkkejä annetaan kuitenkin runsaasti. Ne ovat kiinnostavia ja selkeitä, ja myös niiden glossaus on toteutettu huolellisesti. Esimerkkien avulla tarkasteltavina olevista rakenteista ja niiden ominaisuuksista saa varsin hyvän käsityksen. Artikkelit luettuaan lukija on todennäköisesti oppinut jotain puolan kielestä yleisesti.

Kokoelman seuraavaa artikkelia ei sen sijaan voi moittia vaikeaselkoisuudesta. Olga Kaganin tutkimuksen aiheena on määräisyysvaikutus (Definiteness Effect) venäjän kielessä sekä ns. kiellon genetiivi, joka Kaganin mukaan selittäisi, miksi venäjän eksistentiaali- ja omistuslauseet eivät aina noudata määräisyysvaikutuksen ”sääntöä”. Aluksi Kagan esittelee erittäin lukijaystävällisesti, mistä määräisyysvaikutuksessa on kyse: useissa maailman kielissä vaikuttaa syntaktis-semanttinen sääntö tai periaate, jonka mukaan eksistentiaalilauseen subjekti voi saada vain ns. heikkoja määritteitä tai tarkennuksia. Näihin kuuluvat mm. lukumäärän ilmaukset (*There are **five chairs** in the room*), epämääräiset artikkelit sekä indefiniittiset kvanttorit (*There is **a chair** in the room; There are **some chairs** in the room*). Pois suljettuja ovat sen sijaan ns. vahvat määritteet kuten määräiset artikkelit (**There is **the chair** in the room*) ja määräiset kvanttori-ilmaukset (**There are **most chairs** / **both chairs** in the room*). Vahvoiksi lausekkeiksi katsotaan myös erisnimet, jotka näin ollen eivät sovi eksistentiaalilauseen subjektiksi: **There is John in the room*.

Kaganin esittelyn lähtökohtana on englannin kielestä tehty tutkimus (ks. Milsark 1979), ja vaikka Kagan ei väitäkään määräisyysvaikutusta kieliuniversaaliksi, hän ymmärtää sen kuitenkin laajasti vaikuttavaksi säännöksi, jonka tulisi toimia myös venäjässä. Tilanteita, joissa siitä poiketaan, hän pitää ongelmallisina säännön rikkomuksina (*violations*) ja katsoo niiden vaativan perusteluja. Erityisesti venäjän kielteisissä eksistentiaalilauseissa subjekti voi odotuksenvastaisesti saada myös vahvoja määritteitä tai olla erisnimi. Selitykseksi tähän Kagan tarjoaa mm. kiellon genetiiviä, jonka vaikutuksesta vahvat lausekkeet siirtyisivät merkitykseltään lähemmäksi ominaisuuden ilmauksia. Vahvatkin määritteet ovat nimittäin sallittuja eksistentiaalilauseissa, mikäli ne ilmaisevat ominaisuutta.

Lukijaa jää vaivaamaan, miksi määräisyysvaikutuksesta on väkisin tehty lähes kieliuniversaaliin rinnastettava sääntö, josta poikkeaminen vaatii erityisen selityksen. Olisiko venäjän eksistentiaali- ja omistuslauseista syntynyt monipuolisempi ja vivahteikkaampi kuva, jos määräisyysvaikutusta olisi tarkasteltu ilmiölähtöisemmin ja hyödynnetty esimerkiksi prototyypin käsitettä? Suomalaisen lukijan on varsin helppo keksiä esimerkkejä eksistentiaalilauseista, joissa esiintyy vahva tai määräinen subjekti. Vaikka *Huoneessa on juuri nämä viisi tuolia* ei kenties ole prototyyppinen eksistentiaalilause, on se silti yhtä lailla mahdollinen kuin *Huoneessa on paljon tuoleja*. Voisiko myös venäjän eksistentiaalilauseissa esiintyä vastaavanlaista määräisyyden ja epämääräisyyden kirjoa ja vaihtelua? Olisiko tätä vaihtelua voinut saada näkyväksi esim. korpuspohjaisella tutkimuksella?

Kaganin ansioksi on kuitenkin luettava, että hänen ajatuksenjuoksuaan pystyy vaivatta seuraamaan. Lukijalla on siis mahdollisuus olla lukemastaan myös eri mieltä, koska teksti on kaikilta osin ymmärrettävää, selkeää ja havainnollista.

Kokoelman neljäs slaavilaiskieliä käsittelevä artikkeli vie lukijan jälleen formalististen ja teoreettisten esitystapojen sokkeloihin. Egor Tsedrykin artikkelin aihe on tosin äärimmäisen mielenkiintoinen: kun predikoivat omistusrakenteet maailman eri kielissä edustavat yleensä joko BE-possessiiveja tai HAVE-possessiiveja, valkovenäjässä käytetään kumpaakin tyyppiä rinnakkain. Lisäksi BE-tyyppisistä omistusrakenteista voidaan erottaa eksistentiaaliset ja kopulatiiviset BE-possessiivit. Näiden erityyppisten omistusrakenteiden vaihtelua ja keskinäistä suhdetta Tsedryk tarkastelee – sikäli kuin ymmärsin – formaalin semantiikan keinoin. Teoreettisena viitekehyksenään Tsedryk mainitsee *Distributed Morphology* -menetelmän (Halle & Marantz 1993), ja lähestymistapaansa hän kuvailee pääosin derivationaaliseksi. Hän siis pyrkii selvittämään, mikä on se yhteinen semanttinen juuri, johon valkovenäjän HAVE- ja BE-possessiivit palautuisivat.

Artikkelissa esitellään ansiokkaasti valkovenäjän erityyppisiä omistusrakenteita ja pohditaan mm., millaisia omistussuhteita niillä ilmaistaan tai mitä verbien aikamuotoja kunkin yhteydessä preferoidaan. Myös määräisyysvaikutukselle (DE) ja kiellon genetiiville valkovenäjässä on omat alalukunsa 2.4 ja 2.5. Pidän tätä deskriptiivistä lukua 2 artikkelin parhaana antina. Kiinnostavasti valitut ja huolellisesti glossatut esimerkit läpikäytyään lukija voi todella sanoa oppineensa jotain uutta valkovenäjästä. Lauserakenteiden formaalit kuvaukset erikoisine merkintätapoineen (luvuissa 3–5) eivät sen sijaan vakuuta yhtä lailla, kuten ei myöskään loppupäätelmä (luvussa 6), jonka mukaan HAVE-possessiivit sekä eksistentiaaliset ja kopulatiiviset BE-possessiivit palautuisivat yhteiseen spatio-temporaaliseen juureen √AT, jota edustaisi myös venäjälle ja valkovenäjälle yhteinen prepositio *u* ’at’. Yhteinen juuri ei kuitenkaan vaikuta olevan tämä prepositio eikä mikään muukaan konkreettinen kielenaines vaan generatiivisen kielitieteen olettama abstrakti syvärakenne.

Miksi näin monimutkaista selitysmallia on lähdetty tavoittelemaan, kun kielisukulaisuuteen ja kielikontakteihin perustuva selitys olisi huomattavasti yksinkertaisempi? Venäjässähän käytetään BE-possessiiveja, kun taas puolassa HAVE-possessiivit ovat käytössä. Eikö olisi luonnollista, että valkovenäjässä on piirteitä kummastakin naapurikielestä, jotka lisäksi ovat sen läheisiä sukukieliä? Tsedryk toki myöntää, että historiallinen selitys HAVE/BE-possessiivien vaihtelulle on olemassa. Samaan hengenvetoon hän kuitenkin toteaa: “*Nevertheless, a language-internal explanation is also expected*.” Tahoa, joka tällaista selitystä vaatii, ei Tsedryk

nimeä tarkemmin, mutta koko kielitieteen kenttää se tuskin edustaa. Moni historiallis-vertailevaan kielitieteeseen, kielitypologiaan, historialliseen syntaksiin tai vanhoihin kirjakieliin suuntautunut lukija olisi mielihyvin tyytynyt historialliseen kuvaukseen valkovenäjän HAVE- ja BE-possessiivien kehitysvaiheista ja niiden yhteisestä historiasta venäjän ja puolan vastaavien rakenteiden kanssa.

3 Artikkelit 6–10: predikoivat omistusrakenteet uralilaisissa kielissä

Uralilaisia kieliä käsittelevistä artikkeleista ensimmäinen on Maria Vilkunan käsialaa. Siinä käsitellään ennen kaikkea suomen omistuslausetta eksistentiaalilauseen alatyyppinä, mutta myös muita omistuksen, paikallisuuden tai olemassaolon ilmauksen rakenteita esitellään. Lisäksi pohditaan näiden eri lausehahmojen välisiä sukulaisuussuhteita ja tehdään näkyväksi niissä ilmenevää variaatiota.

Vilkuna on pitkän tutkijauransa aikana hyödyntänyt useita erilaisia teoreettisia viitekehyksiä, hirttäytymättä silti yhteenkään niistä. Tässä artikkelissa hän tukeutuu konstruktiokielioppiin. Esitystapansa puolesta artikkeli on kuitenkin ennen kaikkea deskriptiivinen: eksistentiaalisten ja omituslauseiden kirjoa esitellään lukijalle sekä aineistopohjaisten että konstruoitujen esimerkkien avulla, aina prototyyppisemmistä tapauksista erikoisempiin edeten. Selkeydessään ja lukijaystävällisyydessään Vilkunan artikkeli on tässä kokoelmassa selkeästi ylitse muiden. Niitäkään lukijoita, joille konstruktiokielioppi ei ole entuudestaan tuttu, ei pudoteta kelkasta, mutta ei myöskään aliarvioida tai tukehduteta ylenmääräisellä terminologian läpikäymisellä. Itse asiassa tutkimusalan omaa käsitteistöä olisi ehkä voinut kuljettaa mukana hieman rohkeamminkin ja voittaa näin konstruktiokieliopille uusia ystäviä.

Kokoelman seitsemännessä artikkelissa Gréte Dalmi esittelee unkarin BE-possessiiveja ja vertailee niitä kopulatiivisiin BE-rakenteisiin. Lähestymistapana on unkarin BE-verbin argumenttirakenne, joka on erilainen kopulatiivisissa, eksistentiaalisissa ja omistusta ilmaisevissa rakenteissa, vaikka itse verbinä toimii periaatteessa sama lekseemi. Tarkastelun lopputulema osoittaa Dalmin mukaan, että unkarin possesiivinen ja eksistentiaalinen BE-verbi ovat kumpikin 2-paikkaisia intransitiivisia predikaatteja (*dyadic unaccusative predicates*) eivätkä 1-paikkaisia (*monadic unaccusatives*), kuten aikaisemmin on otaksuttu. Eksistentiaalinen BE saa täydennykseksi lokatiivisen argumentin ja teeman (*Location and Theme*), kun taas possessiivinen BE täydentyy obliikvimuotoisella omistajalla ja teemalla (*oblique Possessor and Theme*).

Dalmin artikkelin tekee jossain määrin vaikeaselkoiseksi se, ettei unkarin erilaisia lauserakenteita verrata pelkästään keskenään, vaan niitä verrataan myös vastaavanlaisiin rakenteisiin lukuisissa muissa kielissä. Luvussa 1, jossa ilmiötä esitellään, valtaosa käyttöesimerkeistä onkin itse asiassa peräisin venäjästä, ja lukijan pitää olla tarkkana, jotta pystyy seuraamaan, mistä kielestä kulloinkin on puhe. Jopa luvun 2 (BE-possessives in Hungarian) avausesimerkit ovat ranskaa, ja vasta alaluvussa 2.2. unkarin kieli alkaa olla selkeästi pääosassa. Lukijalta, joka haluaa tietää, miltä unkarin *olla*-verbi oikeasti näyttää, vaaditaan salapoliisintyötä, sillä siitä puhutaan jatkuvasti vain possessiivisena tai eksistentiaalisena BE-verbinä. Yleiskielitieteellisestä ja vertailevasta näkökulmasta tällainen merkintätapa voi olla perusteltu ja hyväkin. Olisin silti toivonut artikkeliin edes lyhyttä alalukua, jossa olisi keskitytty kuvaamaan nimenomaan unkarin *olla*-verbiä vertaamatta sitä jatkuvasti muiden kielten vastaavaan verbiin.

Alexandra Simonenkon artikkeli *Existential possession in Meadow Mari* on kiinnostava katsaus niittymarin omistuslauseisiin, joita kirjoittajan mukaan on kahta eri tyyppiä, 1) eksistentiaalinen ja 2) predikoiva.

1)	*myj-yn*	*aka-m*	*ulo.*
	I-GEN	sister-POSS.1SG	be.PRES.3SG
	'I have a sister'		

2 a)	*tide*	*pört*	*myj-yn*
	that	house	I-GEN
	'That house is mine.'		

2 b)	*tide*	*pört*	*myj-yn*	*yle.*
	that	house	I-GEN	be.PST.3SG
	'That house was mine.'			

Predikoivaa rakennetta voitaisiin nimittää myös kopulatiiviseksi, kuten muissakin kirjan artikkeleissa tehdään. Yhteistä kummallekin lausetyypille on, että omistajaa ilmaiseva lauseke on genetiivimuotoinen. Kummassakin verbinä esiintyy olla-verbi, mutta preesensmuotoinen predikoiva (tai kopulatiivinen) omistuslause on verbitön. Niittymarin omistuslauseet edustavat siis tyyppiä BE-*possessives*. Simonenko ei kuitenkaan tätä termiä artikkelissaan käytä, mikä hieman harmittaa. Yhdenmukaiset nimitykset olisivat lisänneet artikkelikokoelman sisäistä koherenssia ja auttaneet lukijaa vertailemaan saman ilmiön edustumia eri kielissä. Simonenko tarkastelee artikkelissaan eksistentiaalisten ja kopulatiivisten omistusrakenteiden välisiä yhtäläisyyksiä ja eroja mutta myös omistusliitteiden roolia ja merkitystä sekä määräisyysvaikutusta (DE), joka Simonenkon mukaan vaikuttaa myös niittymarin eksistentiaalilauseissa. Tutkimus on korpuspohjainen, mutta korpusaineistosta nousseita havaintoja on täydennetty myös elisitaatiomenetelmin.

Artikkeli sisältää paljon hyvää ja selkeää kielen kuvausta, ja erityisesti omistusliitteiden merkitykseen ja käyttöehtoihin liittyvät havainnot luvussa 2 ovat äärimmäisen mielenkiintoisia, samoin kuin havainnot omistusliitteiden ja määräisyysvaikutuksen yhteispelistä. Lukujen 3–4 formaalit analyysit, joiden tarkoituksena on näitä havaintoja perustella ja selittää, ovat kuitenkin mustia laatikoita niille, jotka eivät tunne käytettyjä teorioita ja menetelmiä. Puukuvaimien voi toki olettaa olevan tuttuja kaikille syntaksin tutkimukseen perehtyneille, mutta kreikkalaisten kirjainsymbolien merkitykset eivät välttämättä ole universaaleja vaan riippuvat kontekstista, jossa niitä käytetään. Matemaattisten symbolien kuten kuin ∀ tai ∃ ymmärtämiseksi lukija tarvitsisi ehdottomasti lähdeteoksen, sillä artikkelissa niitä ei selitetä millään tavoin. Analyysiluvut tekevät toki vaikutuksen, mutta eivät vakuuta. On varmasti vaatinut paljon teoreettista asiantuntemusta ja ajatustyötä tuottaa lukujen 3–4 formaalit kuvaukset, joita en tässä edes yritä lainata. Itse niittymarin omistusrakenteet, joiden toimintaa olisi tarkoitus havainnollistaa, hukkuvat kuitenkin kreikkalaisten ja muiden kirjainsymbolien, matemaattisten operaattoreiden, ylä- ja alaindeksien, moninkertaisten kaari- ja hakasulkeiden ja muiden arvoituksellisten merkintöjen sekaan. Onneksi lukija voi kuitenkin palata lukuihin 1 ja 2, jos haluaa havainnollisempaa tietoa niittymarista ja sen omistusrakenteista.

Nikolett F. Gulyás lähestyy omaa tutkimusaihettaan perinteisemmästä vertailevasta ja kielitypologisesta näkökulmasta. Hänen tutkimuksensa aiheena on predikoiva possessiivisuus permiläiskielissä. Esimerkkikielinä ovat komipermjakki ja udmurtti. Gulyás kuvailee ja vertailee keskenään predikoivia omistusrakenteita komipermjakissa ja udmurtissa ja kiinnittää huomiota niiden semanttisiin ominaisuuksiin kuten omistussuhteisiin, joita niiden avulla voidaan ilmaista. Myös syntaktisia ominaisuuksia kuten sanajärjestystä, kongruenssia, omistusliitteiden käyttöä ja kieltolauseiden erityispiirteitä käsitellään.

Gulyásin artikkelista puuttuu se järeä teoreettinen koneisto, jolla useissa muissa tämän kirjan artikkeleissa operoidaan. Totuuden nimessä, en kuitenkaan voi sanoa kaipaavani sellaista, vaan pidän artikkelin deskriptiivistä lähestymistapaa varsin onnistuneena. Ainoa asia, josta huomauttaisin, on merkintätapojen puutteellisuus: kun vertaillaan kahta tai useampaa kieltä, joita kaikki lukijat eivät osaa, olisi tärkeää merkitä jokaisen esimerkin kohdalle selkeästi, esim. kirjainkoodeja käyttäen, mitä kieltä esimerkki edustaa. Nyt osaan esimerkeistä oli selkeästi merkitty, oliko kyseessä komipermjakki, udmurtti vai jokin muu kieli. Välillä tieto piti taas etsiä tai ainakin varmistaa leipätekstistä. Toki tieto kielestä lopulta löytyi, mutta olisi ollut yksinkertainen kädenojennus lukijalle merkitä kaikki komipermjakin kieliset johdonmukaisesti vaikkapa K-P:lla ja udmurtinkieliset U:lla. Ne lukijat, jotka kyseisiä kieliä hyvin osaavat, tunnistavat varmasti oikopäätä, kumpaa niistä esimerkit edustavat. Selkeät merkintätavat palvelisivat kuitenkin niitä, joille komipermjakki ja udmurtti ovat uusia tuttavuuksia. Sain myös lukemani pohjalta sen käsityksen, että udmurtin ja komipermjakin omistuslauseet kuuluvat BE-possessiiveihin. Tätä nimitystä olisi ollut hyvä hyödyntää johdonmukaisesti.

Kirjan viimeinen tutkimusartikkeli käsittelee predikoivia omistusrakenteita selkupin murteissa. Beáta Wagner-Nagy esittelee aluksi selkupin kielen ja tekee selkoa sen asemasta uralilaisen kielikunnan jäsenenä. Kuten Wagner-Nagy aivan oikein toteaa, selkuppi ei ole suomalais-ugrilainen kieli vaan samojedikieli. On harmillista, ettei tätä tietoa ole hyödynnetty artikkelikokoelman toimitusvaiheessa eikä teoksen alaotsikkoa ole korjattu asianmukaisempaan muotoon: *The view from Slavic and Uralic.*

Alkuesittelyn jälkeen Wagner-Nagy esittelee lyhyesti tutkimusmenetelmänsä – kyse on korpuspohjaisesta tutkimuksesta – ja siirtyy sitten kuvailemaan selkupin predikoivia omistusrakenteita ja niiden syntaktisia ja semanttisia piirteitä. Artikkelin lopputulema on, että selkupin murteissa esiintyy kaksi pääskeemaa, joiden avulla omistusta ilmaistaan, eli topikaalinen ja lokatiivinen skeema. Topikaalisessa skeemassa lauseen kieliopillinen subjekti on lauseenalkuisessa asemassa, ilmaisee omistajaa ja on nominatiivissa. Myös omistettava on nominatiivimuotoinen, mutta se esiintyy subjektin jäljessä ja voi saada omistusliitteen. Mikäli omistettava on elollinen, omistusliite on pakollinen. Lokatiivinen skeema koostuu puolestaan lauseenalkuisesta omistajan ilmauksesta, joka on paikallissijainen tai paikallisuutta ilmaiseva postpositiolauseke. Omistettava on nominatiivimuotoinen, mutta usein omistusliitteellä merkitty. Mikäli omistettava on ruumiinosa, sukulaistermi tai muu erottamaton omistus, on omistusliite pakollinen. Kummassakin skeemassa verbinä on *ɛ:qo* ’olla’, joka taipuu normaalisti lokatiivisessa skeemassa mutta topikaalisessa skeemassa kongruoi omistettavan kanssa. Näiden kahden pääskeeman lisäksi tunnetaan genetiiviskeema, joka kuitenkin on harvinainen.

Wagner-Nagyn kuvauksen perusteella päättelen, että myös selkupin predikoivat omistusrakenteet edustavat BE-possessiiveja. Nimitystä olisi ollut suotavaa käyttää tässäkin artikkelissa. Kuvaus on kuitenkin selkeä, ja vaikka kysymys on näinkin etäisestä sukukielestä, suomenkieliselle lukijalle esimerkit tuottivat monia iloisia tunnistamisen ja oivalluksen hetkiä.

4 Lopuksi

Approaches to Predicative Possessions oli kiinnostava, vaikka paikoin hämmentäväkin lukukokemus. Otin kirjan innokkaasti arvioitavaksi sen otsikon perusteella: kuuluvathan omistusrakenteet ja niiden pohjalta kehittyneet konstruktiot omiin tutkimusintresseihini. Teoksen luettuani en kuitenkaan ole enää varma, ottaisinko sen uudelleen arvioitavaksi, jos sitä minulle tarjottaisiin. Toki kirjan kannen yläreunassa näkyy selvästi sarjan nimi: BLOOMSBURY STUDIES IN THEORETICAL LINGUISTICS, ja jo tästä olisin voinut päätellä, että joku toinen olisi sopinut tehtävään paremmin. Varoituksen sanat oli kuitenkin painettu pienin, haalein ja hopeanharmain kirjaimin, ja huomasin ne vasta, kun oli liian myöhäistä ja olin lukenut artikkeleista noin puolet. Oma lähestymistapani kieleen ja sen tutkimukseen on ennen kaikkea deskriptiivinen ja historiallinen, joten en todennäköisesti ole tavoittanut kaikkea, mitä kirjoittajat ovat halunneet artikkeleissaan sanoa.

Teos on kokonaisuutena ilman muuta tutustumisen arvoinen ja hyvä tietolähde kaikille omistusrakenteista kiinnostuneille. Kaikissa teoksen artikkeleissa on paljon kiinnostavaa asiaa. Jokainen artikkeli antaa osaltaan yksityiskohtaista ja tarkkaa tietoa siitä kielestä, jota ollaan kuvaamassa. Aivan kaikissa artikkeleissa eivät predikoivat omistusrakenteet tosin ole päärooliss, jos tarkkoja ollaan. Esimerkiksi Geglowskin artikkeli ainoastaan sivuaa niitä, mutta auttaa silti osaltaan rakentamaan kokonaiskuvaa predikoivien ilmausten piirteistä slaavilaisissa kielissä.

Kirjan parhaana antina pidin artikkelien deskriptiivistä osuutta sekä hyvin valittuja ja huolellisesti glossattuja esimerkkejä! Toivomisen varaa sen sijaan jätti termistön ja käsitteistön viimeistely. Toimitusvaiheessa olisi ollut hyvä varmistaa, että esimerkiksi johdantoartikkelissa esiteltyjä nimityksiä HAVE-possessiivit ja BE-possessiivit käytetään johdonmukaisesti kaikissa muissakin teoksen artikkeleissa. Erityisen paljon jäi harmittamaan, että teoksen alaotsikkoon oli virheellisesti jäänyt ”Finno-Ugric”, kun sen varsin helposti olisi voinut korjata muotoon ”Uralic”.

Lähteet

Benveniste, E. 1966. Être et avoir dans leurs fonctions linguistiques. Teoksessa Benveniste, E. (toim.), *Problèmes de linguistique générale 1*, 187–207. Paris: Gallimard.

Bošković, Ž. 2008. What will you have, DP or NP? *Proceedings of the North East Linguistic society* 37. 101–114.

Bošković, Ž. 2009. More on the no-DP analysis of article-less languages. *Studia Linguistica* 63. 187–203.

Halle, M. & Marantz, A. 1993. Distributed morphology and the pieces of inflection. Teoksessa Hale, K. & Keysers, S. J. (toim.), *The view from Building 20: Essays in Linguistics in Honor of Sylvain Bromberger*, 111–176. Cambridge, MA: MIT Press.

Freeze, J. 1992. Existentials and other locatives. *Language* 68(3). 553–595.

Kayne, R. 1993. Towards a theory of modular auxiliary selection. *Studia Linguistica* 47. 3–31.

Milsark, G. 1979. *Existential Sentences in English*. London: Routledge.

Stassen, L. 2009. *Predicative possession*. Cambridge: Cambridge University Press.

Yhteystiedot:

Maria Kok
Itä-Suomen yliopisto
maria.kok@uef.fi

Jaakola, Minna & Onikki-Rantajääskö, Tiina (eds.). 2023. *The Finnish case system: Cognitive linguistic perspectives*. (Studia Fennica Linguistica 23). Helsinki: Finnish Literature Society. Pp. 388. https://doi.org/10.21435/sflin.23

Reviewed by Max Wahlström

1 An ambitious volume on the Finnish case system

The Finnish case system: Cognitive linguistic perspectives is a collective volume dealing with the Finnish case system. The systematic coverage of key areas of the case system in the book suggests it may have potential as a state-of-the-art English language descriptive reference work – this question will be kept in mind throughout the review.

The subtitle *Cognitive linguistic perspectives* implies a shared theoretical basis for the discussions. To contextualize this volume is yet somewhat difficult. Cognitive Linguistics has its origins in a U.S.-centered critical reaction to generative grammar. The two influential strains of Cognitive Linguistics, Langackerian Cognitive Grammar (CG) and Construction Grammar (CxG) both figure in this book, with the majority of articles utilizing Langacker's concepts. Nevertheless, the terminology used in the volume to refer to the morphology and syntax of Finnish does not depart from the mainstream descriptive practice as showcased, for instance, by the pre-eminent journal in the study of Finnish, *Virittäjä*. Nothing in the choice of topics either betrays a particular Cognitive Linguistic agenda or focus. In acknowledgement of this duality, I will mostly hold my assessments of the role of Cognitive Linguistics until the end of this review.

Before moving to the articles of the book, I wish to lay out some considerations for an expert reader who is however not familiar with the idiosyncrasies of the Finnish case system and its study. Finnish is sometimes described as a fairly agglutinative language, with a moderate amount of fused morphemes and relatively little allomorphy (for Finnish in this regard, but also criticism of the Agglutination Hypothesis, see Haspelmath 2009). However, the current case suffixes have developed over a long period of time, and they are sometimes of very different ages. A single nominal often requires different stems to host case morphemes, which is a major factor leading to complexity on morpheme boundaries.

The traditional number of 14 or 15 cases means that not all of them pertain to the marking of core grammatical roles. Yet any attempt to divide the cases into core and adverbial is problematic. Of the "adverbial" local cases the allative encodes the recipient of ditransitives, the elative the complements of verbs of liking, and the adessive the possessor in default predicative possession. However, what is sometimes taken as the fifteenth case, abessive, derives, in fact, an adverb from nouns (*autoi-tta* '[doing something] without cars') and a noun in abessive cannot productively head a noun phrase (NP). The introduction to the book does briefly discuss the inflection–derivation question and notes that in the tradition of the grammatical description of Finnish, inflection and derivation have been seen to form a continuum (p. 26–27).

Another reason why a predominantly derivative morpheme is thought of in terms of case is that almost all non-finite verb forms are morphosyntactically on the nominal spectrum and may be inflected for a few or more forms, which are polysemous with cases – these include for instance the abessive, giving these two uses a significantly higher

frequency together than the abessive as a mere derivative marker. Three non-finite stems called infinitives accept only a limited number of the case-like markers, cannot head an NP, and are clearly semantically verbs sometimes being able to encode arguments with possessive suffixes or genitive NPs. Yet the suffixes of these verbs are often equated with nominal cases (e.g., Example 1e, p. 59; see, however, also p. 153 acknowledging some of these complexities).

Additionally, there are two descriptive traditions differing in whether case should be treated as a syntactic or a morphological category. Therefore "accusative" can alternatively encompass two differential object marking phenomena involving three morphological cases or it can designate a single morphological case. This book predominantly gravitates toward a syntactically based nomenclature (see, especially, Jaakola in the volume), arguing that how cases are defined is not a mere didactic choice.

The book is structured as follows: The introduction to the book, authored by Tuomas Huumo and the two editors of the volume, Minna Jaakola and Tiina Onikki-Rantajääskö, gives a short overview of the Finnish case and introduces the Cognitive linguistic perspective on case. The first part of the book "Cases and core arguments" has two chapters dealing with the partitive and the genitive, respectively. Altogether five chapters make up the second part, entitled "Adverbial cases," and the third part consists of a variety of well-motivated studies, yet slightly more tangential to case. This last part is named "Cases and related phenomena."

2 Core arguments

Tuomas Huumo's chapter on the partitive begins with a well-thought-out and concise presentation of the functions of the partitive as a verbal argument. Huumo argues for four interrelated semantic functions of the partitive: quantification, indication of mass status, aspect, and negation. While this presentation no doubt summarizes two decades of research into the topic, the thought-provoking analyses are nevertheless highly accessible. Only on page 55, I cannot follow his argumentation: In the context of partitive marked S-arguments of the existential clause, Huumo discusses mass nouns that refer to a kind (*milk is good for you*) and states that these types of nouns cannot take the partitive, only the nominative. I take issue with an example he uses: *maito*.NOM *on hyväksi sinulle* 'milk is good for you'. This is not an existential expression, it requires verbal agreement, and it does not allow partitive subjects (cf. Vilkuna et al. 2008: § 893), no matter what the referential scope of the noun as S-argument is or whether it is a count or a mass noun. I agree that kind-referring mass nouns cannot be used as the partitive S-arguments of the existential clause. Yet this has nothing to do with their mass-denoting character: no kind-referring generic expression can be used in existential constructions in the first place, a limitation known, among other, as the definiteness restriction of existential clauses (Milsark 1977: 45). This in mind, Huumo's conclusion that "q-partitives", the partitive marked S-arguments of existential clauses indicate indefiniteness, seems perhaps unsurprising.

Minna Jaakola's chapter makes an intriguing attempt to operationalize the Cognitive Grammar (CG) concept of reference-point construction regarding the adnominal use of the genitive. My summary of what the concept entails is that, allegedly, humans perceive things in binary terms: attention is focused on something familiar, accessible, or recognizable that then in turn helps to define a less familiar entity. This "access point" is termed a reference-point, and, in the case of the adnominal genitive, the genitive marked NP is considered the reference-point, and a concrete linguistic element called

the landmark, contrasting with the head, called the trajector. The reference-point model therefore establishes a structurally stable organization across the grammar, not sensitive to, say, information-structural or semantically motivated considerations. This allows Jaakola to contrast the referents of the proposed landmarks and trajectors in a corpus, and these are classified on a referential hierarchy from inanimates through institutions to humans.

The analysis reveals that the type of referents as genitive modifiers and heads is highly genre-dependent, but the big picture is that humans appear more often as genitive modifiers than heads. After this, Jaakola evaluates the discourse salience of the referents both through assigning them a givenness status and by tracking cataphoric reference. I feel that the results should have been statistically verified, especially in contrast to the type of referents. Statistical testing could reveal otherwise hidden dependencies and perhaps more detailed observations, although I do not doubt the author's broader conclusions: most of the genitive modifiers are identifiable and reference continuity is more often carried by the genitive modifier than the head. I will address the theoretical implications of this analysis later in the review, but a note about the remainder of the chapter: Jaakola assumes that the diachrony of the genitive S-arguments of the necessive clauses is irrefutably established in Inaba (2015). The study is groundbreaking in its material and methodological depth and sets a very high bar for any other study on the topic, but in my view, several key questions still remain open (see Pantermöller 2016).

3 Adverbial cases

Tiina Onikki-Rantajääskö's chapter on the local cases is a clear candidate for an excellent up-to-date description with an extensive bibliography for anybody interested in, for instance, the lexical typological profile of Finnish constructions originating in local expressions. It is an accessible yet detailed account of the local cases with both their more grammatical uses, on the one hand, and abstract, on the other. The article also offers an opportunity to observe another interpretation of the conceptual pair landmark/trajector in action. It seems that the terms are used in the chapter to mechanically name the subject-like arguments as trajectors and the adverbial locative expressions as landmarks. This same use is adopted also in Ojutkangas toward the end of the book. When moving toward more abstract uses of the local cases, one sees fewer references to the landmark. I wonder how the author would treat subject-like uses of elative in expressing the experiencer (e.g. *minu-sta*.ELA *tuntuu hyvä-ltä*.ABL 'I feel good'), not addressed in this article.

The chapter by Eero Voutilainen questions an interpretation laid out in the introduction: The authors of the introduction repurpose an old category of "general local cases" that has been used to represent a set of three historical local cases, essive (location), partitive (source), and translative (goal) (Hakulinen 1979: 100–102). In the modern language, none of these cases mark local relations productively and they have been replaced in these roles by the external and internal sets of local cases. Jaakola, Onikki-Rantajääskö, and Huumo leave the partitive out of their "general" local cases and present essive and translative in terms of local cases. They choose among the current uses *opettaja-na*.ESS 'as a teacher' and *opettaja-ksi*.TRANSL '[become] a teacher'.[1] They argue that the translative represents a metaphorical "goal".

Voutilainen convincingly argues that the uses of the translative can be divided into 1) expressions of actual change (see the previous example), 2) expressions of "fictive" change, as in expressing things turning out to be something, and 3) other closely related

phenomena. Voutilainen himself contrasts these change-centered analyses with the "localist" interpretation offered in the introduction. However, Voutilainen overcomes the seeming discrepancy between these analyses by, in fact, subsuming the traditional six local cases under another division that encompasses also the proposed "general" local cases. Voutilainen categorizes both the essive and the two local cases inessive and adessive as static cases, whereas the rest of the local cases and the translative are grouped as dynamic cases.

Emmi Hynönen's chapter deals with the essive case, which in its primary function expresses non-permanent states such as roles and properties. Unlike several other articles in the volume that revolve around Langacker as their theoretical center of gravity, Hynönen also leans on Laura Janda. Janda's idea, applied here by Hynönen, is to examine those uses of the case that overlap or border that of other cases. The author finds several such instances and, in addition, more complex competing constructions.

Maija Belliard's article reports a data-driven take on the comitative case. Addressing the functions of the case from the perspective of corpus data proves to be fruitful. The findings challenge two common assumptions about the comitative, namely, that the comitative would be most frequently used in its prototypical function expressing accompaniment by a human referent and that the postposition *kanssa* 'with', thought to compete within the same semantic domain, would soon replace the comitative. Belliard shows both that the non-prototypical uses of the comitative outnumber the prototypical and that the uses of *kanssa* only partially overlap with those of the comitative, and, due to the separate domains, comitative is not threatened by extinction.

This part of the book is concluded by Auroora Vihervalli's and Tiina Onikki-Rantajääskö's chapter on the abessive. The paper is in an important contribution to the discussion about the more marginal cases, as it discusses several key characteristics that distinguish ablative from the full-fledged cases. Interestingly, for the authors, productivity seems to be a key argument for case-hood, but its function as deriving adverbs or inability to head an NP are not considered arguments against it. Based on their online discussion data, the authors challenge some previous claims about abessive's unlimited productivity. They find that abessive is overwhelmingly used in specialized and even lexicalized meanings. Yet, like Belliard, the authors do not predict the demise of their case any time soon.

4 Related phenomena

The final part of the book contains four more contributions to the overall discussion about cases. First, Mari Siiroinen's chapter offers a refreshing take on a marginal construction of Finnish expressing change-of-state, *puuro*.NOM *tuli sakea-a*.PART 'the porrige turned thick'. Unlike other articles in the volume, Siiroinen uses dialectal and historical data but discusses the construction also in the context of other Finnic languages. This variety of data allows the author to convincingly demonstrate that the construction deemed marginal in some descriptions was widespread in certain dialects not that long time ago.

The following article by Krista Ojutkangas discusses the co-occurrence of multiple dynamic local cases (in the sense of Voutilainen, same volume). Ojutkangas presents some typological claims about the distribution of source and goal expressions and sets

[1] The table includes also elative in parenthesis 'from [being] a teacher'; a tiny note: the cells for translative and elative have switched places on p. 23.

out to assess these. Yet there is not enough elaboration of these claims to see whether the author operationalizes them in a meaningful way. Some of the results are potentially valuable, but their linguistic motivation remains likewise unclear.

The chapter by Minna Jaakola and Krista Ojutkangas presents interesting details about the system of postpositions in Finnish. Their claim that postpositions are an open class that accepts new expressions without them having to go through additional steps of grammaticalization is intriguing. My intuition is that parallel phenomena with partially open construction types can be found in other European languages that do not have postpositions, but this requires more research.

The closing chapter of the book by Anni Jääskeläinen analyses the *-sti* suffix that derives adverbs mainly from adjectives. In brief, in Finnish, there are a handful of derivative suffixes that forge adverbs, among which *-sti* is the most productive deriving adverbs of manner. In addition to adjectives, the suffix attaches to cardinal numbers forming adverbs of absolute frequency (*kahde-sti* 'twice') and, among nouns, to curse words (*se sattui saatana-sti* 'it hurt like hell'). Since the function of *-sti* seems clearcut, it is perhaps surprising that the investigation is centered around the question of whether as a morpheme attaching to a nominal, *-sti* should be thought of in terms of a nominal case or a derivative suffix. However, as we have seen in this volume, in the descriptive tradition of Finnish, inflection and derivation have been thought of in terms of a continuum.

Jääskeläinen gives the formation of adverbs from comparatives and superlatives and adverbs of absolute frequency as a "strong" argument in favor of the case-like character of *-sti*: it behaves morphosyntactically in a similar manner as traditional case desinences. Yet all Finnish suffixes that derive adverbs attach to one of the same nominal stems used by cases as well, do not allow intervening morphemes, and adapt to vowel harmony. They cannot be used with numerals or comparatives and superlatives (apart from *-in* that in so doing also behaves like *-sti*) – but that is a distributional, not a morphosyntactic property. In the discussion (p. 377), Jääskeläinen presents an important point about *-sti* as a potential case marker: not all the 15 canonical cases of Finnish mark verbal complements. In other words, if we discount *-sti* as a case based on perhaps the most used definition of case, the marking of dependent nouns in relation to their verbal, adpositional, or nominal heads, we should exclude the abessive, at the very least. Also, the abessive cannot generally be used in complex NPs. Jääskeläinen's arguments will likely remain without wider support, but the discussion in this chapter should not be dismissed. It highlights many crucial aspects of the debate involving the more marginal cases of Finnish and case in general.

5 It is not cognitive, but what is it then?

Here I attempt to summarize my criticism regarding the role of Cognitive Linguistics in this volume. Yet a word of contextualization: as has hopefully transpired already, my overall impression of the book is overwhelmingly positive, and the following words do not alter this general judgement.

My first complaints are subjective and perhaps predominantly aesthetic. As a rule, I found it harder to interpret the various diagrams illustrating the semantics of both local and more grammatical uses of case than their explanations in prose. In almost all instances, the text beats the drawing. I understand that the authors do not use these diagrams blindly only because of the traditions of CG; they may genuinely provide information more quickly for a trained reader. In her article, Jääskeläinen argues that the suffix *-sti* contributes to three separate constructions, and, to me, the constructional representation

of the intensifying adverbs, derived from curse words and a handful of adjectives (Fig. 3, p. 376), is informative, compact, and elegant. However, the constructional representations of the adverbs of manner (Fig. 1, p. 370) and adverbs of absolute frequency (called "multiplicative" by Jääskeläinen, a perfectly adequate term as well; Fig. 2, p. 373), seem contrived: these functions of the suffix *-sti* seem merely additive in nature.

Moving toward more conceptual questions, I have tried to track the use of the terminological pair landmark–trajector throughout the book. Whereas the introduction (p. 15) claims that the trajector/landmark alignment "often coincides with the categories of traditional syntax", of all authors Jaakola seems to offer the most elaborated treatment of the landmark–trajector pair, and, in fact, ties this into the basic argumentation of the study reported in the chapter. Jaakola presents a hypothesis: the two syntactic components of the adnominal genitive phrase can be thought of in terms of reference point asymmetry. The extent to which reference point asymmetry represents anything observable regarding human cognition is a further question I will come back to but let us assume for now it does. Whether the reference point is equated with landmark and the target with trajector remains slightly unclear. Yet to me, their definitions seem almost identical, respectively. However, landmark and trajector are predetermined regarding syntax: modifiers are landmarks, whereas heads are trajectors. The last step in the operationalization of the hypothesis is to choose a classification of the inherent salience of referents and metrics to determine the referents' salience in discourse as well.

But then what do the results – genitive marked adnominal nouns are more often human-like than the head noun and they are more often discourse-given and can be picked up easier cataphorically – tell us? A human can potentially be said to possess almost anything, but when *a car* is used adnominally, it cannot signal possession in the same sense. Is it really because humans allegedly conceptualize certain things as the reference point and others as the target that a construction used to signal possession, albeit also various other relations between two nominals, shows certain statistical distributional asymmetries whereby one participant is, on average, more often human-like? Or is it because we simply live in societies where humans possess things and cars do not?

I previously noted a problem with an example in Huumo's paper on partitives – or rather what it is used to illustrate. The discussion on kind-referring mass nouns is preceded by a discussion (pp. 53–55) on the CG account of mass nouns. Huumo refers to Langacker's (2016: 85) claim that proper names and kind-referring mass nouns (*milk is good for you*) both have a "unique reference". This is because, Huumo paraphrasing Langacker, these mass nouns name a substance as an "undifferentiated whole" that is "maximally inclusive". According to Langacker, these mass nouns are "characterized by quality" rather than a spatial manifestation, and therefore the "mass noun referent is unique". Langacker, however, seems to define referential uniqueness quite traditionally, assuming that proper names refer to unique objects. Huumo perhaps identifies the motivation behind Langacker's extraordinary equation of the reference of proper names and kind-referring mass nouns, noting on page 54 that in English the article is omitted from these mass nouns, as well as with proper names. It is not sure whether this indeed prompts Langacker's analysis, but article omission in proper names and kind-referring mass nouns are two unconnected, language-specific diachronic outcomes that must be discussed in the context of the English definite article.

Most of the chapters of the book refer to CG concepts, and by doing so, to Langacker. His academic prose uses virtually no references, not to other linguists, nor to cognitive scientists or psychologists. Langacker's *Nominal structure in cognitive grammar* (2016)

is, nevertheless exceptional because it is compiled from lectures, and each chapter ends with a question–answer session. I will use one of his answers in hope to shed some light on his thinking.

> **Question:** In the beginning of your talk you talked about *this rock* as an instance of grouping. Is there any evidence coming from language psychology that it's really something that is done.
> **Langacker:** Well, surely the answer is no. What psychologist would take seriously this characterization, these outlandish notions, and try to test them out? There's no motivation for people to take up this challenge. All I'm telling you is a pretty story – but it's a coherent story and everything fits. (Langacker 2016: 99.)

In all fairness, Langacker goes on for a few sentences to repeat his account of grouping. Yet he openly admits that his concept of grouping is not based on what is known in psychology about cognition but rather he tells people: "trust me, I know". I have highlighted a few assumptions about human cognition made in the book – not because I do not intuitively believe them or that I could prove otherwise, but because no empirical evidence is presented in their support. In concrete terms, an overwhelming majority of the higher theoretical concepts used in this volume refer to Langacker, but Langacker, in turn, refers to no one.

The volume does distance itself to a degree from making cognitive claims. The authors of the introduction present a cautious formulation in this respect: "this volume focuses on meaning organization construed by the case system of the Finnish language but does not make claims as to its relation to cognition". As I have shown, this is not necessarily true regarding all chapters. Continuing on the topic of the relationship of Cognitive Linguistics with cognition, Silvennoinen (2023) asks whether Construction Grammar is, in fact, cognitive. Drawing from a selection of corpus-based CxG studies, Silvennoinen recommends that "corpus-based construction grammarians can content themselves with describing language as a social phenomenon; argumentative leaps to mental representations should be treated with caution".

6 Final remarks

I set out to evaluate this collective volume as an up-to-date reference work. *The Finnish case system: Cognitive linguistic perspectives* has a coordinated organization and shares key terminology – and explicitly discusses differing terminological and descriptive choices in the literature. It presents both original research and tries to exhaustively present and describe key areas of the case system. In addition, this volume contains an enormous number of glossed examples with English translations and other contextualized data regarding Finnish. I have a sense that this volume will earn a place as a popular reference for typologists and other scholars interested in Finnish. While this volume is exhaustive, some lacunae remain regarding the Finnish case and adjacent phenomena. A second volume could, perhaps, address the diachronic aspects of the case system, its interplay with the possessive suffix, and analyses of the polysemous markers on verbs, originating in nominal cases.

References

Hakulinen, Lauri. 1979. *Suomen kielen rakenne ja kehitys*. 4th edn. Helsinki: Otava.
Haspelmath, Martin. 2009. An empirical test of the Agglutination Hypothesis. In Scalise, Sergio & Magni, Elisabetta & Bisetto, Antonietta (eds.), *Universals of Language Today*. Heidelberg: Springer.
Inaba, Nobufumi. 2015. *Suomen datiivigenetiivin juuret vertailevan menetelmän valossa.* Helsinki: Suomalais-Ugrilainen Seura.
Langacker, Ronald W. 2016. *Nominal structure in cognitive grammar: The Lublin lectures, edited by Adam Głaz, Hubert Kowalewski, Przemysław Łozowski*. Lublin: Maria Curie-Skłodowska University Press.
Milsark, Gary L. 1977. Toward an explanation of certain peculiarities of the existential construction in English. *Linguistic analysis* 3(2). 1–29.
Pantermöller, Marko. 2016. Datiivigenetiivi – ikivanhaa perintöä vai vanhan suomen nuori uudennos? *Virittäjä* 120(3). 441–447.
Silvennoinen, Olli O. 2023. Is construction grammar cognitive? *Constructions* 15(1). 1–17.
Hakulinen, Auli & Vilkuna, Maria & Korhonen, Riitta & Koivisto, Vesa & Heinonen, Tarja Riitta & Alho, Irja. 2008. *Iso suomen kielioppi* [verkkoversio]. Helsinki: Suomalaisen Kirjallisuuden Seura. https://kaino.kotus.fi/visk/etusivu.php (30 May, 2023).

Contact information:

Max Wahlström
Department of Languages
University of Helsinki
max.wahlstrom@helsinki.fi

Jenny Paananen, Meri Lindeman, Camilla Lindholm ja Milla Luodonpää-Manni (toim.). 2023. *Kieli, hyvinvointi ja haavoittuvuus – Kohti kielellistä osallisuutta*. Gaudeamus. 288 s.

Kirjoittanut Maija Yli-Jokipii

1 Johdanto

Jenny Paanasen, Meri Lindemanin, Camilla Lindholmin ja Milla Luodonpää-Mannin toimittama teos *Kieli, hyvinvointi ja haavoittuvuus – Kohti kielellistä osallisuutta* (Gaudeamus, 2023) sisältää kahdeksan tutkimusartikkelia, kirjallisuuskatsauksen (2. artikkeli), yhden kokoavan tarkastelun (7. artikkeli) sekä toimittajien kirjoittaman johdannon ja jälkisanat. Kuten teoksen nimikin antaa ymmärtää, kaikki artikkelit tarkastelevat kielenkäyttöä ja kielellistä vuorovaikutusta erityisesti osallisuuden ja yhdenvertaisuuden näkökulmasta.

Teos asemoituu erityisesti kielisosiologian ja soveltavan kielitieteen kentille, ja keskiössä on erityisen haavoittuvassa asemassa olevien ihmisten ja ryhmien suhde kieleen ja sen käyttöön. Nämä teemat ovat näkyneet viime vuosina myös esimerkiksi Suomen soveltavan kielitieteen yhdistyksen AFinLa:n vuosikirjoissa *Kieli ja osallisuus* (Hynninen ym. 2023) ja *Kielitietoisuus eriarvoistuvassa yhteiskunnassa* (Latomaa ym. 2017) sekä keskusteluissa rasismista (esim. Keskinen ym. 2021), syrjinnästä ja inklusiivisuudesta. Kokonaisuudessaan teos ottaa osaa keskusteluun inkluusiosta ja osallisuudesta ja osoittaa, miten kieli ja kielenkäyttötilanteet voivat sekä sulkea ulos että osallistaa monenlaisia ihmisryhmiä ja yksilöitä. Teoksessa tarkastellaan monien heikossa asemassa olevien ihmisryhmien, kuten kehitysvammaisten, mielenterveyskuntoutujien ja turvapaikanhakijoiden, kokemia kielellisen vuorovaikutuksen haasteita ja sitä, miten kieli voi rajoittaa eri ihmisryhmien osallisuutta ja oikeuksien toteutumista.

Kielellistä osallisuutta ja siihen vaikuttavia tekijöitä nostetaan esiin monien erilaisten kielellisten vähemmistöryhmien kautta. Tekijät toteavat johdannossa, että kieli vaikuttaa yksilöihin ja heidän mahdollisuuksiinsa monella tavalla. Kieli on yhtä aikaa rikkaus ja voimavara, mutta se voi samanaikaisesti rajata yksilön pääsyä tietoon, palveluihin ja työelämään. Kieli on vallankäytön väline, sillä kielenkäytön avulla luodaan sitä todellisuutta, jossa elämme.

Teos on jaettu kolmeen osaan, joista kunkin aloittaa sen tematiikkaa valottava lyhyt johdantoluku. Myös muilla toimituksellisilla ratkaisuilla on pyritty siihen, että teos olisi yksi kokonaisuus eikä kokoelma erillisiä artikkeleita; esimerkiksi lähteet on koottu koko teoksen loppuun yhdeksi lähdeluetteloksi, ja kunkin luvun lopussa on tiivistelmä, joka summaa luvun tärkeimmät havainnot.

2 Teoksen keskeinen sisältö

Teoksen johdanto-osassa toimittajat esittelevät muutamia ydinkäsitteitä, jotka ovat teoksen eri artikkelien yhdistävä tekijä, ja kuvaavat teoksen yleistä taustaeetosta. Keskeisiksi näkökulmiksi on tässä teoksessa nostettu osallisuus, inkluusio sekä haavoittuvuus, joka näyttäytyy kahden edellisen vastakohtana. Osallisuuden käsitettä tarkastellaan erityisesti osana yhteisöihin tai ryhmiin kuulumista. Osallisuus näyttäytyy merkityksellisten

ihmissuhteiden rakentumisena, aktiivisena toimijuutena sekä toimeentulon saamisena ja välttämättömien tarpeiden täyttymisenä. Inkluusio puolestaan nähdään erityisesti yhteisön pyrkimyksenä osallistaa kaikki jäsenensä. Haavoittuvuus-käsitteen määritelmäksi on valittu Terveyden ja hyvinvoinnin laitoksen (2024) muotoilu väestöryhmistä, ”joilla oman vaikutusvaltansa ulkopuolella olevista tekijöistä johtuen ei ole samoja mahdollisuuksia kuin hyväosaisemmilla väestöryhmillä”. Kielellisesti haavoittuvina ryhminä teoksessa käsitellään erityisesti vähemmistökielten käyttäjiä sekä ryhmiä, joiden kielelliset kompetenssit poikkeavat enemmistön kompetensseista. Lisäksi haavoittuvuuden syntyyn vaikuttaa se, että kielen avulla on mahdollista sekä luoda että purkaa yhteiskunnallista eriarvoisuutta. Haavoittuvuus-käsitteen problemaattisuus tunnistetaan, mutta sen käyttö perustellaan osana pyrkimystä tehdä kielenkäyttöön liittyvä syrjäytyminen näkyväksi.

Teoksen johdannossa toimittajat avaavat ajatusta siitä, että teoksella halutaan ottaa osaa keskusteluun suomalaisesta kielipolitiikasta. Kielen käyttö on viranomaisten ja päätöksenteon lisäksi jokaisen kielellisen toimijan, esimerkiksi yrityksen, perheen ja yksilön (Pöyhönen ym. 2019), vastuulla, ja siksi kielellistä osallisuutta rakennetaan moninaisin toimin. Johdanto-osa pohjustaa kutakin teoksen kolmesta osasta, mikä tuntuu osittain turhalta, etenkin kun kunkin osan alussa on vielä erillinen johdantoluku. Toisaalta teoksella tavoitellaan laajempaa monitieteistä lukijakuntaa, joten johdanto-osan laajempi taustoitus palvelee niitä lukijoita, jotka eivät ole seuranneet kielitieteen viimeaikaisia keskusteluja kovin tarkasti.

2.1 Ensimmäinen osa: *Kielellisistä oikeuksista kohti aitoja mahdollisuuksia*

Teoksen ensimmäinen osa on otsikoitu *Kielellisistä oikeuksista kohti aitoja mahdollisuuksia*, ja siinä tarkastellaan kolmea eri kielellisesti hyvin haavoittuvassa asemassa olevaa ryhmää: kehitysvammaisia, viittomakielisiä sekä turvapaikanhakijoita. Leealaura Leskelä ja Camilla Lindholm tarkastelevat ensimmäisessä artikkelissa *Selkopuhe kehitysvammaisen henkilön kielellisen osallisuuden tukena* kehitysvammaisten ja heidän hoitajiensa välistä kielellistä vuorovaikutusta ja osoittavat, että osallisuutta rakentava vuorovaikutus vaatii hoitohenkilökunnalta taitoa ja herkkyyttä kuunnella ja tulkita kehitysvammaisten vuorovaikutusaloitteita, jotta henkilökunta ei määritä vuorovaikutuksen sisältöä ja suuntaa. Leskelän ja Lindholmin keskustelunanalyysiä hyödyntävän tutkimuksen aineistona ovat kehitysvammaisten henkilöiden sekä heidän ohjaajiensa väliset videoidut keskustelut. Tutkimus osoittaa, että kehitysvammaisten kielellinen osallisuus on lisääntynyt selkopuheen käytön lisääntyessä, mutta heidän yhdenvertaiseen osallisuuteensa tulee jatkossakin kiinnittää huomiota ja jatkaa selkopuheen käytön tukemista etenkin kehitysvamma-alalla.

Teoksen toinen artikkeli on muista teoksen artikkeleista poiketen luonteeltaan kirjallisuuskatsaus. Artikkelissaan *Suomen viittomakielten käyttäjien osallisuus ja haavoittuvuus 1850-luvulta nykypäivään* Ritva Takkinen, Karoliina Nikula ja Juhana Salonen käyvät läpi suomalaisen ja suomenruotsalaisen viittomakielen käytön historiaa erityisesti viittomakielisen koulutuksen saatavuuden näkökulmasta. Kirjoittajat muun muassa osoittavat, että teknologian tuomat mahdollisuudet saattavat toisinaan myös kaventaa kielellisten vähemmistöjen oikeuksia, kun esimerkiksi kuulovammaisten oikeus omaan kieleen ja kulttuuriin kyseenalaistetaan, koska sisäkorvaistutteiden avulla kuulo on mahdollista palauttaa ainakin jollain tasolla. Tämä artikkeli on aiheensa ja tarkastelunäkökulmansa puolesta tärkeä nosto kansallisten vähemmistökielten puhujien asemaa ja oikeuksia koskevaan keskusteluun.

Kolmannessa artikkelissa *Haavoittuvuus ja osallisuus turvapaikkapuhuttelussa* Simo Määttä, Eeva Puumala ja Riitta Ylikomi kuvaavat niitä HLBTQI-turvapaikanhakijan turvapaikkapuhutteluun liittyviä kielellisiä ja vuorovaikutuksellisia seikkoja, joilla voi olla vaikutusta turvapaikkapäätöksen saamiseen. Tutkimuksen aineistona on videoitu turvapaikkapuhuttelu, jota kirjottajat analysoivat monimetodisin menetelmin. Määtän ja kumppanien artikkeli tuo esiin, että näennäisesti pientenkin kielellisten ratkaisujen merkitys voi olla huomattavan suuri, etenkin kun tarkastellaan tilanteita, joissa valtahierarkioiden epätasapaino on suuri. Erityisesti tämäntyyppisissä vuorovaikutustilanteissa sekä tulkin ammattitaito että viranomaisen ymmärrys kielellisesti vastuullisen vuorovaikutuksen merkityksestä on tärkeää. Kirjoittajat ovat kuvanneet koko tutkimuksen ja sen teoreettiset ja metodologiset lähtökohdat selkeästi, ja teksti on siksikin vakuuttavaa.

2.2 Toinen osa: *Kielitaito hyvinvoinnin kulmakivenä*

Teoksen toisessa osassa tarkastellaan kielitaidon ja hyvinvoinnin suhdetta kolmesta hieman erilaisesta näkökulmasta. Tämän osan aloittaa teoksen neljäs artikkeli *Hyvinvoinnin asema maahanmuuttajien lukutaitokoulutuksen opetussuunnitelmissa*. Tässä artikkelissa Taina Tammelin-Laine tarkastelee sitä, miten opiskelijoiden hyvinvointi näkyy maahanmuuttajien lukutaitokoulutusten opetussuunnitelmissa. Keskeiseksi taustateoriaksi nousee Allardtin hyvinvoinnin kolmen ulottuvuuden malli (Allardt 1976; 1993), ja sitä olisin suonut käytettävän vahvemmin myös sisällönanalyysiä ohjaavana teoriana, sillä nyt lukijalle jää hieman epäselväksi, miten analyysi on toteutettu. Tammelin-Laine tarkastelee dokumenttianalyysissään opetussuunnitelmia monipuolisesti, ja lukutaitokoulutuksen opiskelijoiden hyvinvoinnin ulottuvuudet esitellään lukijalle kattavasti.

Minna Intke-Hernandézin artikkeli *Yhteinen arkemme kielenoppimisen ja hyvinvoinnin tukena* osoittaa, että kielellinen osallisuus voi lähteä kasvamaan pienistä kielenkäyttötilanteista. Kielenoppijat itse eivät osaa aina edes pitää merkityksellisinä tilanteita, jotka ovat luoneet mahdollisuuden havainnoida kieltä ja päästä osaksi kieliyhteisöä. Usein tällaiset arkiset kohtaamiset ovat saattaneet olla ratkaiseva tekijä kielenoppijan minäpystyvyyden ja motivaation syntymisessä. Intke-Hernandéz kuvaa artikkeliin liittyvää monimenetelmäistä neksusanalyysiä hyödyntävää tutkimusprosessia kiinnostavasti ja havainnollisesti, ja tutkijan oma positio, joka on tässä tutkimuksessa ollut keskeinen tekijä, on sanoitettu selkeästi.

Annmari Sahlstein kuvaa teoksen kuudennessa artikkelissa *Suomenkielisten lääkäriopiskelijoiden asenteet ruotsinkielistä potilastyötä kohtaan* toimintatutkimustaan, jonka aineistona ovat olleet lääketieteen opiskelijoilta kerätyt kyselyvastaukset. Tutkimus on monimenetelmäinen, ja lisäksi siinä on seurantatutkimuksen elementtejä, sillä samoilta opiskelijoilta on kerätty vastauksia useampaan kertaan heidän opintojensa aikana, ja näin on voitu tarkastella asenteiden kehittymistä. Sahlstein on havainnut, että opiskelijoiden asenteet muuttuvat myönteisemmiksi opintojen kuluessa etenkin, jos heillä on mahdollisuus saada kokemusta ruotsinkielisestä potilastyöstä. Aihe on tärkeä, sillä mahdollisuus käyttää omaa äidinkieltä vaikuttaa sekä potilaan ja lääkärin väliseen hoitosuhteeseen että potilasturvallisuuteen. Sahlsteinin artikkeli olisi hyötynyt vielä perusteellisemmasta toimituksellisesta työstä erityisesti sen kuvioiden ja diagrammien osalta, jotta lukijan olisi helpompi seurata taulukoissa esitettyjä huomioita.

2.3 Kolmas osa: *Toiseuttavasta kielenkäytöstä kohti sisällyttävää viestintää*

Kolmannessa osassa tarkastellaan kielenkäyttöön sisältyviä toiseuttavia ja sisällyttäviä käytäntöjä. Leea Lakan artikkeli *Lukutaito ja osallisuus – onko yhtä ilman toista?* pohtii juuri sitä, millä tavoin lukutaito ja osallisuus kietoutuvat toisiinsa. Artikkeli tarkastelee niitä moninaisia tekijöitä, jotka vaikuttavat yksilöiden ja yhteisöjen lukutaidon kehitykseen, ja, kuten kirjoittaja itsekin toteaa, tuo tarkasteluun myös sosiologisia teorioita ja näkökulmia. Lakan artikkeli on vaikeasti määrittyvä, sillä se ei ole tutkimusartikkeli eikä kirjallisuuskatsaus vaan pikemminkin kokoava tarkastelu, jossa kirjoittaja pohtii ja erittelee niitä tekijöitä, joita lukutaidon ja osallisuuden suhteeseen voidaan liittää. Lakan artikkelin näkökulmien runsaus kuvaa hyvin aiheen monitahoisuutta: lukutaitoon liittyy monenlaisia yksilöön ja erityisesti yhteisöihin liittyviä muuttujia. Tärkeänä havaintona artikkelissa nostetaan esiin se, että lukutaidon kehittäminen vaatii pysyviä resursseja ja pitkäjänteistä työtä kaikilla koulutusjärjestelmämme tasoilla.

Anna Weckström puolestaan kysyy kokoelman kahdeksannessa artikkelissa: *Kenelle mielenterveydestä viestitään?* Weckströmin näkökulmana on lukijuus, jota mielenterveysviestinnässä rakennetaan, ja aineisto koostuu mielenterveyskentän toimijoiden sähköisistä, suurelle yleisölle suunnatuista materiaaleista. Weckströmin analyysi osoittaa, että vain osa mielenterveysviestinnästä on suunnattu mielenterveyden häiriöistä kärsiville ja että mielenterveysviestintä tuottaa ulkopuolisuutta ja voi näyttäytyä holhoavana ja leimaavana. Weckströmin valitsema kategoria-analyysi palvelee tutkimusta tarkoituksenmukaisesti, ja se on esitelty lukijalle riittävän tarkasti.

Edellisen artikkelin teemaa jatkaa yhdeksäs artikkeli, *Me ja te mielenterveys-kuntoutuksen päätöksentekokeskusteluissa*. Jenny Paananen, Camilla Lindholm, Melisa Stevanovic, Elina Weiste, Taina Valkeapää ja Samuel Tuhkanen ovat tehneet laajan keskustelunanalyysitutkimuksen ja tarkastelleet sitä, miten persoonapronominien käyttö rakentaa tai estää osallisuuden toteutumista mielenterveyskuntoutuksessa. Tarkastelun kohteena on ollut osallisuutta ja toimijuutta korostava Klubitalo-kuntoutusmalli, ja Paananen ja kumppanit ovat huomanneet, että myös tällaisen toimintamallin puitteissa puhetavat rakentavat institutionaalista hierarkkisuutta. Paanasen ja kumppanien tutkimus on raportoitu selkeästi ja argumentaatio kiinnittyy aiempaan kirjallisuuteen erityisesti keskusteluanalyyttisen tutkimuksen kehyksessä (esim. Paananen ym. 2020; Stevanovic ym. 2022; Urfalino 2014; Valkeapää ym. 2019).

Teoksen kymmenennessä artikkelissa *Sarjakuva, sosiaalihuolto ja saavutettavuus* Laura Kalliomaa-Puha, Anne Ketola ja Eliisa Pitkäsalo esittelevät kiinnostavan näkökulman kielelliseen saavutettavuuteen. Heidän tutkimuksessaan tarkastellaan sarjakuvan käyttöä viranomaistekstien muokkaamisessa saavutettavampaan muotoon. Kaksiosainen tutkimus esittelee ensin intersemioottisen käännösprosessin, jossa viranomaisteksti muutetaan sarjakuvamuotoon, ja analysoi sen jälkeen sarjakuvan vastaanottoa selvittävää kyselyaineistoa. Tutkimus osoittaa, että vaikka sarjakuvan muotoon muokattu teksti auttaa osaa lukijoista, sarjakuva tekstilajina sekä kuvien tulkinta ei ole kaikille kohderyhmän jäsenille tuttua tai ongelmatonta. Kalliomaa-Puhan ja kumppaneiden artikkeli tuo kielellisesti osallistavaan keskusteluun uuden mielenkiintoisen näkökulman, sarjakuvan käytön virallisten tekstien tukena. Samalla huolellisesti toteutettu tutkimus osoittaa, että kriittinen tarkastelu on paikallaan uudenlaisten viestintämuotojen kehittämisessä, sillä esimerkiksi kuvallinen ilmaisu ei ole välttämättä yleismaailmallista ja yksiselitteisesti tulkittavaa.

3 Lopuksi

Paanasen ja kumppaneiden toimittama teos on aiheeltaan ajankohtainen ja valottaa monia kielenkäyttöön liittyviä vallan ja osallisuuden näkökulmia. Yksi keskeinen ja merkittävä huomio, joka toistuu useissa teoksen artikkeleissa, on se, että kielenkäyttöön liittyvää osallisuutta voidaan rakentaa ottamalla kielelliset vähemmistöt ja kielenoppijat sekä muut haavoittuvat ryhmät mukaan kielenkäytön suunnitteluun ja toteutukseen. Kyseessä on selkeästi valtahierarkioihin liittyvä ilmiö, jota voidaan purkaa vain tarkastelemalla hierarkioita kriittisesti ja panostamalla osallisuutta tuottaviin toimintamalleihin. Kuten Minna Intke-Hernandez artikkelissaan toteaa, ”osallisuus on äänettömyyden rikkomista”.

Toimituksellisista ratkaisuista kaikkien artikkelien yhtenäinen lähdeluettelo toimii muuten hyvin, mutta eri kirjallisuuden lajien ja eri julkaisumuotojen luetteloiminen kukin omaan ryhmäänsä tuntuu erikoiselta eikä oikein palvele lukijaa. Lisäksi haastetta tuo se, että joissain tapauksissa lähdeviite ei ole riittävän tarkka: esimerkiksi viitteitä Pöyhönen ym. 2019 on kaksi, mutta niitä ei ole tekstissä tai lähdeluettelossa yksilöity esimerkiksi kirjaintunnistein. Lähdeviitteissä on jonkin verran muitakin epätarkkuuksia. Lisäksi ala- ja loppuviitteiden olisin suonut olevan kunkin luvun yhteydessä; nyt niitä joutuu erikseen etsimään.

Kunkin osan alussa olevat lyhyet johdantoluvut sekä jokaisen artikkelin loppuun koottu tiivistelmä auttavat lukijaa hahmottamaan teoksen olennaiset näkökulmat ja tutkimusten keskeiset löydökset. Tämä tekee teoksesta helposti lähestyttävän muillekin kuin kielitieteen tutkimuskentällä toimiville lukijoille. *Kieli, hyvinvointi ja haavoittuvuus* -kokoelman ansiot ovatkin vahvasti yhteiskunnallisia. Pidän ansiokkaana myös sitä, että kielitieteen alalla halutaan julkaista tärkeistä teemoista suomeksi, suomenkieliselle lukijakunnalle.

Lähteet

Allardt, Erik. 1976. *Hyvinvoinnin ulottuvuuksia*. WSOY.

Allardt, Erik. 1993. Having, loving, being: An alternative to the Swedish model of welfare research. *The quality of life* 8, 88–95.

Hynninen, Niina & Herneaho, Irina & Isosävi, Johanna & Sippola, Eeva & Yang, Mei (toim.). 2023. *Kieli ja osallisuus*. (Suomen soveltavan kielitieteen yhdistyksen julkaisuja 80.) Suomen soveltavan kielitieteen yhdistys AFinLa.

Keskinen, Suvi & Seikkula, Minna & Mkwesha, Faith (toim.). 2021. *Rasismi, valta ja vastarinta: Rodullistaminen, valkoisuus ja koloniaalisuus Suomessa*. Gaudeamus.

Latomaa, Sirkku & Luukka, Emilia & Lilja, Niina (toim.). 2017. *Kielitietoisuus eriarvoistuvassa yhteiskunnassa–Language awareness in an increasingly unequal society*. (Suomen soveltavan kielitieteen yhdistyksen julkaisuja 75.) Suomen soveltavan kielitieteen yhdistys AFinLa.

Paananen, Jenny & Lindholm, Camilla & Stevanovic, Melisa & Valkeapää, Taina & Weiste, Elina. 2020. “What Do You Think?” Interactional Boundary-Making Between “You” and “Us” as a Resource to Elicit Client Participation. Teoksessa Lindholm, Camilla & Stevanovic, Melisa & Weiste, Taina (toim.), *Joint Decision Making in Mental Health: An Interactional Approach*, 211–234. Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-030-43531-8_9

Pöyhönen, Sari & Nuolijärvi, Pirkko & Saarinen, Taina & Kangasvieri, Teija. 2019. Kielipolitiikka ja kielikoulutuspolitiikka monipaikkaisina ilmiöinä ja tutkimusaloina. Teoksessa Saarinen, Taina & Nuolijärvi, Pirkko & Pöyhönen, Sari & Kangasvieri, Teija (toim.), *Kieli, koulutus, politiikka: Monipaikkaisia käytänteitä ja tulkintoja*, 9–24. Vastapaino.

Stevanovic, Melisa & Valkeapää, Taina & Weiste, Elina & Lindholm, Camilla. 2022. Joint decision making in a mental health rehabilitation community: The impact of support workers’ proposal design on client responsiveness. *Counselling Psychology Quarterly* 35(1). 129–154. https://doi.org/10.1080/09515070.2020.1762166

Terveyden ja hyvinvoinnin laitos. 2024. *Keskeisiä käsitteitä*. (https://thl.fi/aiheet/hyvinvoinnin-ja-terveyden-edistamisen-johtaminen/hyvinvointijohtaminen/hyvinvointi-ja-terveyserot/keskeisia-kasitteita) (Viitattu 21.4.2024.)

Urfalino, Philippe. 2014. The rule of non-opposition: Opening up decision-making by consensus. *Journal of Political Philosophy* 22(3). 320–341. https://doi.org/10.1111/jopp.12037

Valkeapää, Taina & Tanaka, Kimiko & Lindholm, Camilla & Weiste, Elina & Stevanovic, Melisa. 2019. Interaction, ideology, and practice in mental health rehabilitation. *Journal of Psychosocial Rehabilitation and Mental Health* 6. 9–23. https://doi.org/10.1007/s40737-018-0131-3

Yhteystiedot:

Maija Yli-Jokipii
Tallinnan yliopisto
Suomen kielen ja kulttuurin vieraileva lehtori
maijayli@tlu.ee

Corrigendum

Yida Cai

The author information in the book review "Dalrymple, Mary & Lowe, John J. & Mycock, Louise (eds.). 2019. The Oxford Reference Guide to Lexical Functional Grammar. Oxford: Oxford University Press. Pp. 835" published in the issue Vol. 35 (2022) of the Finnish Journal of Linguistics was incorrect, the correct information is "Dalrymple, Mary & Lowe, John J. & Mycock, Louise. 2019. The Oxford Reference Guide to Lexical Functional Grammar. Oxford: Oxford University Press. Pp. 835". Consequently, the first sentence of the book review is "The Oxford reference guide to Lexical Functional Grammar is a volume offering wide-ranging and detailed information about Lexical Functional Grammar (LFG)."

Finnish Journal of Linguistics 37 (2024), 205
DOI: 10.61197/fjl.152317