Interaction between case marking, grammatical role and co-reference in Finnish: a corpus study

Hessel Haagsma (s1932683)

Course: LFF999B10 - Ba-scriptie Fins
Supervisor: prof. Cornelius Hasselblatt

Abstract

In this thesis, I present a corpus study investigating the relationship between case marking, grammatical role and co-referential properties in Finnish. The main focus is on the interaction between grammatical role and case marking and if and how this influences the position of the argument on the Noun Phrase (NP) accessibility hierarchy. It is proposed that a ranking based only on grammatical role is too simplistic for Finnish, and that the concept of grammatical role is more fluid than in other, more widely-studied languages. A corpus study was carried out on a collection of novels and newspaper text, extracting referential uses of *hän* and *tämä*. The results suggest influences of non-canonical marking on referential form, but suffer from low occurrences of many specific sentence types in the corpus. This, too, prevented conclusions about the co-referential properties of many non-canonically marked arguments. However, it was found that the argument of an impersonal construction, generally considered an object, behaves very much like a subject in terms of its co-referential properties. Additionally, genitive possessor showed a higher preference for *hän* than would be expected on basis of the NP accessibility hierarchy. Also, the corpus study showed the importance of competitors being present for *hän*/*tämä*-variation to occur, which should be taken into account when doing further research into their distribution.

1 Introduction

The human capacity for establishing relations in language between two or more phrases that refer to the same entity, also known as co-reference, is a remarkable feat and central element of human language processing. Both reference production and resolution require extensive linguistic and background knowledge, either to produce a sufficiently unambiguous reference (also called an *anaphor*), or to identify that which is referred to (the *antecedent*). Even more remarkably, humans manage to do this at great speed, and seemingly without effort.

An obvious explanation for this is that we possess a lot of knowledge about the world, which guides our interpretation, but this alone is insufficient. In addition to world knowledge, there are a lot of cues in language which facilitate co-reference. One case where these linguistic cues play an important role is in the resolution of ambiguous anaphora, i.e. in language comprehension (cf. Example (1a))[1], or conversely, in choosing between two possible reference forms, i.e. in language production (cf. Example (1b,c)), which are situations in which real-world knowledge does not always help.

(1) a. **Anssi** halas-i **Anna-a** ja *hän* läht-i kotiin.
Anssi.NOM hug-PST.3SG Anna-PAR and s/he.NOM leave-PST.3SG home.to
'**Anssi** hugged **Anna** and *s/he* went home.'

b. Vieraa-t halas-i-vat **Anna-a** ja *hän* läht-i kotiin.
guest-NOM.PL hug-PST-3PL Anna-PAR and s/he.NOM leave-PST.3SG home.to
'The guests hugged **Anna** and *she* went home.'

[1] Only relevant morphological information is represented in glosses. Anaphora are marked by *italics* and possible antecedents are in **bold**.

The used gloss abbreviations are as follows: 3 - third person, ACC - accusative case, GEN - genitive case, NOM - nominative case, PAR - partitive case, PL - plural, PST - past tense, SG - singular. For an in-depth discussion of used case definitions, see Section 1.3.2.

c. Vieraa-t halas-i-vat **Anna-a** ja *tämä* läht-i kotiin.
guest-NOM.PL hug-PST-3PL Anna-PAR and this.NOM leave-PST.3SG home.to
'The guests hugged **Anna** and *she* went home.'

In sentence (1a), *hän* can mean both he and she (as Finnish has gender-neutral pronouns only), and therefore is ambiguous: it could refer just as well to **Anssi** as to **Anna**. Nevertheless, *hän* is usually interpreted as referring to **Anssi**, the subject. The reason for this is that *hän* is more often used for reference to the subject than to the object, and therefore *hän* is interpreted as referring to **Anssi** instead of **Anna**, the object (Järvikivi, van Gompel, Hyönä & Bertram, 2005; Kaiser, 2005; Kaiser & Trueswell, 2008). Note, however, that the most natural sentence with **Anssi** as the antecedent would have a zero-pronoun anaphor, as in '**Anssi** hugged **Anna** and ø went home.'.

Sentences (1b,c) show that two different pronouns, *hän* and *tämä*, can be used to refer to **Anna**. *Hän* is a personal pronoun, equivalent to English *he* and *she*, while tämä is a demonstrative pronoun, and is equivalent to English *this*. Contrary to its English counterpart, though, *tämä* can also be used to refer to singular human entities in third person[2]. The traditional view on their distribution holds that *hän* is used to refer to subjects, and *tämä* is used for object-reference (Halmari, 1994, for example). This is not the complete explanation, as more recent research indicates that, although *hän* is used for subject-reference, *tämä* is mainly used to refer to last-mentioned non-subjects in the sentence (Kaiser & Trueswell, 2008). In this light, (1c) seems to be a more natural choice of anaphor than sentence (1b).

Note that, although the use of demonstrative pronouns to refer to humans seems odd in English, it is far from uniquely Finnish. Similar systems, with similar underlying mechanics, can be found in Dutch, for the referential pronoun pairs *hij/zij* and *die/deze* (Kaiser, 2011), in German, for the pronoun triples *er/sie/es* and *der/die/das* (Bosch, Katz & Umbach, 2007) and in Estonian, for the pronouns *ta* and *see* (Kaiser & Vihman, forthcoming).

Because of the common occurrence of such systems, general underlying theories have been proposed to account for their distribution and that of anaphor form variation in general. Most theories focus on grammatical role and linear word order as the largest influences on anaphor choice, and explain this using an interaction between a hierarchy of anaphor forms and a ranking of possible antecedents.

The hierarchies, and the connection between them, are discussed in Section 1.1. A more in-depth discussion of previous research into the *hän/tämä* distinction is presented in Section 1.2. In addition to the well-researched influences of syntactic role and word order, this study will focus on the influence of canonical and non-canonical case marking on co-reference in general and anaphor choice in particular. Section 1.3 will provide a background discussion of Finnish case marking, and will discuss instances of non-canonical marking that might be of interest. In order to explore the hypotheses proposed in that section, a short re-analysis of an existing data set is presented in Section 1.4.

1.1 Referential form and NP accessibility: two hierarchies

To explain the interaction between anaphor form and antecedent properties, several theories have been proposed. These theories aim to correlate a ranking of referential expressions with a hierarchy of possible antecedent NPs, ranked by a certain property. Three variations of such a ranking of referential expressions have been proposed by Prince (1981a), Gundel, Hedberg and Zacharski (1993) and Arnold (2001). In this study, we will be concerned with the ranking devised by Gundel et al., the so-called Givenness hierarchy, as it is widely-used and practical for our purposes. For the ranking of possible antecedent noun phrases, the widely-used NP accessibility hierarchy, devised by Keenan and Comrie (1977) will be used as a starting point, introducing the key concept of accessibility.

1.1.1 The Givenness hierarchy

The Givenness hierarchy ranks anaphor forms in terms of cognitive status. This cognitive status entails how actively present something is in one's mind, how much attention is paid to it, and how

[2]In colloquial spoken language, *hän* can also be used to refer to other animate entities, and sometimes even for inanimate objects (Varteva, 1998).

well-known it is. In this hierarchy, the choice of anaphor form is determined by the cognitive status of the antecedent. For example, usage of the personal pronoun *he* would indicate that the antecedent is something that is well-known, has recently come to attention and is the focus of the current discourse. Similar outlines of antecedent candidates can be given for other referential expressions.

These referential expressions can then be ranked by the cognitive status of their possible antecedents, creating the Givenness hierarchy. For example, *he* will always refer to something with a high cognitive status, as it would otherwise be very ambiguous. Vice versa, a full noun phrase like *the big yellow canary* can be used for something that has a low cognitive status, as it will usually be clear what it refers to. A simplified version of the Givenness hierarchy is shown in Table 1.

	Type of referential form	**English equivalents**	**Finnish equivalents**
more specified	Proper Name	Anne	Anna
↑	Full NP	the/a Noun (the/a dog)	Noun (koira)
|	Demonstrative pronoun	this/that	*tämä*/tuo
↓	Personal pronoun	he/she/it	*hän*/se
less specified	Null pronoun	Ø	Ø

Table 1: A simplified version of the Givenness hierarchy, extended with Finnish anaphoric expressions (adapted from Gundel et al. (1993)).

1.1.2 The NP accessibility hierarchy

Although the Givenness hierarchy generally specifies what kind of referential forms can be used for what kind of NPs, its definitions are too broad for practical use. To classify possible antecedents more precisely, the concept of *accessibility* is introduced. Accessibility is the central notion of the NP accessibility hierarchy, as defined by Keenan and Comrie (1977), which indicates whether a specific NP can be relativized, i.e. whether it is *accessible* for reference using a relative clause. Keenan and Comrie suggested that, if a certain type of noun phrase can be relativized, the same holds for a higher-ranked NP type, and that this is a universal feature of language. The default NP accessibility hierarchy is as follows, with the most accessible form on the left:

Subject > Direct object > Indirect object > Oblique > Possessor > Complement

Figure 1: The default NP accessibility hierarchy, after Keenan and Comrie (1977)

Although it originally described relativization, the NP accessibility hierarchy has been extended to account for co-reference. Then, the hierarchy does not absolutely express what can and what cannot be referred to using a specific referential form, but rather indicates how accessible a certain noun phrase is for reference in general. This is then combined with a ranking of referential forms, such as the Givenness hierarchy, by linking accessibility to cognitive status: the more accessible an NP, the higher its cognitive status. Thus, an anaphoric expression that ranks high on the Givenness hierarchy, e.g. the null pronoun, is most likely to be used for reference to an NP that is high on the accessibility hierarchy, such as a subject. Similarly, a referential expression that is low in Givenness, such as a full, descriptive NP, will usually be used to refer to an entity lower on the accessibility hierarchy, e.g. an indirect object.

This is in accordance with intuitions about grammaticality: one is unlikely to interpret a zero pronoun as referring to an indirect object, such as **me** in the sentence 'He gave the book to **me** and Ø left.', or conversely, to use a full NP to refer to a subject repeatedly, as in the sentence '**John** came into the room, **John** looked at Mary and **John** left.'.

As Figure 1 shows, Keenan and Comrie used a straightforward approach to classifying the NPs, ranking them by only one property, grammatical role. With regard to co-reference, it immediately becomes clear that this approach is too simple. The choice between *tämä* and *hän*, for example, is influenced not only by grammatical role, but also by word order. Therefore, an NP accessibility hierarchy that can account for this distinction should include linear word order too. However, this would still be insufficient to represent accessibility, which should be seen as a compound notion, influenced by many different aspects and linguistic cues. Examples of features that are known to

influence accessibility are animacy (Fukumura & van Gompel, 2011), verbal aspect (Ferretti, Rohde, Kehler & Crutchley, 2009), thematic role (Arnold, 2001), whether the NP occurs in a matrix or a subordinate clause (Kaiser, 2000), distance between anaphor and antecedent (Ariel, 1988) and contrasts in discourse (Kaiser, 2011). As mentioned earlier, the aim of the current study is to answer the question of whether case marking should be added to this list.

1.2 Previous work on *hän* and *tämä*

Initially, word order and grammatical role were not considered as separate influences of accessibility. One of the reasons for this is that, in English, there is no clear distinction between word order and grammatical role. In most sentence types and regular discourse, the default word order of subject-verb-object (SVO) is used. Then, the subject is always the first-mentioned NP, and the object is always the last-mentioned NP, so word order and grammatical role are indistinguishable.

Finnish, on the other hand, allows a lot more variation in the order of grammatical arguments. Under the right discourse conditions, all six orders (SVO,SOV,VSO,VOS,OSV,OVS) are grammatically valid, although some are quite rare (Vilkuna, 1989, p. 9). As Finnish has a complete case marking system that assigns grammatical role, word order variation does not generally cause ambiguity and occurs commonly in natural language (cf. Hakulinen, Karlsson and Vilkuna (1980) for quantitative data). This allows researchers to study the effects of word order and grammatical role as separate factors.

This makes Finnish, and the *hän/tämä* distinction in particular, an interesting research topic. Halmari (1994) was the first to do a corpus study investigating the relationship between grammatical role and different anaphor forms in Finnish. She found that the proposed relationship between the NP accessibility and the Givenness hierarchies also held for Finnish: referential forms lower in Givenness were used more often for highly accessible antecedents, and vice versa. However, Halmari did not take word order into account, barring her from drawing any definite conclusions about the influence of grammatical role by itself.

Kaiser (2000) did a similar corpus study, narrowing the scope by only taking the *hän/tämä*-variation into account, and additionally looking at the effect of clause type (matrix or subordinate). She found that *hän* is used more for subjects and *tämä* is used more often to refer to objects. Additionally, she found that *hän* often has antecedents in main clauses, and that *tämä* has more antecedents in subordinate clauses. This might be explained by competitor effects: when there is an antecedent in a subordinate clause, it is more likely that there is also a possible antecedent in the main clause, requiring the use of *tämä* to disambiguate between the two. Kaiser did not look into word order in the corpus either, as there were no instances of object-subject word order (OSV, OVS or VOS) in her corpus. She did carry out a small-scale sentence-continuation experiment, which suggested that linear word order has some influence on anaphor form, but it was too limited in scope to draw any definite conclusions.

Later work has focused mainly on interpretation of anaphor form, as opposed to choice of anaphor form, using both sentence-completion and eye-tracking experiments. Järvikivi et al. (2005) looked into the processing of referential *hän*, using eye-tracking. They found that both word order and grammatical role had a significant influence on interpretation: *hän* was most often interpreted as referring to the subject and first-mentioned argument. However, in non-canonical OVS-order sentences, they found that there was no such preference for first-mentioned arguments, showing that grammatical role is the most important linguistic cue when interpreting *hän*. Sentence-completion and eye-tracking experiments by Kaiser, Runner, Sussman and Tanenhaus (2005) and Kaiser and Trueswell (2008) confirmed this, finding that *hän* is mainly interpreted as referring to subjects, and *tämä* is mainly interpreted as referring to post-verbal or second-mentioned referents, objects and discourse-new entities.

A different approach was taken by Seppänen (1998), who investigated the referential properties of *hän* and *tämä* in spoken discourse. She found that *tämä* is generally used for conversation participants who have been speaking earlier. Regarding *hän*, the tendencies were less clear, but she concluded that it was used in cases where the speaker expressed the referent's viewpoints or opinions.

1.3 Case marking, subject and object in Finnish

Given the amount of research on the *hän/tämä*-variation, it is remarkable that all have focussed on regular-type sentences that have canonical case marking. For example, the fact that the majority of non-canonical partitive case subjects occur in a non-default word order (Hakulinen, Karlsson & Vilkuna, 1980), might have been of interest for researchers looking into word order effects. A similar link between case marking and word order was found by Hyönä and Hujanen (1997), who concluded that syntactic processing of sentences is considerably easier when overt case marking is used, but only in sentences using a non-SVO word order.

Grammatical role, too, is closely tied in with case marking, as it is the main method of indicating the grammatical role of an argument. However, this is not a one-to-one relationship, as the same grammatical role may be marked using different cases (the object mainly occurs in the partitive, genitive, nominative and accusative cases, for example, cf. Section 1.3.3). This brings up the following question: does different case marking influence grammatical role perception and by that, accessibility? And is this reflected in its co-referential properties? These are the main questions we are concerned with in this paper, and which we will try to answer in the corpus study.

The idea that non-canonical sentence types and case marking influence the subject- and object-hood of an argument is not completely new. Suggestions in this direction have already been made by Karlsson (1983), who found that, between the prototypical or canonical subjects and objects in Finnish, there is a grey area, which he describes as an intersection between subject and object. The partitive argument in an existential sentence, for example, deviates from a prototypical subject in so many aspects, that it becomes syntactically neutralized, being somewhere between subject and object. It would be interesting to see how this interacts with co-reference, which, in current theory, deals only with strictly separate notions of subject and object.

Before we can start answering these questions, we will have to take a closer look at the Finnish case system, its non-canonical case marking, irregular sentence types and the nature of the Finnish subject and object. The purpose of this is to identify those sentence types and arguments which are most likely to differ from regular sentence in their accessibility properties, and thus should be selected for the corpus study. Existing literature on non-canonical marking in Finnish will be discussed to find potentially interesting sentence types. First, a quick overview of the Finnish case system will be presented, as necessary background information for a discussion of non-canonical marking

1.3.1 The Finnish case system

There are 15 different cases in Finnish, which are traditionally divided into three categories: the grammatical cases, the locative cases and the marginal cases (Hakulinen, Vilkuna et al., 2004, p. 1173-74). The grammatical cases are the nominative, the genitive, the partitive and the accusative[3]. The locative cases consist of the illative, inessive, elative, allative, adessive, ablative, essive and translative, and are generally used to indicate locations or abstract relationships, like prepositions in English. The marginal cases are the comitative, abessive and instructive. These are only rarely used, hence the name, and are not relevant for this study.

We are mostly concerned with the four grammatical cases, as these mainly function to assign grammatical roles. Generally, the subject is marked as nominative (the 'unmarked' case, with no case ending in the singular), while the object is interchangeably marked as either partitive, genitive or accusative. Additionally, the genitive is used to mark possession. Statistics mostly seem to support this view: the subject is in the nominative case in approximately 85% of the time, in the partitive case approximately 8% of the cases and as genitive in approximately 7% of cases (Hakulinen, Vilkuna et al., 2004, p. 1181-82).

The statistics for the object are less clear, with it being in the partitive approximately 58% of the time, in the nominative 21% of the time, in the genitive case 20% of the time, and in the accusative case 1% of the time (Hakulinen, Vilkuna et al., 2004, p. 1181-82). As the nominative case is the second most frequent object case, it seems odd that it is not considered a regular object marking case. The main reason is that the partitive and genitive are used predictably and interchangeably to

[3]Multiple definitions of the accusative case exist. Here, the accusative is taken to be restricted to the *-t* marked form of personal pronouns. For an overview, see Section 1.3.2

mark certain properties of the object, while the nominative object only occurs in sentences where the subject is not in the nominative case. One important exception to this is the case of plural objects, which are either in the nominative or partitive. We consider the plural and singular separately, and therefore regard the nominative as the regular case for plural objects.

1.3.2 The Finnish case discussion

Previous investigations into non-canonical marking in Finnish have reached different conclusions (Itkonen, 1974; Hämäläinen, 2005, for example). One cause of these differing views is the view on the Finnish case system that is taken. Two competing views exist, as there are two main ways of classifying Finnish grammatical cases: the traditional view, which has a large role for the accusative case, and an alternative view, which limits the accusative case to personal pronouns. These differences are caused by the fact that the accusative is a heterogeneous case in the traditional view, having the alternating case endings *-n* or *-ø* for singulars, *-t* for plurals and *-t* for both singular and plural personal pronouns. As it happens, the *-n* ending is identical to that of the genitive singular, the *-ø* is identical to the unmarked nominative singular and the *-t* is identical to the nominative plural ending. This has lead some scholars to analyse these forms as the nominative or genitive case, instead of the accusative, restricting the accusative to only the *-t* suffix of personal pronouns. Views on this topic differ strongly, and will not be discussed in detail here (for discussion, see Sands and Campbell (2001), Vainikka (1989), Toivainen (1993), Maling (1993), Reime (1993), Bielecki (2009), Kiparsky (2001)).

For this study, the details of the different case analyses are not relevant. However, a good working definition of case categories is necessary for a definition and discussion of what non-canonical marking entails. Essentially, the different viewpoints on the accusative originate from different takes on the concept of case. This difference boils down to what will be called the *morphological viewpoint* and the *syntactic viewpoint* here. In the morphological viewpoint, cases are defined by their morphological markings, regardless of the syntactic positions in which they usually occur. From such a viewpoint, all words marked in a case with *-n* are called the genitive, no matter if the case is used to mark possession, objecthood or a relation to an adposition. In the syntactic viewpoint, cases are defined by both their morphological markings and their syntactic function. From this viewpoint, a *-n* marked word should be analysed as genitive case if it is a possessor, but should be seen as accusative case if it is an object. This links cases to certain functions, in which it can take different forms, such as the accusative *-n, -t* and *ø*.

In this study, we will take the morphological viewpoint, as this is the most compatible with the idea of non-canonical marking. Non-canonical marking, by definition, implies that there is no fixed relationship between syntactic function and case, as non-canonical marking denotes those situations where a different case than usual is used to mark a certain function. Therefore, only the *-t* marking on personal pronouns will be analysed as accusative, while other suffixes will be analysed as genitive or nominative, according to their morphological endings. Mentions of case in previous sections have reflected this stance, already. Nevertheless, as grammatical role will be recorded alongside case in the corpus study, the case categories corresponding to the syntactic viewpoint could easily be derived from the data, if necessary.

1.3.3 Types of non-canonical case marking

If non-canonical case marking of grammatical arguments in Finnish is defined as subjects that are not marked nominative and objects that are not marked genitive or partitive, we find many instances of what seems to be non-canonical marking. For example, there are genitive-marked subjects in necessive constructions, nominative objects in various sentence types and elative subject-like NPs in 'feeling'-constructions. Previous overviews of this topic exist, and Sands and Campbell (2001) and Hämäläinen (2005) identify and analyse most, if not all, of these possible instances of non-canonical marking. In this section, we will look at different instances of non-canonical marking and see whether they might have any relation to the matters of subjecthood, objecthood, accessibility and co-reference.

Following Sands and Campbell, the following types of possible non-canonically marked constructions and sentence types can be identified: partitive 'subjects' in existential sentences (Ex-

ample 2) and causatives of feeling (Examples 3a-b), genitive subjects in necessive, infinite and 'for'-constructions (4a-d), nominative objects in imperative, non-finite and impersonal/passive constructions (5a-c), locative object constructions (6a,b), locative marking with perception verbs (6c) and adessive marking of 'have'-constructions (6d). All example sentences are taken from Sands and Campbell (2001), unless noted otherwise.[4]

The most-widely studied and discussed type of the ones mentioned above is surely the partitive-marked NP in existential constructions, such as *lapsia* in Example (2) (cf. e.g. Helasvuo (1996), Tiainen (1997), for an overview of literature). Most discussion has focussed on whether the partitive marked NP in such constructions should be seen as object or subject, or as a neutralization of the two. Regardless of whether this counts as non-canonical marking (which it would be if it were a subject, but not otherwise), it is certainly relevant for the current study. Many arguments have been made for both views, but, assuming that it has characteristics of both the object (Wiik, 1974) and subject (Hakulinen, 1982), we conclude for now that it falls somewhere in between, as Karlsson (1983) suggests. If this conclusion is right, its co-referential properties should be between that of subjects and objects, and the current study should show whether this is the case.

(2) *Ulkona leikki-i laps-i-a.*
outside play-3SG child-PL-PAR
'Outside children are playing.'/'There are children playing outside.'

Another type of partitive argument occurs with the so-called causatives of feeling, e.g. *minua* in Example (3). Intuitively, regarding word order and the translation, the partitive marked NP resembles a subject, and therefore seems to be non-canonically marked. However, this initial analysis is false, as Siiroinen (2001) and Sands and Campbell (2001) show that this argument is actually a topicalized object in a subject-less sentence. Both topicalization and genuinely subject-less sentences are common in Finnish, and this analysis can be made explicit by looking at Example (3b). This shows that this NP is a canonically marked object, and is not relevant for the current study. However, as the object is topicalized in such sentences, such sentences might be useful for research on non-default word order.

(3) a. *Minu-a pelotta-a.*
1SG-PAR frighten-3SG
'I am frightened (by something).'/'Something frightens me.'

b. *Minu-a pelotta-vat koira-t.*
1SG-PAR frighten-3PL dog-NOM.PL
'I am frightened by the dogs.'/'The dogs frighten me.'

The other main category of non-canonically marked subjects is that of genitive subjects that occur in necessive, non-finite and so-called 'for'-constructions. As before, the key question is whether the analysis of these arguments as subjects is a correct one. In the case of the necessive construction (4a), the answer is not straightforward. According to Sands and Campbell (2001), experts on Finnish have analysed this argument as either subject of the necessive verb, subject of the infinitive (and the infinitive as subject of the necessive verb) or a dative-adverbial. Sands and Campbell show, by contrasting the genitive form with a true dative-adverbial form in the allative case, that this last view is incorrect. This makes the genitive NP either a non-canonically marked subject of a finite necessive verb, or a subject of an infinite verb form, for which the default case is the genitive. Whether the genitive marking of subjects of non-finite verbs constitutes non-canonical marking depends on the exact definition of non-canonical marking that one maintains (cf. Hämäläinen (2005), p.41-44). Arguments for both remaining analyses can be made, but the genitive NP is a non-prototypical subject in any case, and is therefore of interest to the current study.

The analysis of the genitive argument as the subject of a non-finite verb can be applied, undisputedly, to other sentences with non-finite verb forms, such as (4b) and similar clauses. Similarly, the focus of the corpus study should be extended to include these sentence types. The third type of genitive subject, in the 'for'-construction (4c) is a different matter. Although these forms are

[4] Additional abbreviations used: 1 - first person, 2 - second person, ADE - adessive case, ALL - allative case, ELA - elative case, INF - infinitive, ILL - illative case, PARTIC - participle, PASS - impersonal/passive, QU - question marker

sometimes interpreted as being subjects, when compared to equivalent sentences such as (4d), it quickly becomes clear that the correct analysis is that of them being dative-adverbials. As such, these forms are expected to behave like regular oblique NPs in terms of co-referential properties and are not selected for the corpus study.

(4) a. *Sinu-n pitä-ä men-nä.*
2SG-GEN must-3SG go-INF
'You must go.'

b. *Luule-t-ko minu-n tietä-vän tämä-n?*
think-2SG-QU 1SG-GEN know-PARTIC this-GEN
'Do you think I know this?'

c. *Minu-n tul-i rakko jalka-an.*
1SG-GEN come-3SG.PST blister.NOM foot-ILL
'I got a blister on my foot.'

d. *Minu-lle tul-i rakko jalka-an.*
1SG-ALL come-3SG.PST blister.NOM foot-ILL
'I got a blister on my foot.'

In the case of object marking, we consider every object that is not in the partitive, genitive or accusative case a possibly non-canonically marked argument. Then, the most frequent group of non-canonically marked objects is that of nominative objects, which can occur in different sentence types, illustrated in Examples (5a-c). These sentence types are the imperative (5a), non-finite constructions,here combined with an imperative verb (5b) and the passive or impersonal construction, which is used to put the emphasis on the action and patient, by avoiding explicitly mentioning the agent (5c). Despite their differences, these constructions have one thing in common: they lack a nominative subject. Similarly, the object of a necessive construction, is in the nominative (or the partitive, depending on the type of object, negation, among other things). An 'explanation' for this would be, as the function of object marking is to distinguish it from the subject, there is no need to explicitly mark the object for contrast, when there is no nominative subject, leaving it in the 'default' nominative case, such as in (5a,c). In cases with a genitive subject, one could say that the nominative case is still 'free' to be assigned to the object, and, combined with the partitive, contrasts clearly with the subject (5b). This intuition seems to be confirmed by the fact that not all non-finite complements take nominative objects, but only those that have a main sentence without a nominative subject (Sands & Campbell, 2001, p. 280-282).

With respect to co-referential properties, a difference is to be expected between nominative objects in subject-less sentences and those in sentences with non-nominative subjects. In the first case, the object might behave more like a regular subject, as it is in the default subject case, and because the referential expression does not have to disambiguate between object and subject antecedents. As such, references to arguments in sentences such as (5a,c) are of particular interest for the current study. For the sentences with a non-nominative subject, there is no such clear expectation, as these are more similar to prototypical sentences, having the same subject/object-contrast, only with reversed case markings. This makes these sentences very useful to study the effect of case marking exclusively, as this is the only irregular aspect of their subjects and objects. Therefore, the objects of these sentences will be selected for the corpus study, as well as their subjects, mentioned above.

(5) a. *Sano se uudestaan.*
say.IMP it.NOM again
'Say it again!'

b. *Anna Marja-n osta-a auto.*
let.IMP Marja-GEN buy-INF car.NOM
'Let Marja buy a car.'

c. *Kirja pan-tiin pöydä-lle.*
book.NOM put-PST.PASS table-ALL
'The book was put on the table.'

A very different category of non-canonical case marking concerns those grammatical arguments that are not marked with grammatical cases, but with locative cases, such as the illative, elative and adessive. One type of this, the locative marking of what seem to be direct objects, occurs across all sentence types, and depends on the verb. The verb *pitää*, for example, always takes a complement in the elative case, as in Example (6a). The complements in such sentences are very much like direct objects, as they bear the direct effect of the verb and are required by the transitive verb. Nevertheless, they are only considered objects by some, while others see them as regular oblique NPs or refuse to label them directly (Sands & Campbell, 2001, p. 286-87). An object analysis seems plausible though, for example in sentences such as (6b), in which a locative marked complement can also be marked in the canonical partitive case without significant change in meaning. As such, we expect these arguments to behave like objects with respect to co-reference. To see whether this expectation holds true, such arguments will be of particular interest for the corpus study.

The other two sentences differ from Example (6a,b) in that their locative marked arguments more closely resemble subjects. As for the first type, the locative argument of perception verbs (6c), it may seem unlikely that the locative argument is a subject, but in translation, such an analysis may be suggested, if translated as 'I think the picture is crooked'. Nevertheless, on closer inspection, it quickly becomes clear that this locative marked NP is a complement that can be added to or removed from a well-formed matrix clause without problem.

The last example sentence, (6d), presents a more interesting challenge for analysis, as its adessive NP is semantically close to a subject, and its sentence structure is close to the existential sentence of Example (2). In translation, the adessive NP translates to the subject of a 'have'-verb, and semantically it has the role of possessor, which is more subject- than object-like. A second argument for a subject analysis would be that, if occurring with a personal pronoun, the personal pronoun takes the accusative case, as if it were an object. Alternatively, when this construction is seen as a subtype of the existential sentences, the possessed NP can be seen as subject-like.

As no straightforward answer can be provided, we follow Sands and Campbell (2001) here in the conclusion that no clear label should be applied to this argument. This renders this sentence type especially interesting for the study of co-reference, as it possesses two arguments which are neither clear subjects nor objects (cf. discussion of Example (1)), and thus allowing no predictions of co-referential properties.

(6) a. *Hän pitä-ä sii-tä.*
3SG.NOM like-3SG it-ELA
'S/he likes it.'

b. *Koputa sinä ove-a/ove-en/ove-lle.*
knock.IMP 2SG.NOM door-PAR/door-ILL/door-ALL
'Knock on the door!' (Hakulinen & Karlsson, 1979, p. 176)

c. *Minu-sta kuva on vinossa.*
1SG-ELA picture.NOM be.3SG askew
'To me, the picture is crooked.'

d. *Juka-lla ol-i avaime-t.*
Jukka-ADE be-3SG.PST key-NOM.PL
'Jukka had the keys.'

1.4 Pilot study: examining existing data

Before starting the corpus study, it would be welcome to have some evidence of the existence of an influence of case marking on co-referential properties. Luckily, the author has collected a set of co-reference data for Finnish for a previous Bachelor's thesis. These data were gathered by inquiring native speakers about their interpretations of referential expressions. No data on case marking was recorded then, but was added for this study. The dataset is small, and contains 236 co-referential pairs, 139 with *hän* as anaphor and 97 with *tämä*. There was no focus on the non-prototypical sentence types discussed in Section 1.3.3, and as a result there are only few instances of non-canonical case marking in the data. Nevertheless, we will examine these few instances for indications of the current hypothesis' correctness.

Antecedent Case	Gramm. Role	*hän*	*hän*-%	*tämä*	*tämä*-%	Total
Ablative	Other	1	25.0	3	75.0	4
Adessive	Subject	1	50.0	1	50.0	2
Allative	Other	2	16.7	10	83.3	12
Elative	Subject	1	100.0	0	0.0	1
	Object	1	33.3	2	66.7	3
	Other	2	40.0	3	60.0	5
Genitive	Subject	2	66.7	1	33.3	3
	Object	1	16.7	5	83.3	6
	Other	18	62.0	11	38.0	29
Illative	Object	4	100.0	0	0.0	4
	Other	2	100.0	0	0.0	2
Inessive	Other	0	0.0	1	100.0	1
Nominative	Subject	95	88.0	13	12.0	108
	Object	2	40.0	3	60.0	5
	Other	7	25.0	21	75.0	28
Partitive	Subject	2	100.0	0	0.0	2
	Object	3	17.6	14	82.4	17
	Other	1	25.0	3	75.0	4
Total:		145	61.4	91	38.6	236

Table 2: Pilot data, split out by antecedent case and grammatical role.

Before looking at the actual data, reproduced in Table 2, an often over-looked distinction has to be made between two ways of interpreting it. One way is to look at the preferences of certain anaphor forms, e.g. seeing that 74% of *hän*'s antecedents have the nominative case. This expresses the probability $P(nominative|hän)$, which could be seen as aiding in language comprehension: given a certain anaphoric expression, e.g. *hän*, it might be useful to know that the antecedent is most likely one that has the nominative case. The reverse probability, $P(hän|nominative)$, relates to language production. This is the information that is the goal of this study, which investigates the influence of case marking on accessibility, and thus on the probability of producing *hän* or *tämä*.

A first examination of the data from this perspective shows the expected pattern: summed over all grammatical roles, nominative arguments are usually referred to by *hän* (74% vs. 26% *tämä*), partitive arguments prefer *tämä* (74% vs. 26% *hän*), as do locative-marked arguments (76% vs. 24% *hän*). The only surprise is found with genitive arguments, which seem to have no clear preference (55% *hän* vs. 45% *tämä*). The general pattern is unsurprising, as *hän* prefers subjects, which usually take nominative case, and *tämä* prefers non-subjects, which usually take non-nominative cases.

For a more detailed analysis, we look at the interplay between grammatical role, case and anaphoric expression. As expected, we see only few non-canonical markings (which renders calculations of percentage data unnecessary): 2 partitive subjects, 2 adessive subjects, 3 genitive subjects, 7 locative objects and 5 nominative objects. The locative objects clearly prefer *tämä* (1-6), like other objects, but objects in nominative case do not (2-3), which could be interpreted as confirming our expectations. As for subjects, adessive (1-1), and genitive (2-1) show similar non-canonical tendencies, while partitive subjects (2-0) do not. A final interesting thing, the unclear preferences of genitive antecedents, can be explained by looking at the oblique (non-subject, non-object) category, where genitives are split (62% vs. 38%). A quick analysis of these antecedents indicates that *hän* is used mostly for genitive possessors, while *tämä* is used for other genitive complements.

In conclusion, the pilot study shows indications that the hypothesis is correct, and that case marking does indeed influence co-reference. However, the size of the dataset is simply too small to draw any meaningful conclusions. The proposed targeted corpus study should deal with this problem. Details of the corpus and its annotation are presented in Section 2.

2 Methods

The current corpus study is based on both newspaper texts and novels from the Finnish Text Collection of the CSC - IT Center for Science[5]. The novels were taken from the Otava 1998/99 and the WSOY 1996/97/98 subcollections. Not all material from these collections was used, only novels (as opposed to informative books and other types) were selected. For a list of selected novels, see the 'Sources'-section of the bibliography. The newspaper texts were taken from the Aamulehti 1999-01 sub-collection. In total, the used novels contain approximately 0.66 million tokens and the newspaper texts contain about 1.30 million tokens, making for a total of 1.96 million tokens in the whole corpus. The use of both newspaper texts and novels provides a broader coverage of written Finnish than using only one type and allows for comparison of the two genres.

In order to extract referential pairs containing *hän* or *tämä* from the corpus, a simple exhaustive search method was used. After preprocessing steps, such as conversion to plain text and to a one-sentence-per-line format, all instances of *tämä* and *hän*, in all their case inflections, were extracted, together with one succeeding and three preceding sentences. This was all done using Unix command-line tools. Then, a self-written Python program was used to present and annotate each extracted fragment. All annotation was done manually by the author. Naturally, all wrongly extracted excerpts, such as those containing a non-referential, determinative use of *tämä*, or *häntä* in the meaning of 'tail', were discarded. In addition, all referential pairs where *tämä* referred to an inanimate antecedent were discarded, as reference to inanimate arguments is not a situation in which both *hän* and *tämä* can be used. Of each referential pair, the following properties were recorded: root form, case and part-of-speech of both anaphor and antecedent, sentence argument order (SOV, OVS, etc.), non-canonical sentence type (cf. Section 1.3.3) and matrix/subordinate clause type for the sentence(s) containing the anaphor and antecedent, and animacy, gender (human, animal or inanimate) and presence of a competitor (a second potential antecedent) for the antecedent. The competitor annotation was split up in two parts: the occurrence of an intervening competitor (a competitor between antecedent and anaphor) and the occurrence of a competitor in the excerpt as a whole.

The goal was to collect approximately 400 co-referential pairs, of which half with antecedents in regular-type sentences and half with antecedents in non-canonical-type sentences. Additionally, these were to be equally distributed over newspaper texts and novels. In order to achieve these numbers, all co-referential pairs with antecedents in non-canonical-type sentences were collected from the whole corpus. Regular-type sentences occur much more frequently, so these were collected from a smaller (10-15% of the original), randomly selected subset of the material. The subsets were taken from the original material in such a way that the distribution of styles and authors was equivalent to that of the whole set, in order to allow for valid comparisons with the non-canonical-type-annotations, which draw on all the material.

3 Results

An overview of co-reference pairs annotated during the corpus study is presented in Table 3. We see that this matches the goal, as 460 co-referential pairs were annotated, equally distributed over both genres and sentence types (i.e. with and without non-canonical marking). Additionally, this basic overview shows a difference between genres, as there is a significantly higher percentage of *tämä*-references in the novels than in the newspaper texts (16.7% vs. 9.3%). Also, there seems to be a slightly higher use of *tämä* when referring to arguments in 'irregular' sentences (15.4% vs. 11.2%).

More importantly, Table 3 shows a large difference between *hän* and *tämä*: the vast majority of references use *hän*, and the use of *tämä* is relatively minor (87.0% vs. 13.0%). This is an important baseline for the rest of the analysis, as it suggests that the counts of specific combinations of *tämä* in specific sentence types or case markings will be low. Also, this implies that it makes more sense to speak of a relatively lower or higher preference for *hän*, instead of a preference for either *hän* or *tämä*.

In order to get an idea of whether case marking influences anaphor form, Table 4 presents the interaction between anaphor form and antecedent case. In general, the patterns are in accordance

[5]More information on the Finnish Text Collection (in English) can be found at the website of the CSC, www.csc.fi/english/research/software/ftc.

Genre	Sentence type	*hän*	*hän*-%	*tämä*	*tämä*-%	Total
Novel	Regular	115	86.5	18	13.5	133
	Irregular	80	79.2	21	20.8	101
Newspaper	Regular	115	91.3	11	8.7	126
	Irregular	90	90.0	10	10.0	100
Total:		400	87.0	60	13.0	460

Table 3: Number of co-referential pairs annotated, by genre, antecedent sentence type and anaphor type.

with expectations. Locative case-marked arguments have an above-average preference for *tämä*, with the exception of adessive- and elative-marked arguments, which were mostly subjects, and as such prefer *hän*. Furthermore, nominative-marked arguments are almost exclusively referred to by *hän*, where genitive- and partitive-marked arguments have *tämä* as their anaphor more often, albeit still only in 20-30% of cases.

Antecedent Case	*hän*	*hän*-%	*tämä*	*tämä*-%	Total
Ablative	3	50.0	3	50.0	6
Adessive	47	94.0	3	6.0	50
Allative	8	53.3	7	46.7	15
Elative	16	94.1	1	5.9	17
Genitive	73	79.3	19	20.7	92
Illative	4	30.8	9	69.2	13
Nominative	217	97.3	6	2.7	223
Partitive	32	72.7	12	27.3	44
Total:	400	87.0	60	13.0	460

Table 4: Number of co-referential pairs annotated, by antecedent case and anaphor form.

3.1 Sentences with canonically marked arguments

To confirm that there is indeed an effect of case marking, it has to be viewed in the context of grammatical role and sentence type. Due to the many different sentence types with non-canonical marking, the interaction between grammatical role, case marking and anaphor form in the so-called 'regular' sentences, as presented in Table 5, will be considered first to get a baseline. This shows patterns that are in line with expectations. Subjects are overwhelmingly referred to with *hän*, where other arguments show a more balanced preference. Genitive-marked objects show a higher preference for *tämä* than partitive-marked objects.

Nominative-marked objects are a special case, as these fall under non-canonical case marking. Here, the two nominative-marked objects were found in so-called 'generic' sentences, sentences with no overt subject. One of these is reproduced in sentence (7). As these were not one of the non-canonical sentence types identified in Section 1.3.3, they were annotated as regular sentences. Interestingly, these nominative-marked objects are referred to by *hän*, which may indicate an influence of case marking, but which is most likely due to generic sentences having no subject, removing the need to disambiguate between possible antecedents.

(7) **Tommi Pohjola** pit-i otta-a kiinni, jotta *hän* [...]
Tommi.Pohjola.NOM have.PST.3SG take-INF shut so.that s/he.NOM [...]
'**Tommi Pohjola** has to be grabbed hold of, so that *he* [...]'

Indirect objects, marked with allative case, show a clear preference for *tämä*, but the number of occurrences in the corpus is too low to draw any definite conclusions. Possessors, in the genitive case, do not suffer from this problem, with 38 occurrences. Surprisingly, given the low position of

possessors in the NP accessibility hierarchy, these show a strong preference for *hän*. A tentative explanation for this is that these are often possessors of inanimate things, making *hän* the least ambiguous referring expression. Arguments classified as having the 'other'-role, which are often arguments related to locations, modifiers and complements, do not show clear patterns, partly due to the low number of occurrences in the corpus and their varied nature.

Grammatical Role	**Antecedent Case**	***hän***	***hän*-%**	***tämä***	***tämä*-%**	**Total**
Subject	Nominative	169	99.4	1	0.6	170
Object	Genitive	4	50.0	4	50.0	8
	Nominative	2	100.0	0	0.0	2
	Partitive	13	65.0	7	35.0	20
Indirect Object	Allative	1	20.0	4	80.0	5
Possessor	Genitive	32	84.2	6	15.8	38
Other Role	Allative	1	100.0	0	0.0	1
	Elative	3	75.0	1	25.0	4
	Genitive	5	50.0	5	50.0	10
	Illative	0	0.0	1	100.0	1
Total:		230	88.8	29	11.2	259

Table 5: Number of co-referential pairs annotated, by antecedent case marking, grammatical role and anaphor form.

3.2 Sentences with non-canonically marked arguments

Whereas case marking by itself does not seem to have a strong effect on accessibility and anaphor form, it might do so in non-canonical sentence types. In Table 6, the data for non-canonically marked arguments in 6 types of sentences is presented, split by grammatical role and case marking of the antecedent. Accidentally, 9 co-reference pairs with a canonically marked indirect object or possessor were annotated. These have been left out here. The first important observation is the low number of occurrences of some sentence types. For example, in almost 2 million words, only three times a reference is made to the object of an imperative sentence using *hän* or *tämä*. Unfortunately, this number is too low to draw any significant conclusions with regards to anaphor form variation. The same holds for objects of necessive sentences and other non-finite constructions, locative objects and objects of so-called 'have'-constructions. Additionally, we see low counts (< 5) for partitive NPs in existential sentences, adessive-marked subjects of necessive and other non-finite constructions and allative- or elative-marked objects in regular sentences. Nevertheless, some interesting patterns seem to be present in the low numbers. We see that for the e-NP, the object in necessive constructions and the object of other non-finite constructions, case marking seems to be a relevant factor: partitive-marked arguments are referred to with *tämä* in 1/1, 2/2 and 1/2 cases, while non-partitive arguments show a 100% *hän*-preference, which suggests a more balanced preference for partitive-marked antecedents.

Considering the sentence types and arguments with more occurrences, we see that all show an overwhelming preference for *hän*, except for illative- and ablative-marked objects, which show a majority-preference for *tämä*. Additionally, we see that the subjects of necessive sentences behave exactly like regular subjects, in that they show an almost complete ($> 98\%$) preference for *hän*. The other non-canonically marked subjects, in the elative case, have the same preference. Genitive-marked subjects of other non-finite constructions show a slightly lower *hän*-preference (86.4%) than regular subjects. The remaining category of subject-like arguments, the adessive-marked arguments of 'have'-constructions, also show a strong *hän*-preference, but again slightly lower than regular subjects (93.7%). The only remaining category is the most numerous, and is that of the objects of passive or impersonal sentences. Perhaps surprisingly, these show more subject- than object-like preferences (90.7% *hän*). The preference is stronger for partitive- than for nominative-marked objects, which is opposite to the pattern that other objects showed.

Sent. Type	Gramm. Role	Antec. Case	*hän*	*hän*-%	*tämä*	*tämä*-%	Total
Existential	Subject (e-NP)	Nominative	10	100.0	0	0.0	10
		Partitive	0	0	1	100.0	1
Necessive	Subject	Adessive	1	100.0	0	0.0	1
		Genitive	11	100.0	0	0.0	11
	Object	Nominative	2	100.0	0	0.0	2
		Partitive	0	0	2	100.0	2
Other non-finite	Subject	Adessive	1	100.0	0	0.0	1
		Genitive	19	86.4	3	13.6	22
	Object	Ablative	1	100.0	0	0	1
		Genitive	2	100.0	0	0.0	2
		Partitive	1	50.0	1	50.0	2
Imperative	Object	Elative	1	100.0	0	0.0	1
		Nominative	1	100.0	0	0.0	1
		Partitive	1	100.0	0	0.0	1
Passive	Object	Nominative	32	88.9	4	11.1	36
		Partitive	17	94.4	1	5.6	18
Locative Case	Subject	Elative	8	100.0	0	0.0	8
	Object	Ablative	2	40.0	3	60.0	5
		Allative	1	33.3	2	66.7	3
		Elative	3	100.0	0	0.0	3
		Illative	4	33.3	8	66.7	12
Have-construction	Subject	Adessive	45	93.7	3	6.3	48
	Object	Nominative	1	100.0	0	0.0	1
Total:			164	85.4	28	14.6	192

Table 6: Number of co-referential pairs annotated, by sentence type, grammatical role, antecedent case and anaphor form (subjects and objects only).

4 Discussion

The investigation of the complex interaction between grammatical role, (non-canonical) case marking and anaphor form using a corpus study has proven to be a difficult task. As with any corpus study, the lack of control over additional factors and sentence specifics makes it difficult to get a clear image of the influence of the factors under investigation. Additionally, the required material not occurring in large enough amounts in the corpus forbids definitive conclusions about specific combinations of factors. More concretely put: it will be difficult to get any definite answers with regards to the interaction between grammatical role, case marking and anaphor form from this study, as some sentence types and case markings are very rare and because additional factors, such as whether or not there is a competitor in the sentence, cannot be controlled for and might obfuscate the results. An experimental study could control for these factors and yield more insight, but has the drawback of having artificially designed materials, whereas the corpus study has the benefit of using only naturally produced written language, albeit redacted and revised.

Despite the downsides of the current method, plenty of observations can be made on the data. For one, the overall statistics show that, in regular written text, the use of *tämä* to refer to people is rare, and even more so in newspaper texts than in novels. This is most likely explained by an inherent difference between the two text genres: novels generally revolve around interactions between two or more characters, where newspaper articles are short, segmented stories which have only one or a few main characters in it. Additionally, this touches on one of the major points about the *hän*/*tämä*-variation, namely that it is only applicable when there are two or more animate entities in the discourse.

4.1 The influence of competitors

The importance of the presence of multiple possible referents in the discourse is illustrated by the data gathered on competitors, (Table 7), which shows the usage of *hän* and *tämä*, as influenced by the presence of a competitor in the excerpt used for annotation (i.e. in the current or 3 previous sentences) and the presence of a competitor between the anaphor and antecedent (an intervening competitor). There is a clear pattern: almost all occurrences of *tämä* are in excerpts where there is only a non-intervening competitor, i.e. the antecedent is the last possible referent, e.g. sentence (8a). Even under these conditions, *hän* is still used in the majority of cases, but usage of *tämä* is a lot more substantial.

Competitor	***hän***	***hän*-%**	***tämä***	***tämä*-%**	**Total**
Intervening competitor	48	98.0	1	2.0	49
Competitor, not intervening	159	74.6	54	25.4	213
No competitor	193	97.5	5	2.5	198
Total:	400	87.0	60	13.0	460

Table 7: Number of co-referential pairs annotated by competitor type

Before looking at the *hän*/*tämä*-variation under these conditions in more detail, the uncommon usage of *tämä* in the other two conditions will be considered. Even though it seems like *tämä* is only used when there is need to disambiguate between multiple possible antecedents, it is not wholly impossible for it to occur when there is no possible ambiguity. In the current corpus, we find 5 instances of this. Looking at these excerpts in more detail, it becomes clear that these are genuine usages of *tämä* in cases where *hän* would be a more common and ambiguous alternative. One such sentence is reproduced in (8b), which had no possible antecedents in its preceding text. Of the 5 antecedents, 1 is used in a comparison, 2 are objects and 2 are possessors. As all are non-subjects, this suggests that, sometimes, factors such as grammatical role can override the default use of *hän*, even when *hän* would be the more common and equally clear alternative.

(8) a. **Ruotsalainen** hämmästel-i **Blatteri-n** suitsutus-ta uudelle idealle ennen kuin
Swede.NOM wonder-PST.3SG Blatter-GEN support-PAR new.to idea.to before that
tämä [...]
this.NOM [...]
'**The Swede** was astonished at **Blatter's** support for the new idea before *he* [...]'

b. [...] komitea kalust-i KOK:sta tiistaina eronne-en
[...] committee.NOM furnish.PST.3SG IOC.from Tuesday.on resigned-GEN
suomalaisjäsen-en **Pirjo Häggmani-n** ja *tämä-n* silloise-n aviomiehe-n
Finnish.member-GEN Pirjo.Haggman-GEN and this-GEN then-GEN husband-GEN
Bjarne-n asuintalo-a.
Bjarne-GEN house-PAR
'[...] the committee furnished the house of **Pirjo Häggman**, the Finnish IOC member who resigned on Tuesday, and *her* then husband Bjarne.'

The other unusual *tämä*-reference is the one with an intervening competitor in the sentence, see (9a). We see that there are two possible referents, *miehen*, 'the man' and *vaimon*, 'the wife'. Obviously, real-world knowledge prevents the interpretation where 'the wife' is the antecedent, but apart from that, 'the wife' seems the most probable referent. Based on its grammatical role, it has low accessibility, being part of the complement, while 'the man' is a genitive-marked subject of a non-finite construction, which is higher on the NP accessibility hierarchy. Additionally, *vaimon* is the last-mentioned possible referent before the anaphor in a sentence with a competitor, similar to almost all antecedents of *tämä* in the corpus. If the sentence is adapted to make 'the wife' a semantically plausible antecedent, cf. sentence (9b), this seems to be confirmed. However, this unconventional use of *tämä* is not wholly inexplicable. Normally, 'the man' and 'the wife' are equal in terms of cognitive status for co-reference, but here, 'the man' is said to be in a coma. As such, 'the man'

can be seen as inanimate, while 'the wife' is animate. It is known, from research on English, that animate entities are more accessible than inanimate entities (Fukumura & van Gompel, 2011). This might have lowered the accessibility of 'the man' so, that *tämä* is a more suitable anaphor form than *hän*, which could possibly be interpreted as referring to 'the wife', which is more accessible. This effect of animacy seems to be confirmed when looking at the objects of passive sentences, where, out of 5 *tämä*-references, the antecedent was unconscious once and dead in 3 cases.

(9) a. **Miehe-n** jouduttua koomaan lääkäri-t ott-i-vat **vaimo-n** toivomuksesta
man-GEN end.up.after coma.into doctor-NOM.PL take-PL-3 wife-GEN wish.from
tä-ltä spermaa talteen.
this-ABL sperm safekeeping.into
'After the **man** fell into a coma, doctors retrieved sperm from *him*, upon the **wife's** request.'

b. **Miehe-n** jouduttua koomaan lääkäri-t läht-i-vät huoneesta **vaimo-n**
man-GEN end.up.after coma.into doctor-NOM.PL leave-PL-3 room.from wife-GEN
toivomuksesta ja *tämä* kuol-i.
wish.from and this.NOM die.PST.3SG
'After the **man** fell into a coma, the doctors left the room upon the **wife's** request and *she* died.'

Another interesting use of *tämä* is presented in sentence (10), which was not annotated in the corpus, as it does not have a non-canonically marked antecedent nor did it occur in the subset used for the annotation of regular sentences. Here, we see *hän* and *tämä* used in direct succession, referring to the same entity, *miehen* 'the man'. This is highly unusual, as one would expect *hän* and *tämä* to be used in very different contexts, not to refer to the same antecedent. There is no clear reason for using this alternation, and there are two alternative options: "[...] seuratessa *häntä* asuntoonsa [...]", where asuntoonsa translates to 'his house', without using a possessive pronoun. A different option would have been to use *hän* twice, as in: "[...] seuratessa *häntä hänen* asuntoonsa [...]". As no clear reason for the alternation can be found, this occurrence will have to be considered as an unexplained anomaly.

(10) Uhkaava tilanne synty-i päivystävän poliisipartion kiinnitettyä huomiota
threatening situation arise-PST.3SG on-call police.patrol fix.after notice
miehe-n käytökseen ja seuratessa *hän-tä* *tämä-n* asuntoon [...].
man-GEN behaviour.to and following.while s/he-PAR this-GEN house.to [...]
'A threatening situation arose after the on-duty police patrol took notice of the **man's** behaviour and while following *him* to *his* house [...].'

4.2 Independent effects of case marking and sentence type

Given the importance of a competitor as a prerequisite for *hän*/*tämä*-variation to occur, it is useful to exclude it as a factor when looking at case marking and grammatical role. As 54 out of 60 *tämä*-references occur in excerpts where there is a competitor, but not an intervening competitor, we consider only co-reference pairs found under the same conditions. This leaves 213 pairs, the details of which are presented in Table 8 (only subjects and object included).

The patterns seen in Table 8 are similar to those in 6, but are more extreme and accentuate the *hän*/*tämä*-variation, although the cell counts are even lower. We can now see a 50/50 preference when referring to regular sentence objects, with the only exception being the nominative-marked object, of which there is only one instance in the corpus. This suggests that case marking does not have an influence on accessibility and co-reference in regular sentences, or at least too small an influence to be discernible without experiments using case-variation in tightly-controlled contexts.

Regarding the non-canonical sentences, two questions have to be answered: does case-variation have an effect on accessibility and co-reference and how are the arguments of these sentences treated, i.e. more or less subject-/object-like than would be expected?

Sent. Type	Gramm. Role	Antec. Case	*hän*	*hän*-%	*tämä*	*tämä*-%	Total
Regular	Subject	Nominative	50	98.0	1	2.0	51
	Object	Genitive	3	42.9	4	57.1	7
		Nominative	1	100.0	0	0.0	1
		Partitive	7	50.0	7	50.0	14
Existential	Subject (e-NP)	Nominative	4	100.0	0	0.0	4
		Partitive	0	0.0	1	100.0	1
Necessive	Subject	Genitive	5	100.0	0	0.0	5
	Object	Nominative	1	100.0	0	0.0	1
		Partitive	0	0	2	100.0	2
Other non-finite	Subject	Adessive	1	100.0	0	0.0	1
		Genitive	9	81.8	2	18.2	11
	Object	Genitive	1	100.0	0	0.0	1
		Partitive	0	0.0	1	100.0	1
Passive	Object	Nominative	10	76.9	3	23.1	13
		Partitive	9	90.0	1	10.0	10
Locative Case	Subject	Elative	4	100.0	0	0.0	4
	Object	Ablative	0	0.0	2	100.0	2
		Allative	1	33.3	2	66.7	3
		Elative	1	100.0	0	0.0	1
		Illative	2	20.0	8	80.0	10
Have-construction	Subject	Adessive	23	88.5	3	11.5	26
Total:			132	78.1	37	21.9	169

Table 8: Number of co-referential pairs annotated from excerpts with a non-intervening competitor only, by sentence type, grammatical role, case marking role and anaphor form.

The first question is hard to answer, as it suffers from low counts for some arguments in specific sentence types. Nevertheless, some patterns can be discerned. In existential sentences, nominative-marked e-NPs prefer *hän*, while the partitive-marked e-NP prefers *tämä*. In necessive sentences, partitive-marked objects prefer *tämä* too, while nominative-marked objects, again, prefer *hän*. For the objects of non-necessive non-finite constructions, we see a similar pattern: partitive-marking increases *tämä*-preference, while genitive-marking increases *hän*-preference. For the subjects of these constructions, case seems to have no influence, *hän* is the majority-preference all around. In passive or impersonal sentences, we see a reversed pattern: nominative-marked objects have a weaker *hän*-preference than partitive-marked objects. Case, also, seems to have an effect on locative-marked objects, where allative- and illative-marked show a strong *tämä*-preference, while elative-marked objects show a tentative *hän*-preference. Considering all this, the answer to the question has to be that there are indications that case marking in non-canonical sentences has an effect on accessibility and co-reference, which it does not have in canonically-marked sentences, but that this is based on too low a number of occurrences of certain case markings to draw any conclusions. Further research, experiment- instead of corpus-based, will have to show whether or not this influence of case marking is a real thing.

The second question suffers from the same problems, but other collapsing numbers across different case markings, an image arises nevertheless. Disregarding the low counts, we get the following ranking of arguments by *hän*-preference: Necessive subjects (100.0%) = Locative-marked subjects (100.0%) > Regular subjects (98.0%) > 'Have'-construction-subjects (88.5%) > Other non-finite subjects (83.3%) > Passive objects (82.6%) > Existential e-NPs (80.0%) > Regular objects (50.0%) = Other non-finite objects (50.0%) > Necessive objects (33.3%) > Locative-marked objects (25.0%).

Seemingly, the arguments can be divided into two groups: one with a strong *hän*-preference (over 80%), and a second one with no preference for *hän* (less than or equal to 50%). Most objects fall in the second category, while most subjects fall in the first category, but a few argument types stand

out. Subjects of necessive and other non-finite constructions seemingly behave like regular subjects, and their objects behave like regular objects, indicating that, although the case marking is reversed in these sentence types, this does not influence perceptions of grammatical role or accessibility. Regarding the adessive-marked subjects of 'have'-constructions, we come to a similar conclusion. Even though they are superficially similar to the locative arguments of existential sentences, they behave like regular subjects in terms of accessibility and co-reference, regardless of their non-canonical case marking.

The well-discussed e-NP falls into the 'subject-like'-category here, which supports the analysis of this argument as a subject. However, it should be noted that a large part of the discussion about the subject- or objecthood of this argument has revolved around its partitive-marked variant, and the one partitive-marked e-NP in this study had *tämä* as its anaphor. In conclusion, no real additions to the e-NP debate can be taken from this, but it indicates that a targeted study into the e-NP from a co-reference perspective, taking case marking-variation into account, could shed new light on the discussion.

The last, and perhaps most surprising argument in the high *hän*-preference group is the object of the impersonal sentence type. This argument behaves very much like a subject with respect to co-reference, after competitor effects are factored out, even though it is seen as an object. This shows that a simple hierarchy based on grammatical role is insufficient for Finnish. Alternatively, it suggest that the object of passives is more subject-like than generally thought.

Overall, this study has shown both the benefits and drawbacks of a corpus study into a phenomenon, accessibility, that is a compound of so many different factors. On the one hand, low occurrences of the intended research material in the corpus are a risk that forbids any decisive conclusions. Additionally, any patterns that are clear and numerous, may be discounted by additional influences that cannot be controlled for, as a result of the natural variation in language production. As such, before any definitive conclusions can be drawn, more controlled, experiment-based studies need to be carried out. On the other hand, natural variation is one of the strong points of a corpus study. After studying a large amount of data, many suggestions and indications of possible interactions, influences and effects come up. Not constrained by the research set-up, a well carried-out corpus study provides many insights into possible new directions of research and suggests new views on existing theories.

Abbreviations

1 - First Person
2 - Second Person
3 - Third Person
ACC - Accusative Case
ADE - Adessive Case
ALL - Allative Case
ELA - Elative Case
e-NP - Main argument of the existential sentence (avoiding the terms 'subject' and 'object')
GEN - Genitive Case
ILL - Illative Case
INF - Infinitive
NOM - Nominative Case
PAR - Partitive Case
PARTIC - Participle
PASS - Impersonal or Passive Voice
PL - Plural
PST - Past Tense
QU - Question Marker
SG - Singular

Sources

Frangén, S., Heikura, P. & Liikka, J. (1998). *Alivaltiosihteeri: nuoret viralliset miehet*. Helsinki: Otava.

Härkönen, A.-L. (1998). *Avoimien ovien päivä*. Helsinki: Otava.

Holappa, P. (1998). *Ystävän muotokuva*. Juva: WSOY.

Kellokumpu, H. (1998). *Lasteen alaisia*. Juva: WSOY.

Lehtinen, T. (1998). *Sara@crazymail.com*. Helsinki: Otava.

Mäkelä, H. (1998). *Pelin henki: love/40 - erään ottelun tarina*. Helsinki: Otava.

Mörö, M. (1997). *Vesipajatso*. Juva: WSOY.

Paasilinna, A. (1996). *Lentävä kirvesmies*. Juva: WSOY.

Pakkanen, M. (1998). *Täysillä, Mika!* Juva: WSOY.

Ranivaara, J. (1997). *Poisliitävä Anne Lee*. Juva: WSOY.

Schuurman, N. (1998). *Vahinkorakkaus*. Helsinki: Otava.

Snellman, A. (1999). *Paratiisin kartta: romaani*. Helsinki: Otava.

Ylikangas, H. (1996). *Ilkkaisen sota*. Juva: WSOY.

References

Ariel, M. (1988). Referring and accessibility. *Journal of Linguistics*, *24*, 65–87.

Arnold, J. E. (2001). The effect of thematic roles on pronoun use and frequency of reference continuation. *Discourse Processes*, *31*(2), 137–162.

Bielecki, R. (2009). On the nature of the accusative in Finnish. *Lingua Posnaniensis*, *51*, 19–38.

Bosch, P., Katz, G. & Umbach, C. (2007). The non-subject bias of German demonstrative pronouns. In M. Schwarz-Friesel, M. Consten & M. Knees (Eds.), *Anaphors in text: cognitive, formal and applied approaches to anaphoric reference* (pp. 145–164). Amsterdam, Philadelphia: John Benjamins.

Ferretti, T. R., Rohde, H., Kehler, A. & Crutchley, M. (2009). Verb aspect, event structure and coreferential processing. *Journal of Memory and Language*, *61*, 191–205.

Fukumura, K. & van Gompel, R. P. (2011). The effect of animacy on the choice of referring expression. *Language and Cognitive Processes*, *26*(10), 1472–1504.

Gundel, J. K., Hedberg, N. & Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language*, *69*(2), 274–307.

Hakulinen, A. (1982). Subjektin kategoria vai nominaalijäsenten subjektimaisuus? In *Lauseenjäsennyksen perusteet: seminaari Seilissä, 9-10.9.1982* (pp. 17–34). Turku: Suomen Kielitieteellinen Yhdistys.

Hakulinen, A. & Karlsson, F. (1979). *Nykysuomen lauseoppia*. Helsinki: Suomalaisen Kirjallisuuden Seura.

Hakulinen, A., Karlsson, F. & Vilkuna, M. (1980). *Suomen tekstilauseiden piirteitä: kvantitatiivinen tutkimus*. Helsinki: Helsingin Yliopisto.

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

Halmari, H. (1994). On accessibility and coreference. *Nordic Journal of Linguistics*, *17*, 35–59.

Hämäläinen, J. (2005). *Non-canonical marking of grammatical subjects*. (Master's thesis, University of Helsinki, Faculty of Humanities).

Helasvuo, M.-L. (1996). Ollako vai eikö olla - eksistentiaalilauseen subjektin kohtalonkysymys. *Virittäjä*, *100*, 340–356.

Hyönä, J. & Hujanen, H. (1997). Effects of case marking and word order on sentence parsing in Finnish: an eye fixation analysis. *The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, *50*(4), 841–858.

Itkonen, T. (1974). Ergatiivisuutta suomessa. *Virittäjä*, *78*, 379–398.

Järvikivi, J., van Gompel, R. P. G., Hyönä, J. & Bertram, R. (2005). Ambiguous pronoun resolution: contrasting the first-mention and subject-preference accounts. *Psychological Science*, *16*(4), 260–264.

Kaiser, E. (2000). Pronouns and demonstratives in Finnish: indicators of referent salience. In P. Baker, A. Hardie, T. McEnery & A. Siewierska (Eds.), *Proceedings of the Discourse Anaphora and Anaphor Resolution Conference* (Vol. 12, pp. 20–27). Technical Papers. University Center for Computer Research on Language. Lancaster, UK.

Kaiser, E. (2005). When salience isn't enough: pronouns, demonstratives and the quest for an antecedent. In R. Laury (Ed.), *Minimal reference: the use of pronouns in Finnish and Estonian discourse* (pp. 135–162). Helsinki: Suomalaisen Kirjallisuuden Seura.

Kaiser, E. (2011). Saliency and contrast effects in reference resolution: the interpretation of Dutch pronouns and demonstratives. *Language and Cognitive Processes*, *26*(10), 1587–1624.

Kaiser, E. & Trueswell, J. C. (2008). Interpreting pronouns and demonstratives in Finnish: evidence for a form-specific approach to reference resolution. *Language and Cognitive Processes*, *23*(5), 709–748.

Kaiser, E. & Vihman, V. (forthcoming). *On the referential properties of Estonian pronouns and demonstratives*. To appear in the Proceedings of the 22nd Scandinavian Conference of Linguistics, Aalborg University, June 2006.

Kaiser, E., Runner, J. T., Sussman, R. S. & Tanenhaus, M. K. (2005). What influences the referential properties of reflexives and pronouns in Finnish? In C. Ebert & C. Endriss (Eds.), *Proceedings of Sinn und Bedeutung 10: annual meeting of the Gesellschaft für Semantik* (Vol. 44, pp. 155–169). ZAS Papers in Linguistics.

Karlsson, F. (1983). Prototypes as models for linguistic structure. In F. Karlsson (Ed.), *Papers from the 7th Scandinavian Conference of Linguistics, Hanasaari, Finland, December 17-19, 1982* (pp. 583–604). Helsinki: University of Helsinki.

Keenan, E. L. & Comrie, B. (1977). Noun phrase accessibility and universal grammar. *Linguistic Inquiry*, *8*(1), 63–99.

Kiparsky, P. (2001). Structural case in Finnish. *Lingua*, *111*(4), 315–376.

Maling, J. (1993). Of nominative and accusative: the hierarchical assignment of grammatical case in Finnish. In A. Holmberg & U. Nikanne (Eds.), *Case and other functional categories in Finnish syntax* (pp. 49–74). Berlin: Mouton de Gruyter.

Prince, E. F. (1981a). On the inferencing of indefinite-*this* NPs. In A. K. Joshi, B. L. Webber & I. A. Sag (Eds.), *Elements of discourse understanding* (pp. 231–250). Cambridge: Cambridge University Press.

Reime, H. (1993). Accusative marking in Finnish. In A. Holmberg & U. Nikanne (Eds.), *Case and other functional categories in Finnish syntax* (pp. 89–109). Berlin: Mouton de Gruyter.

Sands, K. & Campbell, L. (2001). Non-canonical subjects and objects in Finnish. In A. Y. Aikhenvald, R. M. W. Dixon & M. Onishi (Eds.), *Non-canonical marking of subjects and objects* (pp. 251–306). Amsterdam: John Benjamins.

Seppänen, E.-L. (1998). *Läsnäolon pronominit: tämä, tuo, se ja hän viittaamassa keskustelun osallistujaan.* Helsinki: Suomalaisen Kirjallisuuden Seura.

Siiroinen, M. (2001). *Kuka pelkää, ketä pelottaa*. Helsinki: Suomalaisen Kirjallisuuden Seura.

Tiainen, O. (1997). Suomen eksistentiaalilause - päättymätön tarina. *Virittäjä*, *101*, 563–571.

Toivainen, J. (1993). The nature of the accusative in Finnish. In A. Holmberg & U. Nikanne (Eds.), *Case and other functional categories in Finnish syntax* (pp. 111–128). Berlin: Mouton de Gruyter.

Vainikka, A. M. (1989). *Deriving syntactic representations in Finnish.* (Doctoral dissertation, University of Massachusetts, Amherst).

Varteva, A. (1998). Pronominit *hän* ja *tämä* tekstissä. *Virittäjä*, *102*(2), 202–223.

Vilkuna, M. (1989). *Free word order in Finnish.* Helsinki: Suomalaisen Kirjallisuuden Seura.

Wiik, K. (1974). *Suomen eksistentiaalilauseiden "subjekti"*. Turku: Turun yliopiston fonetiikan laitos.