Grammatical role and word order predict referential form in Finnish: an annotated corpus of Finnish

Bachelor's project

Hessel Haagsma, s1932683, h.haagsma.1@student.rug.nl

Abstract: In this project, a corpus study using crowd-sourced annotation was done to gain new insights in the distribution of two Finnish pronouns: the demonstrative *tämä*, 'this', and the personal pronoun *hän*, 's/he'. A non-tagged, non-annotated corpus was morphosyntactically tagged, the relevant pronouns were extracted and Wordrobe, a game-like, online system, which allows people with no linguistic or annotation expertise (but high language-competence) to use their knowledge for annotations, was used to acquire co-reference annotation. The expectation is that a significant relationship between grammatical role, linear word order and the usage of *tämä* and *hän* will be found, with *tämä* used for second-mentioned antecedents, and *hän* mainly used for subject antecedents. Results show low frequencies of OS-word order, so no clear distinction between the factors can be made. Sentences with only one argument and non-subject/object arguments were also studied, and an unexpectedly high number of *tämä* with intransitive subject antecedents, which cannot directly be explained. This warrants further study of these previously largely ignored sentence types.

1 Introduction

Humans are masters at understanding language, sometimes without even realizing it. One example of this is reference resolution, which we do constantly, without even being aware of it. Take, for example, sentences (1a-c), which illustrate three different cases of reference. In all sentences, the anaphor (the referring expression) is marked in *italic*, while the antecedent (that which is referred to) is marked in **bold**. In (1a), the third-person singular feminine pronoun *she* can only refer to **Maria**, and making this inference will generally not cost a speaker of English any conscious effort. The reason for this is that language users can extract simple cues, e.g. from the form of the referring expression, to create intuitions about which of the possible antecedents is the right one. In this case, the fact that *she* is feminine and **Maria** is a female name and **Allan** is not makes the choice easy.

In sentence (1b), the right antecedent is not so obvious. In this case, no explicit referent expression (also called a null pronoun, indicated by Ø) is used, so the anaphor provides no information about gender or number of the antecedent. Nevertheless, it is still obvious who left the scene, namely **Allan**. Again, we automatically extract cues from the referential form, here the null pronoun, which can only refer to a subject, which in this case is **Allan**. This contrasts with (1a), where *she* could have referred to both subject and object if they were feminine. Example (1c) illustrates this, showing that usage of *she* in cases were there are two singular female entities leads to ambiguity, as it provides no cues regarding grammatical role. Of course this does not necessary cause difficulties in daily language use, as we have intuitions and information about the real world events the sentence describes, and this helps us interpret cases of ambiguous reference.

(1) a. Allan waved at **Maria** and *she* left immediately.
b. **Allan** waved at Maria and then *Ø* left immediately.
c. **Ann** waved at **Maria**. *She* then left immediately.

In Finnish, the language studied in this project, there is an extra possible referential form in the kind of sentences presented in (1a-c). In Finnish, both the personal pronoun *hän* - 's/he'[1] and the demonstrative pronoun *tämä* - 'this' can be used to refer to singular human entities like Allan, Maria and Ann in (1a-c). Both pronouns provide speakers of Finnish with different cues regarding the antecedent. Sentences (2a,b)[2] illustrate this: in (2a), *hän* is used and generally interpreted as referring to the subject, **Anssi**, while in (2b), *tämä* is used and generally interpreted as referring to the object, **Anna**, even though there is nothing else barring the opposite interpretations (Kaiser, 2000). The use of demonstrative pronouns to refer to humans is not unique to Finnish and neither is their selection of different antecedents based on grammatical role. Similar distinctions are found in languages such as Dutch, with the referential pronouns *hij/zij* and *die/deze* (Kaiser, 2011), German, with *er/sie/es* and *der/die/das* (Bosch et al., 2007) and Estonian with the pronouns *ta* and *see* (Kaiser and Vihman, 2006)

(2) a. **Anssi** halasi **Anna-a**, *hän* oli lähdö-ssä.
Anssi.NOM hugged Anna-PART, s/he.NOM was departure-INE.
'**Anssi** hugged **Anna**, *he* was leaving.'

b. **Anssi** halasi **Anna-a**, *tämä* oli lähdö-ssä.
Anssi.NOM hugged Anna-PART, this.NOM was departure-INE.
'**Anssi** hugged **Anna**, *she* was leaving.'

1.1 The Givenness and NP Accessibility hierarchies

There are different theories concerning the relationship between the choice of referential form and the properties of the antecedent, e.g. Gundel et al. (1993); Prince (1981); Arnold (2001). A general consensus among these is that referential forms can be ranked by a certain criterion, and that the position on these ranking determines what kind of antecedent a certain referential form can refer to. One influential theory, that of Gundel et al. (1993), ranks referent forms in terms of cognitive status to create a hierarchy of referential forms called the Givenness hierarchy. The cognitive statuses in question consist of information about memory location and attentional state of the referent, which the addressee uses to distinguish a set of possible referents. For example, usage of the demonstrative determiner *that* for reference indicates to the addressee that it is familiar with the referent and can identify it. A generalized version of this hierarchy is shown in Table 1. A higher position on the hierarchy means that the anaphor has a more restrictive cognitive status, i.e. a status which allows only a smaller set of possible referents. These forms are less linguistically specified, and therefore can be referred to only by highly activated referents.

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

Table 1: A hierarchy of referential forms, with Finnish and English examples (adapted from Gundel et al. (1993))

In this hierarchy, referential forms are ranked by the size of the set of possible referents they define. Another characteristic, informativeness, can be used for the same purpose. When ranking referential forms by informativeness, the referential form that provides the most information about

[1] There is no gender marking on Finnish pronouns, so *hän* can mean both *he* and *she.*

[2] List of abbreviations used in glosses:
NOM - Nominative case, PART - Partitive case, GEN - genitive case, ALL - Allative case, INE - Inessive case, ACC - Accusative case

its antecedent is placed first in the hierarchy. This notion is approximately inversely related to the Givenness hierarchy: the more information about the referent is given, the smaller the set of possible referent becomes, so a referential form high on the informativeness hierarchy will usually be low on the Givenness hierarchy and vice versa.

This hierarchy of referential forms is usually correlated with a hierarchy of NPs, to determine which kind of NPs can be used with which referential forms. This hierarchy is based on the notion of accessibility, and is called the NP Accessibility Hierarchy. In the original Accessibility Hierarchy by Keenan and Comrie (1977), accessibility expressed the possibility of relativization for certain NP positions, making generalizations about the types of NPs that can be relativized in a language. In the study of co-reference, accessibility is used in relation to the possibility of reference to a NP in a certain position, thus giving an indication of which NP positions are most easily referred to. The standard Accessibility Hierarchy is as follows, with the most accessible form on the left:

Subject > Direct object > Indirect object > Oblique > Genitive > Complement
(after Keenan and Comrie (1977))

Many theories on co-reference propose a combination of this notion of informativeness or givenness with the NP Accessibility Hierarchy: if the anaphor is less informative or has a lower cognitive status, the antecedent has to be highly accessible to make reference possible, and vice versa if the anaphor is highly informative or has a high cognitive status.

In Finnish, however, a hierarchy based on informativeness collides with the givenness hierarchy. In Finnish, the personal pronoun *hän* is more specific, and thus more informative than *tämä*. *Hän* can only be used to refer to humans of any gender[3], while *tämä* can be used for the same, but also non-human and inanimate entities, as in (3a). It also has a more general demonstrative use, and can function as an abstract object anaphor, as in (3b), referring back to an event or situation.

(3) a. *Tämä* on tyhmä peli.
This.NOM is dumb.NOM game.NOM
'*This* is a dumb game.'

b. *Tämä* voi sattua kene-lle tahansa.
This.NOM can happen anyone-ALL
'*This* could happen to anyone.'

Nevertheless, the personal pronoun *hän* is placed higher than the demonstrative *tämä* in the hierarchy in Table 1, which thus conflicts with an informativeness hierarchy. It is known from previous research(e.g. Kaiser (2000); Halmari (1994)) that *hän* and *tämä* have different distributions, and that syntactic role of the antecedent plays an important role in this, but the exact relationship is unknown. A more detailed insight in this relationship is one of the goals of this study, and could provide information about whether the place on the givenness or informativeness hierarchy is a better predictor of the distribution of these pronouns.

In this study, we look at a corpus consisting of Finnish novels in order to find out which factors influence the *hän/tämä* distribution, mainly focusing on grammatical role and linear word order. One of the reasons for studying Finnish in particular is that Finnish has relatively free word order, i.e. all six different orders of subject (S), verb (V) and object (O) can be used if the context is appropriate (Vilkuna, 1989). This allows for a separation of grammatical role and word order, as both object and subject can occur in both the first and last position in the sentence, as opposed to the relatively fixed word order of languages such as English. Novels in particular are of interest, as these are expected to have more colloquial-style and varied discourse, which should allow for more instances of non-standard word order. In this paper, we hope to shed light on the issue of cognitive status vs. informativeness, and on two other theoretical debates, about the influence of likelihood and structural factors and the relationship between hierarchies, as discussed in sections 1.2 and 1.3.

[3] However, in colloquial spoken language, *hän* can also be used to refer to other animate entities, and sometimes even to inanimate entities (Varteva, 1998).

1.2 Influences on accessibility and referential form

Apart from the three hierarchies discussed in the previous section, other factors have been proposed to influence referential form. Arnold (2001) suggested that the likelihood of an entity being referred to in further discourse influences accessibility of that entity. We assume a relationship between referential form and NP accessibility, and therefore this likelihood should influence referential form, too. The same likelihood is assumed to influence the choice of referent, i.e. the choice of which entity out of multiple options to refer to in a follow-up sentence.

Arnold based her theory on a study involving a sentence-completion experiment and a corpus study. She focused on the influence of thematic roles, such as goal and source and found that NPs in the 'goal' thematic role, but only in transfer-of-possession events, are more likely to be referred to by pronouns instead of full NPs or names. She explains the influence of thematic role and the effect of other factors, such as grammatical role and recency by combining them into the compound notion of likelihood of reference.

If this theory of reference likelihood having an influence on accessibility is true, this would suggest that the factors researched in this study, grammatical role and linear word order, cannot fully explain the distribution of the referent forms *hän* and *tämä.* If results are found where these structural factors cannot explain the distribution of *hän* and *tämä,* this would support Arnold's theory and have to prompt research into other, semantic factors, such as thematic role in combination with certain verb groups.

More recently however, researchers have found evidence that supports a theory in which only structural factors, and not likelihood, influence choice of referent form (Fukumura and van Gompel, 2010; Kaiser et al., 2011). Fukumura and van Gompel (2010) suggested that although semantic factors, such as the semantic type of verb or connective, do indeed influence antecedent selection in sentence continuation experiments, they do not influence referential form at all and that only structural factors affect choice of referential form. To investigate this, they performed two sentence-completion experiments, one regarding influence of verb type, the other regarding the type of connective used. Both experiments found no significant effect of these factors on choice of referent form, but did find a strong effect of structural properties: pronouns were used for first-mentioned noun phrases (NP1, e.g. the subject in a sentence with SVO word order) and subjects more often, while names were used for second-mentioned noun phrases (NP2, e.g. the object in a sentence with SVO word order) and objects more often. Fukumura and van Gompel explain this lack of influence of semantic factors by suggesting that these do not affect accessibility of referents, while structural properties do. They argue that sentence structure production is dependent on the level of activation, while it is unlikely that verb or connective type are influenced by the activation level of discourse entities. Therefore the structure of the sentence containing the referent influences comprehenders' expectation and activation of discourse entities, but semantic factors do not. This is supported by other research that has found that highly activated discourse entities precede less active entities in sentence production (Ward and Birner, 2004).

Fukumura and van Gompel's findings cannot be fully corroborated in the current study, as only referent form, and not antecedent selection is studied. Nevertheless, a finding that grammatical role and linear word order work well as predictors of referent form would confirm their hypothesis that only structural factors influence referent form. Therefore, at least a partial answer could be given to the question of which of these theories is correct.

1.3 Three different approaches

Regardless of the question of **what** exactly influences anaphor and referent choice, another debate concerns the question of **how** these factors are related to referential form. There are three different approaches to this: the single-factor approach, the multi-factor approach and the form-specific approach (Kaiser and Trueswell, 2008). The single-factor approach corresponds to the idea of the strict NP accessibility- referential form hierarchy, where only one factor influences the choice of referential form. It has been used to analyze German (Strube and Hahn, 1996) and Turkish (Hoffman, 1998), but did not fully account for the distribution of referent forms. With regards to the *hän*/*tämä* distribution, proponents of the the single-factor approach only seem to occur in older sources (e.g. (Hakulinen and Karlsson, 1979; Sulkala and Karjalainen, 1992)), which refer to subjecthood or recency as the sole deciding factor for the usage of *hän* and *tämä*.

The multi-factor approach is a more recent hypothesis, and embodies the idea that such a simple hierarchy alone cannot account for all data, and that accessibility, and thus referential form, is influenced by more than one factor. Evidence for multiple factors influencing choice of referent form has been found from experiments on Finnish (Järvikivi et al., 2005), who found that grammatical role and word order together are used for resolution of ambiguous pronouns. This is close to the current hypothesis about the *hän*/*tämä* distribution, but does not account for specific variations in sensibility of *tämä* and *hän* to each of these factor, as found in (Kaiser and Trueswell, 2008). The same approach has been used to explain the use of gender and other structural information in English (Arnold et al., 2000; Gordon et al., 1993).

The form-specific approach proposed by Kaiser and Trueswell (2008) is a variant of the multi-factor approach and states that different, informationally equivalent, referential forms attach different weights for the multiple factors that make up salience, e.g. that *hän* is used for subjects, while *tämä* is used for last-mentioned entities, thus showing that the usage *hän* is conditioned mainly by the grammatical role of the antecedent, while *tämä* is influenced more by the linear position of the NP that is referred to. It is supported by findings in Dutch, with grammatical role and topicality as the deciding factors (Kaiser, 2011), English, where it was found that not only saliency, but also syntactic role influences pronoun resolution (Brown-Schmidt et al., 2005), and previous work on Finnish, discussed in section 1.4.

1.4 Previous work on *hän* and *tämä*

Due to the interesting properties of Finnish, quite some research into the *hän/tämä* distribution has already been done. Halmari (1994) was the first to investigate the relationship between referential form and grammatical role by doing a corpus study. She found a clear relationship between the two, indicating that the NP accessibility hierarchy and the givenness hierarchy are inversely related: the more accessible an antecedent, the lower its referent is on the givenness hierarchy. She failed to control for interfering factors, such as word order, and therefore did not draw any definite conclusions. Kaiser (2000) did a similar corpus study focusing on *hän* and *tämä* exclusively, looking into grammatical role and the main/subordinate clause difference. She confirmed the finding that *hän* is used more for subjects, and *tämä* for objects, and also found that *hän* is used more often for antecedents in main clauses, with *tämä* being used more for antecedents in subordinate clauses.

Other work on Finnish has mainly focused on antecedent choice, as opposed to choice of referential form. Järvikivi et al. (2005) looked into the influence of word order and grammatical role on the resolution speed of referential *hän*, using an eye-tracking experiment. They found that both factors had a highly significant influence on pronoun interpretation: *hän* was interpreted as referring to the subject and to the first-mentioned argument. However, in OVS sentences, this preference for first-mentioned arguments was not present, indicating that grammatical role is the most important factor for identifying the antecedent of *hän.* This confirms the theory of Fukumura and van Gompel (2010), who suggested that only these kind of structural factors determine the distribution of referent forms. This was also found by Kaiser (2005), based on two sentence-completion experiments, and by Kaiser and Trueswell (2008) who did a sentence-completion and an eye-tracking experiment and found that *hän* is indeed mostly interpreted as referring to subjects, and that *tämä* prefers post-verbal referents, discourse-new referents and objects. Seppänen (1998), who looked into spoken discourse, found that *tämä* is generally used for conversation participants who have been speaking earlier. Regarding *hän*, she found no such clear tendency, but concluded that it was used in cases where the speaker expressed the referent's viewpoints or opinions.

2 Methods

2.1 The corpus data

The current corpus study is based on 20 different texts[4] by 18 different (groups of) authors from the Otava 1998/1999 and WSOY 1996/ 1997/1998 parts of the Finnish Text Collection[5]. All texts were novels, comprising many different genres. The texts were not specifically selected, the only selection criterion used was that they were novels. For every text all instances of both *hän* and *tämä* and the two sentences preceding the reference were extracted. In the case of *hän*, the word was extracted in all possible cases, while for *tämä*, only the instances in nominative cases were extracted, as the percentage of referential use in other cases was marginal, while this was not an issue with *hän.* Then, morphological tagging was added using the Constraint Grammar-based FinCG-tagger[6], as obtained from the Giellatekno research group of the university of Tromsø. This morphological information was then used to perform basic NP-chunking and to extract possible NP-antecedents from the text fragments, using a straightforward Python script. The same script was used to filter out the cases were *tämä* was used as a determiner, and not as an independent referring expression, and to limit the research data to the first 15 usable 'hits' in each text. As not every novel contained 15 instances of each expression, this resulted in 239 excerpts containing *tämä*, and 266 excerpts containing *hän.* This data was then converted for use with the Wordrobe platform, which is further discussed in section 2.2. Syntactical information regarding word order and grammatical role was added manually.

2.2 Wordrobe

To collect the actual co-reference data, i.e. data about what each anaphor refers to, the Wordrobe[7] platform was used. Wordrobe is a collection of linguistic 'games with a purpose', games which can be played by non-linguists, whose answers can then be used to gather linguistic data. Games on word senses, compounds, named entities and pronouns and reference exist. In this study, we created a Finnish version of the game on pronouns and reference, called 'Viittaukset'. There are several advantages of this approach over gathering the co-reference data manually ourselves, as previous studies have done. First of all, using Wordrobe allowed for the use of native Finnish speakers' knowledge for annotation, without having to be in Finland. Secondly, Wordrobe allows us to use the intuitions of non-linguist, but expert language users, which eliminates the risk of possible biases caused by knowledge of the phenomena in question here. A third advantage is that Wordrobe allows for multiple answers to each 'question' (one referring expression and its possible antecedents, presented in context), from which then a majority vote can be deduced, yielding more accurate answers than relying on just one result. Participants for the game were recruited by word-of-mouth, through friends and other contacts, and via public online platforms, such as Reddit.

[4] The used texts were: Anna-Leena Härkönen - *Avoimien ovien päivä* (1998), Simo Frangén, Pasi Heikura (1998), Jyrki Liikka, - *Alivaltiosihteeri: nuoret viralliset miehet* (1998), Hannu Mäkelä - *Pelin henki: Love/40 - Erään ottelun tarina* (1998), Tuija Lehtinen - *Sara@crazymail.com* (1998), Anna-Liisa Suni - *Tassuterapeutti: koirista ja vähän muistakin eläimistä* (1998), Nora Schuurman - *Vahinkorakkaus* (1998), Anja Snellman - *Paratiisin kartta: romaani* (1999), Olli Jalonen - *Yksityiset tähtitaivaat* (1999), Arto Paasilinna - *Lentävä kirvesmies* (1996), Heikki Ylikangas - *Ilkkaisen sota* (1996), Sirpa Puskala - *Villimies ja pakkopaita* (1997), Jorma Ranivaara - *Poisliitävä Anne Lee* (1997), Mari Mörö - *Vesipajatso* (1997), Mitri Pakkanen - *Täysillä, Mika!* (1998), Arto Salminen - *VARASTO* (1998), Arto Paasilinna - *Hirttämättömien lurjusten yrttitarha: rosvoromaani* (1998), Raili Manninen - *Sinisen ponitallin ratsastajat* (1998), Mari Mörö - *Kiltin yön lahjat* (1998), Helmi Kellokumpu - *Lasteen alaisia* (1998), Pentti Holappa - *Ystävän muotokuva* (1998)

[5] More information on the Finnish text collection can be found at the website of the Finnish IT Center for Science, www.csc.fi

[6] More information at http://giellatekno.uit.no

[7] www.wordrobe.org

3 Results

3.1 General distributional information

The corpus, comprising 20 novels, contains approximately 1.1 million words. In this text collection, *tämä,* in the nominative form, occurs 960 times, of which less than 497 instances are referential. *Hän*, in all its conjugated forms, occurs 13258 times. The frequency of usage for these anaphors differs strongly across texts, with *tämä* occurring once every 6500 words in one text, and once every 800 words in another. The same holds for *hän*, occurring once every 40 words in one text, and not at all in another, where the colloquial alternative *se*, 'it', is consistently used instead. To get similar amounts of data on both *hän* and *tämä*, only the first 15 instances of each pronoun were selected for each text.

3.2 Distributional results from annotated data

As mentioned in 2.1, 239 instances of *tämä* and 266 instances of *hän* were investigated, making a total of 505 questions. Over a period of two weeks, 1144 answers were given. An overview of the number of answers received per question is shown in Table 2 below.

# of answers	# of questions	% of questions
0	145	28.7
1	95	18.8
2	76	15.0
> 2	189	37.5

Table 2: The number of questions with a certain number of answers received

The relatively high amount of questions with 0 answers does not indicate that these questions were never presented to participants, however. Instead it indicates that these questions were consistently skipped because they either had no proper antecedent among the answer options, or because the word in question was not an actual referring expression.

This is due to the antecedent not occurring in the two sentences before the anaphor, which can be solved by taking a bigger fragment, or to mistakes in processing the data for use in Wordrobe. In the second case, problems were caused by faults in the tagging, e.g. tagging *häntä* as the partitive form of *hän*, and not as the nominative form of *häntä,* 'tail', and by mistakes in NP-chunking and filtering of the *tämä* instances to include only referential uses, resulting from that.

Because of this, only the data from questions with at least two answers will be considered. Due to the low number of answers per question, a simple majority measure (proportion of same answer over 50%) is used to decide on the correct answer, and thus the correct antecedent of a given instance of *tämä* or *hän.* Using this criterion, 236 questions received a definite answer, 97 instances of *tämä* and 139 instances of *hän*.

Investigation of the sentences that received inconclusive answers revealed several reasons for this lack of cohesion in the answers. In some cases, answers were split between a proper name, identifying a referent and another referring expression, referencing the same entity, for example '**sen**' and '**Stiina**' in sentence (4) (the anaphor is marked in *italic*, the antecedents in **bold**).

(4) **Stiina** vielä jää. **Se-n** vaattee-t roikku-vat kortte-i-tten päällä kuin pyykki
Stiina.NOM still stays. It-GEN clothes dangle horsetails top.on like laundry
kuivama-ssa. *Hän* kiirehtii omat vaatteet sylissä...
drying. S/he.NOM hurries own clothes lap.in
'**Stiina** still remains. **Her** clothes dangle on top of the horsetails like drying laundry. *She* hurries with her own clothes in her lap...'

Another reason is that, in cases where the referent was not among the answer options, a different, faulty answer was selected instead of skipping, due to lack of clear instructions. An example of this can be seen in sentence (5), which does not contain the antecedent of *hän*.

(5) Olisi jäänyt kotiin, suosiolla. Miettinyt vähän. *Hän* karkotti ajatuksen
Would stayed home, rather. Thought little.bit. S/he.NOM chased thought
saman tien.
immediately
'You should have stayed home, rather. Thought for a bit. *S/he* banished the thought immediately.'

A third cause of inconclusive answers are sentences where the referent is mentioned twice, and both mentions were selected as the correct answer, disregarding the instructions to select the antecedent closest to the pronoun in those cases. An example of this is '**Mummo**' and '**mummolle**' in sentence (6).

(6) **Mummo** ei tunne sinu-a, Oilimaija sanoi. Meiju on aina
Grandma.NOM not know you-PART, Oilimaija.NOM said. Meiju.NOM is always
toimitellut **mummo-lle** asioita ja käy nytkin katsomassa *hän-tä*.
carry.out grandma-ALL things and goes now.even visit her-PART
'Grandma does not know you, Oilimaija said. Meiju has always ran errands for Grandma and nowadays still goes to visit *her*.'

Another cause is faulty NP-chunking. In these cases, the correct antecedent NP is not presented as a single answer, but parts of it are presented as two different answers. This occurred often with proper names, in which the first and last name were presented as separate answer options, as with '**Antti**' and '**Laurinpoika**' in (7). Also, in some cases, a plain wrong answer was given, which distorted the results, without any clear explanation for the mistake.

(7) **Antti Laurinpoika** ei ole muuttunut sitten vuoden 1592, jolloin Ilkka on
Antti Laurinpoika.NOM not is changed since year 1592, when Ilkka.NOM is
häne-t viimeksi nähnyt.
he-ACC last seen.
'**Antti Laurinpoika** hasn't changed since 1592, when Ilkka saw *him* for the last time.'

3.3 Wordrobe results

The possible answers were tagged with information regarding grammatical role: subject, object or other (e.g. genitive possessor or indirect object) and information regarding word order: NP1 (e.g. the subject in an SVO-sentence), NP2 (e.g. the object in an SVO-sentence) or NPx (e.g. the object in an imperative sentence without a subject, where no SO/OS-order can be determined). The distribution of these attributes can be seen in Table 3 and 4, below.

	Subject	**Object**	**Other**	**Total**
hän	101 (72.7%)	7 (5.0%)	31 (22.3%)	139 (100%)
tämä	15 (15.4%)	28 (28.9%)	54 (55.7%)	97 (100%)
total	116 (49.2%)	35 (14.8%)	85 (36.0%)	236 (100%)

Table 3: Distribution of the grammatical roles of the antecedents of *hän* and *tämä*

A Fisher's exact test, performed on the data in Table 3, showed that the percentages of the grammatical roles of the antecedents differed significantly by pronoun ($p < 0.001$). The same was found for the data in Table 4, for which the same test was performed. This revealed that the percentages of

	NP1	NP2	NPx	Total
hän	41 (29.5%)	4 (2.9%)	94 (67.6%)	139 (100%)
tämä	3 (3.1%)	24 (24.7%)	70 (72.2%)	97 (100%)
total	44 (18.6%)	28 (11.9%)	164 (69.5%)	236 (100%)

Table 4: Distribution of the linear position of the antecedents of *hän* and *tämä*

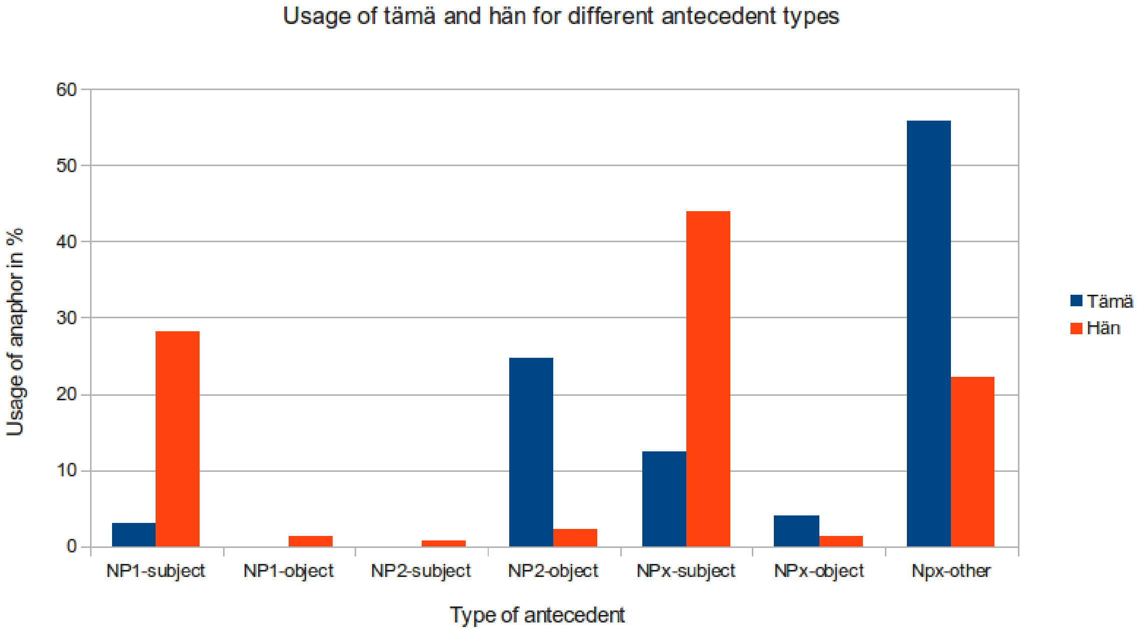

Figure 1: Usage of *tämä* and *hän* in percentages across antecedent types (adapted from Table 5)

the linear position of the antecedents differed significantly by pronoun too ($p < 0.001$). The purpose of this study is to look into the relationship between anaphor choice, grammatical role and word order, for that, the data on non-subject and non-object antecedents is not directly relevant. However, to give a complete image of the results, these figures are included too to yield Table 5. Figure 1 displays the same data as Table 5, but provides a clearer overview of *hän*/*tämä* preference across antecedent types.

	NP1-subj.	NP1-obj.	NP2-subj.	NP2-obj.	NPx-subj.	NPx-obj.	NPx-oth.	Total
hän	39 (28.1%)	2 (1.4%)	1 (0.7%)	3 (2.2%)	61 (43.9%)	2 (1.4%)	31 (22.3%)	139 (100%)
tämä	3 (3.1%)	0 (0%)	0 (0%)	24 (24.7%)	12 (12.4%)	4 (4.1%)	54 (55.7%)	97 (100%)
total	42 (17.8%)	2 (0.9%)	1 (0.4%)	27 (11.5%)	73 (30.9%)	6 (2.5%)	85 (36.0%)	236 (100%)

Table 5: Distribution of grammatical role and argument positions of antecedents of *hän* and *tämä*.

4 General discussion

Several interesting observations can be made on the data in Tables 3 and 4. In general, subjects were more often referred to than objects or arguments in other grammatical roles. This falls in line with the general tendency to continue with subject referents than with object or oblique referents, as found in other corpus studies (Arnold, 2001). Another explanation for this is that NPx-sentences were also counted in these figures, which includes intransitive sentences with only a subject, thus distorting the figures. See sentence (9a), for an example of a NPx-subject sentence.

Table 3 shows a clear preference of *tämä* for antecedents with grammatical functions lower on the saliency hierarchy: it is not frequently used for subjects, while it is predominantly used for direct objects and arguments with other grammatical functions, such as genitives and obliques which are even lower on the hierarchy (cf. section 1). *Hän* shows a reversed pattern, with a clear preference for subjects. However, it is used proportionately more for 'other' arguments than for objects, which contradicts the ranking of the saliency hierarchy.

A similar, counterintuitive pattern was found by Halmari (1994). Halmari, who split the 'other' category into four more specific categories, found an unexpected high usage of pronouns to refer to genitive, i.e. possessor arguments. She accounted for this by stating that, when the antecedent is a human possessor, the possessed is usually non-animate, so not a candidate for reference. This leaves the possessor as a highly salient antecedent, especially if no other human candidates occur in the same sentence. An example of this can be seen in sentence (8a), where the genitive '**Tanjan**' is the only animate entity in the sentence, although it is not used as a possessive in this case. Similar explanations can be given for other antecedent types, such as indirect objects, which might be the only human entity in the sentence, e.g. when the object is inanimate, and there is no (overt) subject due to passivization or pronoun-dropping, such as in sentence (8b).

(8) a. Haluan nauttia **Tanja-n** kanssa niin täysillä kuin mahdollista. Haluan tutustuttaa
Want enjoy Tanja-GEN with so fully as possible. Want introduce
häne-t ...
s/he-ACC ...

'I want to enjoy my time with **Tanja** as fully as possible. I want to introduce *her* to ...'

b. **Jalmari Jyllänkedo-lle** tarjottiin mahdollisuutta majoittua *hän-tä* varten varattuun
Jalmari Jyllänketo-ALL offered possibility occupy he-PART for reserved
huoneeseen ...
room ...

'**Jalmari Jyllänketo** was offered the option of staying in the room that was reserved for *him* ...'

Table 4 shows a similarly clear image as Table 3: *tämä* is used predominantly for NP2-arguments, while *hän* is used more for NP1-arguments. For NPx-arguments, i.e. arguments in sentences which either lack an overt subject or object, both the pronoun and the demonstrative are viable options. Again, this matches with expectations from previous research that *hän* prefers first-mentioned antecedents, while *tämä* favors second-mentioned antecedents. The equal usage regarding NPx-arguments can be explained by the nature of these arguments. A lot of these antecedents are subjects in intransitive sentences, which are usually referred to by *hän*, as can be seen in Table 5 under NPx-subject. An example of this is **Vitikainen** in example (9a). As there is only a low number of NPx-objects, the rest of these arguments consists of 'other'-type arguments such as genitives and obliques. An example of such a genitive antecedent is **Risto Salon** in example sentence (9b). These arguments are more frequently referred to by *tämä*, accounting for the unexpected balance in the data for NPx-arguments.

(9) a. **Vitikainen** käveli Annea kohti, ja Anne huomasi *häne-t*.
Vitikainen.NOM walked Anne towards, and Anne noticed s/he-ACC

'**Vitikainen** walked towards Anne, and Anne noticed *him*.'

b. Näen **Risto Salo-n** kasvoista, että *tämä* on erittäin tyytyväinen
See Risto Salo-GEN face, that this.NOM is very pleased
pelaajatilanteeseen.
player.situation.

'I see from **Risto Salo's** face, that *he* is very pleased with the current player situation.'

However, the most interesting findings can be found in Table 5 and Figure 1, which relate directly to the research question. First of all, a clear image arises which corroborates findings from previous studies on Finnish, namely that *hän* is predominantly used for subjects and NP1-arguments, while *tämä* is predominantly used to refer to objects and NP2-arguments. This shows up clearly in the data for NP1-subjects and NP2-objects, where *hän* is frequently used for the first, and *tämä* is most frequently used to refer to the latter. Similar subject- and object-preferences show in the NPx-data, where *hän* is used more for NPx-subjects and *tämä* is the anaphor of choice for NPx-objects.

Although the general tendency follows expectations, two observations stand out: first, and most importantly, the low number of instances of arguments from OS-sentences, with just 2 NP1-objects and 1 NP2-subject among the antecedents. Second, the relatively high usage of *tämä* to refer to NPx-subjects, which contradicts with the general tendency for subjects to have a strong preference for *hän*.

The unexpected usage of *tämä* to refer to subjects of intransitive sentences cannot be directly explained. The NP accessibility hierarchy makes no distinction between subjects of transitive and intransitive sentences, although the explanation could well lie here. It is possible that intransitive subjects are less accessible to reference than subjects of transitive sentences, falling somewhere in between (transitive) subjects and objects on the NP Accessibility. Treating intransitive and transitive subjects differently is known to happen in other languages, such as those with an ergative alignment. Further research into the influence of verb transitivity as an additional structural influence on referent form could provide more insights into this, and might indirectly shed light on the influence of grammatical role, which is directly related to this.

Both the NP1-objects and NP2-subjects occurred as antecedents of *hän*, but no significant meaning can be attached to this, due to the low number of instances of both. This is problematic, as data on these antecedent types is necessary to separate the influences of grammatical role and word order. Therefore, no meaningful answer to either the question of which factors influence anaphor choice or the question of which of the three approaches to referent form - factor interaction is the right one can be given.

However, the low occurrence of both these antecedent types is an interesting fact in itself. Fact is that, although the number of 236 antecedents is not very great, only 3 instances of non-SO word order is very low. In general, among the possible NP-antecedents selected for this study, NP1-objects and NP2-subjects only occur 14 and 22 times, respectively, out of a total of 3715 antecedent candidates. This indicates that the low numbers are not due to arguments in an OS-sentence not being referred to, but to this order not being found in the sentences before a reference, in general. This low frequency is supported by Kaiser (2000), who looked at 100 instances of both *hän* and *tämä*, but found no reference to first-mentioned objects or last-mentioned subjects. This low occurrence of OS-type sentences probably explains why other studies on word order and grammatical role have used other types of experiments, e.g. eye-tracking or sentence-completion, as opposed to a corpus study. This raises questions about the frequency of non-SO word order in Finnish, and the suitability of Finnish as a language for studying grammatical role and word order separately. Also, it raises the question how the influence of word order on e.g. the distribution of *tämä*, which has been established in experimental results, comes about.

Contrary to these findings, other research such as the corpus study (not focused on co-reference) by Hakulinen et al. (1980), indicates a much higher frequency of OS-word order in written Finnish, reporting that around 16% of all sentences have the subject as last-mentioned argument, but making no distinction between adverbials, predicates and objects, thus providing no exact count for OS-word order, although it is likely to be significantly higher than the ~1% found in this study. Similarly, Hakulinen and Karlsson (1979) state that the OVS-order is by no means unusual in written transitive sentences, so it is expected to occur in both newspapers and novels. On top of that, this non-canonical word order is even more common in spoken, colloquial language, showing that these word orders are an intrinsic part of the Finnish language (Vilkuna, 1989).

Then why is the frequency of objects preceding subjects so low in the co-reference related corpus studies? One possible explanation is text type: for this corpus, only novels were used, while in the

Hakulinen et al. (1980) study, four different genres, non-fiction prose, encyclopedias, editorials, and newspaper articles on culture, humor and news, but not novels were studied. However, the similarity in order frequencies between those genres and the relative textual freedom of literary prose makes it unlikely to have accounted for much of the difference.

To achieve a better explanation for this discrepancy, more specific research into word order frequency in relation to co-reference is necessary. Nevertheless, given the evidence for the intrinsic nature of 'free' word order, corpus studies on Finnish can still provide valuable insights in the questions regarding anaphor choice and accessibility hierarchies. A study similar to this one, using a larger corpus and higher-accuracy NP-chunking and tagging should yield more OS-order sentences and thus give a better detailed image of the *hän/tämä* distribution.

Acknowledgements

I would like to thank everyone who played the Viittaukset game for their input and interesting feedback. Also, thanks to Trond Trosterud for his patient help with the FinCG-parser, to Pirita Pyykkönen-Klauck for her early comments on the game and to Johan Bos and Valerio Basile for access to and help with Wordrobe.

References

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

Arnold, J. E., Eisenband, J. G., Brown-Schmidt, S., and Trueswell, J. C. (2000). The rapid use of gender information: evidence of the time course of pronoun resolution from eyetracking. *Cognition*, 76:B13–B26.

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

Brown-Schmidt, S., Byron, D. K., and Tanenhaus, M. K. (2005). Beyond salience: Interpretation of personal and demonstrative pronouns. *Journal of Memory and Language*, 53:292–313.

Fukumura, K. and van Gompel, R. P. G. (2010). Choosing anaphoric expressions: Do people take into account likelihood of reference? *Journal of Memory and Language*, 62:52–66.

Gordon, P. C., Grosz, B. J., and Gilliom, L. A. (1993). Pronouns, names, and the centering of attention in discourse. *Cognitive Science*, 17:311–347.

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

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

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

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

Hoffman, B. (1998). Word order, information structure and centering in Turkish. In Walker, M. A., Joshi, A. K., and Prince, E. F., editors, *Centering Theory in Discourse*, pages 251–272. Clarendon Press, Oxford.

Järvikivi, J., van Gompel, R. P. G., Hyönä, J., and 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 Baker, P., Hardie, A., McEnery, T., and Siewierska, A., editors, *Proceedings of the Discourse Anaphora and Anaphor Resolution Conference*, volume 12 of *Technical Papers*, pages 20–27, Lancaster, UK. University Center for Computer Research on Language.

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

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., Holsinger, E., and Li, D. C.-H. (2011). It's not the words, but how you say them: Effects of referential predictability on the production of names and pronouns. In Hendrickx, I., Branco, A., Devi, S. L., and Mitkov, R., editors, *Anaphora Processing and Applications, Lecture Notes in Artificial Intelligence*, volume 7099, pages 171–183. Springer, Heidelberg.

Kaiser, E. and 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. and Vihman, V. (2006). On the referential properties of Estonian pronouns and demonstratives. In *Proceedings of the 22nd Scandinavian Conference of Linguistics*.

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

Prince, E. F. (1981). Toward a taxonomy of given-new information. In Cole, P., editor, *Radical Pragmatics*, pages 223–256. Academic Press, New York.

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

Strube, M. and Hahn, U. (1996). Functional centering. In *Proceedings of ACL '96*, pages 270–277.

Sulkala, H. and Karjalainen, M. (1992). *Finnish.* Routledge, London.

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

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

Ward, G. and Birner, B. J. (2004). *Information Structure*, chapter 7, pages 153–174. Basil Blackwell, Oxford.