Choosing between zero and pronominal subject: Modeling subject expression in the 1st person singular in Finnish conversation

Abstract

The variability of subject expression has been extensively investigated across languages. Previous studies, however, have primarily considered only a handful of variables at a time. We present a large-scale multivariate statistical analysis of the choice of subject expression in the 1st person singular in spontaneous Finnish conversation, with a focus on the choice between pronominal and zero subject. Spoken Finnish represents an interesting case, as the dominant type of subject expression is double-marking, i.e. the combination of a pronominal subject marker (subject pronoun) and a verbal subject marker (person marking). Siewierska (1999) notes that this type of marking is typologically rare (see also Dahl 2000). Our findings indicate that the choice of subject expression is affected by both constructional and cognitive/discourse factors, and that an important role in the choice of subject expression is played by the sequential structure of the conversation.

Keywords: multivariate analysis, zero subject, pronominal subject, subject marker

1 Introduction

Across languages there is considerable variation in the ways in which pronominal subjects are expressed. First, there is variation regarding separate pronominal markers: there are languages, for example Swedish, in which pronominal subjects are normally if not obligatorily expressed in subject position, while in others, such as Japanese, pronominal subjects are usually not expressed. Secondly, the subject may be indexed with a pronominal affix on the verb. In conversational Hebrew, for example, the subject is indexed on the verb with an affix on the verb in the past and future tenses, Hacohen and Schegloff (2006). Dryer (2013) offers a typological overview of the expression of pronominal subjects.

Our focus in this article is on Finnish, where the subject is marked with an affix on the verb. From a typological perspective, Finnish is a mixed-type language in terms of subject marking, where the 1st and 2nd person behave differently compared to 3rd person marking (cf. Dryer 2011). In the third-person, subject is normally expressed with a separate subject pronoun in addition to the verbal subject marker. First- and second-person subjects, on the other hand, show interesting variation. In the written standard language the pronominal subject is usually omitted in the 1st and 2nd person, while in casual conversation it is much more common to have "double-marking" of the subject, i.e. a pronominal subject marker together with the verbal subject marker. Siewierska (1999) notes that this type of "grammatical agreement" is typologically rare (see also Dahl 2000). Even in conversational language, however, it is possible to omit the pronominal subject, and in some contexts this is the preferred alternative (see Helasvuo 2014a for a discussion of these contexts). Our purpose here is to explore the concurrent influence of different contextual factors, so as to determine when the pronominal subject is likely to be expressed and when it is omitted.

We approach these questions from the perspective of a large-scale statistical analysis that seeks to model the choice of subject expression. More specifically, we are interested in the variation between the presence and absence of a pronominal subject. The latter will be referred to as the “zero subject”.[1] Using a corpus of conversational Finnish, we focus on subject expression in the 1st person singular.

Sacks and Schegloff (1979) propose that there is a general preference for recipient design in conversation, according to which speakers will use reference forms allowing the recipient to recognize who is being referred to. In addition, reference to person is subject to the principle of minimization, according to which reference to person is “preferredly done with a single reference form”. (Sacks and Schegloff 1979: 16). More recently, Levinson (2007) has pointed out that there are several and sometimes conflicting principles at work. He suggests that optimization of expressions of reference to person is primarily governed by three principles: economy (e.g. Sacks and Schegloff’s minimization), recognition (cf. recipient design), and circumspection. According to the principle of circumspection, speakers should avoid over-reducing the set of referents. We might say that the principles of both economy and circumspection aim at a form of reference that is sufficient for the needs of the participants. The question, however, remains: what exactly constitutes “sufficient” reference?

Recent empirical studies suggest that the realization of a subject argument cannot be attributed to a single variable but instead is primarily influenced by a constellation of factors (cf. Kibrik 2011; Travis and Torres Cacoullos 2012). These can be divided into two groups: constructional and discourse/cognitive factors. The former are related to the morphosyntactic and semantic properties of the clause, the latter to the general properties of discourse and cognitive abilities. However, it is an open question whether previously established tendencies also extend to a dominantly double-marking language variety. In this respect, our purpose here is to contribute to the linguistic discussion on cross-linguistic tendencies regarding patterns of subject expression.

Our article is structured as follows: in Section 2, we give an overview of the patterns associated with subject expression in Finnish from a cross-linguistic perspective. This is followed in Section 3 by a description of the data and the operationalization of the factors used in this study; we first discuss the constructional factors that have been shown to influence subject expression in spoken language, followed by a consideration of the role of discourse and cognitive factors. In Section 4, the data are modeled using mixed-effects logistic regression. The benefits of a multivariate statistical model for this type of data are twofold; it enables us to determine whether or not the patterns observed in the data are the product of chance, and it allows simultaneous evaluation of multiple competing hypotheses. Finally, in Section 5 the results and implications of the model are laid out.

2 Subject expression in Finnish from a cross-linguistic point of view

In Finnish, the predicate verb agrees with the subject in number (singular versus plural) and person (1st, 2nd, and 3rd) (see e.g. Sulkala and Karjalainen 1992). Only nominative subjects trigger agreement in the predicate; in other words, nominative subjects enable the presence of both the verbal and the pronominal subject marker. In this article, we focus on nominative subjects. In the 3rd person singular the pronominal marker is usually expressed, while the 1st and 2nd person show variation. In written Finnish, 1st and 2nd person subjects are generally

[1] This is merely a convenient form of shorthand. It is important to note that clauses with no pronominal subject still contain a verbal subject marker the indexing 1st person singular subject (see Section 2 for a more detailed discussion).

not expressed and the verbal marker encodes the subject; in conversational Finnish, it is common for both the pronominal and the verbal subject marker to occur.

Standard Finnish includes a norm concerning the use of pronominal subjects in the 1st and 2nd person. This norm evolved gradually from the 17th century onwards, and was debated in prescriptive writings especially in the 19th century. The arguments presented to support the avoidance of pronominal subjects included economy and avoidance of redundant markings (Strellman 2005). Interestingly enough, this norm has been widely adopted in the written varieties of Finnish; not only in those varieties which are expected to follow the standard, such as journalistic writing, but also in more recent colloquial varieties which otherwise do not adhere to the standard, such as text messages (Helasvuo 2014c) or chat discussions (Meriläinen 2011; Helasvuo 2014d). With respect to the expression of pronominal subjects in the 1st and 2nd person, written Finnish thus tends to follow the principle of minimization/economy (Sacks and Schegloff 1979; Levinson 2007), and is similar to that of conversational Hebrew (Ariel 1990: 48–49; Hacohen and Schegloff 2006). If the subject is overtly expressed in Standard Finnish, it usually serves some specific discourse function, such as contrast (see Helasvuo 2014b for a discussion of this norm). As Travis and Torres Cacoullos (2012: 714–715) point out, however, the concept of contrast has rarely been investigated quantitatively. Their statistical analysis shows that contrast, operationalized using three different measures, is not a statistically significant predictor in their data from conversational Columbian Spanish.

While written Finnish tends to avoid pronominal subjects in the 1st and 2nd person, the same does not hold for conversational Finnish. Rather, there appears to be a preference for "double-marking" in the 1st and 2nd person; in other words, subjects tend to be encoded by both a pronominal and a verbal marker. This means that the preference for minimization in referring to persons, as described by Sacks and Schegloff (1979), and Levinson (2007), does not hold in the 1st or 2nd person singular. "Single-marking" (i.e. verbal subject marker only) is possible in certain specific conversational contexts, such as in cases of same-subject coordination (see example 1) or list construction (example 2). (See Helasvuo 2014a for an in-depth discussion of these contexts).

Example 1 (SG151)
1 Anni: sit mie ha-i-n se-n
then I fetch-PST-1SG it-ACC
'Then I fetched it'

2 ja **laito-i-n** kauheesti °sii-he hoitoaine-tta°
and put-PST-1SG awfully it-ILL conditioner-PAR
'and applied an awful lot of conditioner to it.'

Example 2 (SG151)
1 Anni: ja mie ot-i-n semmose-n valtava-n valkose-n lanka-vyyhi-n
and I take-PST-1SG such-ACC big-ACC white-ACC wool-coil-ACC
'and I took a big white hank of wool'

2 Anni: ja **laito-i-n** pää-hä-ni
and put-PST-1SG head-ILL-PX1SG
'and put it on top of my head'

3 Anni: ja **irvist-i-n** [Jusu-lle] sillee ((GRINS))

and grin-PST-1SG NAME-ALL like.that
‘and grinned at Jusu like this’

4 Sanna: [myhyh]
‘Uh huh.’

Example 1 illustrates a case with a pronominal subject in the first part of the coordinated compound (line 1); in the latter part of the compound, however, the pronominal subject is not repeated, but is replaced by a zero subject (line 2). This is a classic case of same-subject coordination. Example 2 illustrates a list construction: the first item on the list (line 1) has a pronominal subject, while the second and third item (lines 2 and 3), both have zero subject.

Single-marking is also possible in adjacency pairs: in the latter part of the adjacency pair, a zero subject is commonly used (see Helasvuo 2014a). In Finnish question-answer adjacency pairs, for example, the speaker can respond to a polar question either with a particle (comparable to the English ‘yes’/‘no’) or with just a repeat of the finite verb (see Sorjonen 2001). Consider example 3:

Example 3 (SG151)
1 Sanna: nii oo-t sie jo tä-nä aamu-na ol-lu jo sali-lla,
PTC be-2SG you.SG already this-ESS morning-ESS be-PCP already gym-ADE
‘So have you already been to the gym this morning already [sic]’
2 Anni: **oo-n**.
be-1SG
‘(Yes) I have.’

In (3), Sanna poses a polar question to her co-participant in line 1. The question contains a 2nd person singular form of the auxiliary verb (*oot*) and a 2nd person singular pronominal subject (*sie*). In the answer part of this question-answer adjacency pair (line 2) only the finite verb is repeated, now in the 1st person form (*oon*). Line 2 forms an affirmative answer to the question in line 1.

In sum, Finnish displays a system where both double- and single-marking are grammatically possible. Our focus here is on the variation between zero versus pronominal subjects in the singular first person in Finnish conversational data. We consider only contexts which allow for nominative subjects.

3 Data and variables

The data for this study, comprising approximately seven hours of recordings, come from spontaneous Finnish face-to-face conversations (*n* = 12) among friends and family members (*n* = 58). The data were extracted from the Spoken Language Archives at the University of Turku and the Conversation Analysis Archives at the University of Helsinki. The data were initially transcribed and segmented into syntactic units: clauses, free/unattached NPs, or particles forming utterances of their own (see Helasvuo 2001: 21–33, 105–13).[2] The data contain a total of 15,337 syntactic units and 64,906 words. From these data, all clauses with 1st person singular verb forms were extracted, amounting to a total of 1,788 syntactic units. Since only nominative subjects trigger agreement in the verb, only clauses which allow for

[2] We gratefully acknowledge the help of Milja Väänänen, Pilvi Mattila and Jenny Niemelä who worked as research assistants for the project.

nominative subjects have the possibility of being either double-marked (pronominal and verbal subject marking) or single-marked (verbal subject marking alone). We therefore restricted our focus to clauses which have or could have nominative subjects. This allowed us to investigate possible systematic tendencies in the choice between zero subject versus pronominal subject in spontaneous Finnish conversation.

The data set was further analyzed for type of subject expression (pronominal vs. zero subject) and was encoded for several factors, which constitute the independent variables in our statistical model (see section 4). First, however, we discuss in detail constructional factors in 3.1, followed by discourse and cognitive factors in 3.2. Finally, information on the variables available in the data set is summarized in Section 3.3.

3.1 Constructional factors

Subject expression is intertwined with the semantic properties of verbs. With the growth of usage-based studies on conversational discourse, evidence has accumulated across languages for a positive association between 1st person subjects and verbs of cognition, for example in American English conversation (Kärkkäinen 2003, 2007; Scheibman 2002: 63; see also Tao 2001), British English (Kaltenböck 2007), Estonian (Keevallik 2003), Mandarin (Tao 1996: 25, 26, 124; Endo 2010, 2013), Spanish (Travis 2007:115–16; on Colombian Spanish, Torres Cacoullos and Travis 2011: 252 on New Mexican Spanish; Posio 2011, Posio 2014 on Peninsular Spanish) and European Portuguese (Posio 2014).

Another line of research has focused on the use of 1st and 2nd person subjects in conjunction with verbs of cognition. They have been shown to form fixed units that can be described as prefabricated expressions or prefabs (see e.g. Travis and Torres Cacoullos 2012, Posio 2011; for the term prefabricated expression or prefab, see Erman and Warren 2000). Helasvuo (2014b) offers evidence that several prefabs, formed with 1st person pronominal subjects and certain verbs of cognition, are found in Finnish. Although these constructions with 1st person pronominal subjects and verbs of cognition have become crystallized, they nevertheless have their own internal structure and are not fully lexicalized units. The components of these constructions retain associations with other occurrences of the same lexical elements. (Helasvuo 2014b: 77) This is typical of prefabs in general: prefabs are associated with the more general construction from which they have arisen (e.g. Bybee 2010: 36).

In our study, all verbs were analyzed in terms of their semantics and argument structure (Dixon 2005; Pajunen 2001). We considered three levels for the variable verb type: verbs of cognition, verbs of motion, and other. Verbs describing emotional states and processes were included under verbs of cognition. The motivation behind this coding scheme is twofold. First, it aligns with previous studies: the association between verbs of cognition and 1st person singular pronoun has primarily been investigated in contrast to other verb types. Second, Helasvuo (2014b) offers evidence that verbs of motion show a positive association with zero subjects in Finnish conversation. This three-way encoding schema allows us to contrast verbs of cognition and motion while controlling for other verb types, at the same time avoiding issues related to the sparseness of the data. Based on previous research, we can thus expect the pronominal subject to favor verbs of cognition, while the zero subject will be associated with the verbs of motion.

In a classic article, Hopper and Thompson (1980) analyzed the role of transitivity in grammar and discourse. To achieve a better account of cross-linguistic patterns of transitivity, they proposed that transitivity is best viewed as a scalar phenomenon. They further suggested that clauses high on the transitivity scale are foregrounded in discourse, while clauses with low transitivity remain in the background. One of the factors they associated with high transitivity was the number of participants in the situation depicted in a clause (Hopper and Thompson 1980). In a more recent paper, Thompson and Hopper (2001) revisit the notion of transitivity but now from the perspective of conversational discourse. Based on an analysis of American English conversation, they show that clauses in conversational discourse are overall quite low on the transitivity scale. In their data, intransitive verbal clauses, copular clauses and epistemic/evidential clauses dominate (Thompson and Hopper 2001: 51).[3] The last-mentioned type involves verbs of cognition. Thompson and Hopper (2001: 54) conclude that everyday (English) conversation mainly consists of one-participant clauses and prefabs.

Since our focus here is on clauses with 1st person singular subjects, we have operationalized transitivity somewhat differently from Thompson and Hopper (2001). Earlier research has shown that copular clauses are quite rare in the first person, while clauses with transitive verbs are much more frequent (Helasvuo 2001: 85–88). It is important to note that the verbs most commonly used in transitive clauses in conversational Finnish are verbs which in principle take clausal complements. However, when used as prefabs they commonly do not take any complements at all (Helasvuo 2014b). In the analysis, transitivity was encoded for two levels, transitive (tr) and intransitive (intr), with copular clauses grouped together with intransitives. In the light of prior research, zero subjects are more likely to occur with intransitive verbs.

Travis and Torres Cacoullos (2012) report a combined effect of tense, aspect, and mood on the realization of 1st person subjects in Colombian Spanish. They relate this effect to the conceptualization of the event as backgrounded or foregrounded. In Finnish, however, grammatical aspect is not morphologically marked on the verb. Instead, aspectual distinctions are expressed for example through nominal encoding, such as object marking (for a recent discussion, see Huumo 2010), or through adverbs carrying aspectual meanings. The coding of aspect is thus more diverse, and aspect as a variable would be hard to operationalize in our model; we therefore do not include it in our analysis. Based on previous research, there is reason to believe that tense may be a relevant variable; Lindström et al. (2009) found that in their dialectal data for Estonian, a language closely related to Finnish, 1st person zero subjects were favored over pronominal ones in the past tense. We therefore included the following levels for tense in the analysis: present (prs), past (pst), present perfect (prf), and past perfect (psp). Zero subjects were expected to be favored over pronominal ones in the past tense.

Finally, polarity was encoded for two levels: affirmative (aff) and negative (neg). Negation is generally considered to have a contrastive function in discourse (e.g. Sun and Givón 1985: 346). If the contrastive function is, indeed, the primary interactional pattern associated with clausal negation, it is plausible to assume that pronominal subjects will form the preferred type of subject expression with negation (cf. section 2 for the relationship between pronominal subjects and contrastive function). However, Thompson (1998) shows with support from cross-linguistic literature that standard negation does not participate in any systematic interactional patterns. Using empirical data from American English conversation, Thompson shows that instead, negative clauses are primarily used to deny an event or state which is not usually made either explicit nor implicit in the context. (Thompson 1998: 325–326). Similarly, Travis and Torres Cacoullos (2012) note that polarity does not influence the realization of 1st person subjects in Colombian Spanish. It thus seems to be an open question

[3] It is worth noting that the database used in Thompson and Hopper (2001) was relatively small (446 clauses).

whether polarity and specifically negation affects the realization of subject expression in the 1st person. If polarity is, indeed, associated with contrastive function, we would expect zero subjects to be more likely to appear in affirmative syntactic units compared to pronominal ones. sd

3.2 Discourse and cognitive factors

There is a mass of evidence showing that the placement and choice of referring expressions is related to the structure of discourse and to the general cognitive capacity of speakers (see Kibrik, 2011). As a general framework, Givón (1983) proposed the concept of referential continuity, encompassing the placement of referring expressions and the type of grammatical devices utilized to encode them in discourse (see also Garnham et al. 1982). One aspect of referential continuity is referential distance, i.e. the distance between the current occurrence of a referring expression and its previous mention in the discourse (Givón 1983: 13).

Lindström et al. (2009) studied referential distance and its possible impact on subject expression in Estonian, whose coding strategies for subject are similar to those of Finnish. More specifically, they investigated usage patterns of 1st person verbal and pronominal subject markers in Estonian dialects. They found that the use of the 1st person subject pronoun depended on whether or not the 1st person referent was referred to in the preceding clause.

In addition to referential distance, referential continuity also connects to the concept of the mental accessibility of referring expressions. This concept has been developed in the framework of accessibility theory (Ariel 1988, Ariel 1990, Ariel 2004). According to accessibility theory, different encodings of referring expressions correspond to how easily a referent can be retrieved from memory. These coding strategies can be ordered on a scale from the most accessible to the least so. The most accessible referents are encoded with the least amount of material, i.e. zero forms, followed by pronominal forms. The least accessible referents are encoded with the most material, i.e. definite descriptions such as proper nouns.

In our study, referential distance was calculated for both pronominal and zero subjects. Distance is a continuous variable, measuring the number of syntactic units between the current occurrence and the previous mention of the referent in the discourse. Subject expressions which were either first mentions or could not be traced back in the conversation were encoded with a value of zero (*n* = 19). This type of realization is illustrated in example (4), which belonged to the very first syntactic units in the recorded conversation. Example (5) illustrates a case with several references to the same referent.

Example 4 (SG398)

1 Kati: niin to:ta: **mä** **e-n** tunne tä-tä kirjailija-a
PTC PTC PRO.1SG NEG-1SG know-CONNEG this-PAR author-PAR
‘So I don’t know’
2 ollenkaa et#
at.all COMP
‘this writer at all’

Example 5 (SaPu118)
1 Ulla: ku sato niim paljo
when rain.PST.3SG so much
‘As it was raining so heavily’
2 ni **mä lait-i-n** tota noi,

so I put-PST-1SG PTC PTC
'I put'
3 heh siis sato ihan nii
PTC rain.PST.3SG quite so
'huh huh so it was raining so [hard]'
4 et laineht-i lattia,
COMP flood-PST.3SGS floor.NOM
'that the floor was flooded'
5 ni **mä pisti** niinko muavikassi-st-ki sit
so I put-PST.1SG PTC plastic.bag-ELA-CLI then
'so I put on from a plastic bag then'
6 kengä-t **te-i-n** niinko itse-lle-ni kahde-st muavikassi-st
shoe-PL make-PST-1SG PTC myself-ALL-1SG.PX two-ELA plastic.bag-ELA
'I made shoes for myself out of two plastic bags'
7 ett-ei ny ihan kastu-nuj ja,
COMP-NEG.3SG now quite get.wet-PCP and
'so that one [i.e. I] wouldn't get all wet and'

Example (4) illustrates a pronominal subject which is produced in the first clause of the opening turn of the recorded conversation. The referential distance of the pronominal subject *mä* 'I' on line 1 was coded as 0. In Example (5) there are several instances of first person subjects (lines 2, 5, 6). The pronominal subject on line 5 has a referential distance of 3, since there are two syntactic units between it and the previous mention of the referent on line 2. The two intervening syntactic units are in lines 3 (*siis sato ihan nii* 'so it was raining so hard') and 4 (*et lainehti lattia* 'that the floor was flooded'). Line 6 contains a first-person form of the verb *tein* 'made' with a zero subject whose referential distance is one, since the previous mention of the same referent is in the immediately preceding syntactic unit (line 5).

The measurement we have used differs from that proposed by Givón (1983) in one crucial aspect, i.e. the minimum value the measurement can receive. Givón (1983: 13) proposes that the minimum value of this measure is 1. In his analysis, referents which cannot be located in previous discourse are given the maximum value used to define the referential distance. However, at least for pronouns, we consider a minimum value of zero to be a more natural interpretation based on the concept of accessibility. If a pronoun is indeed used for a first mention in a discourse, its referent must be assumed to be sufficiently identifiable. For the purposes of this article, we are limiting our focus to 1st person subjects, which as a rule are used by speakers to refer to themselves. In our analysis, a referential distance of 1 means that the referent can be located in the immediately preceding syntactic unit, whereas a distance of 0 means that the occurrence is the first mention of this referent in the recorded conversation. In practice, this means that it is a particular speaker's first reference to him- or herself. In this vein, the scale used in this study is a genuine continuous measurement, incorporating the concept of accessibility and distance. On the basis of previous work on referential distance, zero subjects are expected to be favored in environments where the referential distance is slight.

The complexity of a syntactic unit has previously been attributed to differences in production. Sternberg et al. (1978), using word lists, demonstrated that the more words in an utterance, the longer it takes speakers to initiate it. From this perspective, it has been proposed that the minimization of syntactic complexity for comprehension is one of the crucial factors influencing syntactic choices in production (Arnold et al. 2000; Hawkins 2004). In previous studies, a number of different measures of complexity have been proposed: for example the number of (phonological) words, the number of nodes in a dependency tree (Ferreira 1991), or the degree of "embeddedness". For the last-mentioned type, complexity would be defined, for example, by comparing the number of embedded/ subordinate structures relative to simple or conjoined structures in a discourse as previously proposed inter alia by Beaman (1984) and Givón (1991). Different measures of syntactic complexity, however, appear to be highly correlated (cf. Szmrecsányi 2004; Wasow 2002). The simplest possible measure of complexity is perhaps the number of words. It is easy to obtain, thus facilitating for instance cross-linguistic and/or genre-specific comparisons. Syntactic units containing zero subject are inherently shorter ($M = 3.9$, $SD = 1.9$) than units with a pronominal subject ($M = 5.27$, $SD = 2.24$): $t(277.06) = 9.45$, $p < 0.0001$). To avoid showing this fairly trivial relationship, the potential role of complexity was operationalized as a relative measure (see Bresnan & Ford, 2010). More specifically, we calculated this variable as the difference in length of a syntactic unit and the immediately preceding syntactic unit expressed on a scale of natural logarithm. This is a ratio variable $\log(\mathrm{A}) - \log(\mathrm{B}) = \mathrm{A}/\mathrm{B}$, where negative values indicate that the preceding syntactic unit was relatively longer and vice versa. In the following, we refer to this variable as relative s-unit length.[4] Zero subjects are expected to be more likely to be chosen in less complex syntactic contexts than pronominal subjects.

An important aspect related to syntactic production is the speakers' tendency to recycle already produced material. This phenomenon is referred to as persistence or priming. It has been linked to syntactic alternations inter alia by Bresnan and Ford (2010), and to person marking by Torres Cacoullos and Travis (2011; see also Travis 2007; Gries 2005). Previous studies indicate that the effect of persistence is semantic in nature; it does not depend on lexical repetition, but is enhanced by it. Indeed, Travis and Torres Cacoullos (2012) show that persistence also influences the occurrence of 1st person singular zero subject in Spanish conversation. In the case of zero subjects, there is no lexical overlap, while for function words, such as pronouns, persistence has been shown to correlate with accessibility (Ferreira, 2003). It is important to bear in mind that 1st and 2nd person pronouns differ from 3rd person pronouns in one important respect: they index speech act participants, while 3rd person pronouns typically index previously mentioned referents. Due to issues related to data sparseness, persistence was encoded as a binary variable. If a 1st person singular pronoun was used in the immediately preceding syntactic unit, this was coded as "pronoun"; in all other instances the label "other" was used. A 1st person singular pronominal subject is thus expected to be chosen more often if the immediately preceding syntactic unit also contains a 1st person singular pronominal subject.

The final variable to be considered in this study is turn length: a continuous measure of the number of syntactic units produced during one turn by a speaker in a given conversation. If economy was indeed a factor affecting production, we would expect an increase in the probability of choosing a zero subject with greater turn length (cf. Levinson, 2007 on economy). In the following, we refer to this variable as turn length.

3.3. Summary

We have analyzed our data with regard to constructional and discourse/cognitive factors. The choice of these factors was based on the previous literature, as discussed in Sections 3.1 and

[4] In three conversations, a first person subject was produced in the first syntactic unit. In these cases, the length of the preceding syntactic unit is missing. To avoid having missing values for this variable, the mean length of the syntactic units in that particular conversation was used to impute these three missing values.

3.2. Although these factors have been proposed in previous studies, they have not been applied in a systematic manner. Before investigating the contribution of these factors to the choice of 1st person singular nominative subject expressions (see Section 4), possible issues related to data sparseness were explored. With regard to tense, only nine instances were attested for past perfect. These data points were removed from the data set because a realistic estimation cannot be expected. Further removal of data was not carried out. The final data set thus consisted of 1779 syntactic units.

A summary of information regarding the variables available in this final data set is given in Table 1. The variables are grouped into categorical and continuous ones. For categorical variables, treatment (dummy) encoding was used where the levels of the categorical variable are compared to the reference level of the variable. The reference category was chosen not only on the basis of theoretical considerations but also to ensure that the summary output of the statistical model presented in Section 4 will show those contrasts which are the most relevant for the purposes of the current study. The first level of each variable given in Table 1 is the reference level for that particular categorical variable.

Part A: Continuous variables	min.	max.	*M*	*SD*
Referential distance	0	4	2.23	0.94
Relative s-unit length	− 2.2	2.64	0.27	0.79
Turn length	1	29	2.89	3.2
Part B: Categorical variables	*n*			
Subject				
pronoun	1577			
zero	202			
Clause type				
transitive (tr)	921			
intransitive (intr)	858			
Tense				
present (prs)	930			
past (pst)	689			
present perfect (prf)	160			
Polarity				
affirmative (aff)	1320			
negative (neg)	459			
Verb type				
cognition	566			
motion	414			
other	799			
Persistence				
other (than 1st person singular pronoun)	1421			
1st person singular pronoun	358			

Table 1: Summary information of the variables available in the final data set.

Table 1 shows that pronominal subjects (*n* = 1577) are favored in conversational Finnish over zero subjects (*n* = 202) (see also Duvallon 2006, Lappalainen 2004; Helasvuo 2014a, Helasvuo 2014b). This distribution is highly skewed towards double-marking, indicating that it is indeed conventionalized in conversational Finnish. It is an open question, however, which factors influence the choice of subject expression and under which specific circumstances single-marking is chosen. This question is explored in the following section, where a mixed-effects logistic regression is fitted to the data.

4 Mixed-effects logistic regression model of subject expression in the first person singular

We employed a mixed-effects logistic regression analysis for the data. In general, a logistic regression is a model for the probability of an event. It can be used to model a categorical binary response variable and to test whether the response variable is mediated by the predictors included in the model (Baayen et al. 2008; Harrell 2001; Jaeger 2008). In this study, we were interested in modeling the probability of a zero subject in a given syntactic unit; our purpose was to show how the zero subject is related to the predictors described in Section 3. In this respect, a logistic model is closely related to variable rules analysis (Varbrul) (see Sankoff 1988; Tagliamonte and Baayen 2012), which has been used to model subject expression in coversational data (see Travis and Torres Cacoullos 2012; Torres Cacoullos and Travis 2011).

4.1. Why use mixed-effects logistic regression model?

There are a number of benefits related to using a logistic model. First, we can simultaneously include multiple predictors instead of testing them individually. This makes it possible to test multiple competing hypotheses while controlling for others. For example, we can test whether referential distance influences the choice between pronominal vs. zero subject, above and beyond the possible effect of tense or verb type. Second, we can estimate the direction of an effect, i.e. whether a given predictor has a positive or negative influence on the choice of subject (pronoun vs. zero). Third, a logistic model can handle both categorical variables such as tense and continuous ones such as referential distance. The discretization of continuous variables can lead to a loss of statistical power, so that the true effect is obscured (Cohen 1983).

A mixed-effects model is an even more flexible alternative to the traditional logistic regression model, allowing us to include both fixed- and random-effects to handle the inherent variability of spontaneous conversational data (see Tagliamonte and Baayen 2012 for discussion of variation in general). Fixed-effects for categorical variables exhaust all possible levels of the variable and are estimated to represent the population as a whole. In contrast, random-effects constitute only a subset of the larger population. For example, individual speakers sampled in conversational studies represent only a small proportion of a larger population of speakers. Furthermore, certain speakers can produce more syntactic units than others, thus increasing their contribution to the probability of producing syntactic units with zero subjects. Finally, some speakers may simply prefer using the zero subject as an idiosyncratic preference. For example, the data in this study comprise 48 individual speakers, with the number of syntactic units produced by them ranging from 1 to 116. A mixed-effects model allows us to include these sources of variation in the model.

In addition to individual speakers, other sources of variation should also be considered when working with spontaneous conversational data. The conversations themselves represent another source of variation. One such source is the duration of the conversations included in the database; variation in duration influences the number of syntactic units produced, ranging from 58 to 283. Finally, in spontaneous conversational data the rate of verbs produced cannot be controlled, representing yet another source of variation. The data set contains 267 unique verbs (lemmata). However, only a small number of verbs appeared frequently in the data; these included *tietää* 'to know' (*n* = 279) and *ajatella* 'to think' (*n* = 132). Taken together, these three factors create dependences among the data points, thus challenging the assumption of most statistical methods, according to which the data points are assumed to be independent. A mixed-effects model allows us to include and handle these dependencies in a principled manner, consisting of both fixed- and random-effects in a single model. However, it is an empirical question whether a random structure is warranted by the data, and if so how complex it should be. Specifically, the aim is to achieve a balance in the model between precision and parsimony, given the data.

4.2 Model fitting and results

We followed the model fitting procedure proposed in Zuur et al. (2011). All subsequent statistical models and visualizations were carried out in R (R Development Core Team, 2014). According to this procedure, the structure of the random effects is estimated first to avoid attributing any of the variability that could be explained by the fixed effects to the random structure of the model. This is achieved by fitting a maximally complex fixed effect structure afforded by the data and then estimating the random effect structure. Before estimating the random effects, we visually inspected the variation in subject marking across speakers, verbs and conversations. Based on this, the variable speaker appeared to display the greatest variation, suggesting that at least random intercepts for speakers might be required for these data. First, an initial mixed-effects logistic regression model was fitted to the data using the function glmer in the R package lmer4 (Bates et al. 2014). In this model, the response variable subject (either pronoun or zero) was modeled as a function of referential distance, turn length, persistence, polarity, verb type, clause type, tense, and relative s-unit length. Additionally, random intercepts were included in this model for the variable speaker. Second, a model was fitted to the data which also included random intercepts for the variable verb. However, this model failed to converge indicating that the model might be too complex for the data.

To simplify the model, a fixed-effects logistic regression model was fitted using the R package rms (Harrell 2014). A backward elimination of the predictors was carried out, in which the least significant predictor was first removed from the model at the conventional α-level of 0.05, after which the data were refitted (Harrell 2001: 58–59). This procedure was carried out until only statistically significant predictors remained in the model. The following predictors were removed from the model: turn length, persistence, polarity and tense. This fitting procedure yielded a model where the response variable subject was modeled as a function of referential distance, verb type, clause type, and relative s-unit length (results not shown). We will refer to this model as the fixed-effects model. It is worth noting that although this elimination procedure did not contain any random effects, the eliminated predictors did not reach statistical significance in the initial mixed effects model.

Given this simplified fixed-effects structure, we continued with the model fitting procedure by estimating the random effect structure by fitting the following models: a random intercept model for speaker, a random intercept model for speaker and conversation, a random intercept model for speaker and verb, and, finally, a full random intercept model for speaker, conversation and verb. All these models had the same, previously estimated fixed-effects structure. We used the Akaike information criterion (AIC) to rank the models in terms of how much information is lost when a specific model was used to approximate the full reality given the data (Akaike 1974). The AIC values of these five models are shown in Table 2. ΔAIC is provided as well, showing the difference between a particular model and the best-fitting model, in our case the mixed-effects model with random intercepts for speaker. Burnham and Anderson (2002) have proposed that as a rule of thumb ΔAIC can be interpreted as follows: $\Delta < 2$ suggests substantial evidence for a particular model, values ranging between 3–7 indicate considerably less support, and $\Delta > 10$ indicates that the model in question is very unlikely.

	AIC	ΔAIC
mixed-effects (speaker)	1123	NA
mixed-effects (speaker and conversation)	1123	0
mixed-effects (speaker and verb)	1124	1
mixed-effects (speaker, conversation and verb)	1124	1
fixed-effects	1137	14

Table 2: Ranking of the fitted models based on AIC and ΔAIC, rounded off to the nearest whole number.

The ranking of these models indicates that the random intercepts for the variable speaker were fully warranted by the data, compared to the model that contained the fixed effects alone. As can be seen in Table 2, the difference in the AIC is large, 14. This difference can also be expressed in terms of AIC weights, which indicate how often, given the data, a particular model would be selected as the most likely in a set of models. The mixed-effects model with speaker has an AIC weight of 0.99 relative to the fixed-effects model. It is thus estimated to be the most likely model 99% of the time compared to the fixed-effects one. In contrast, inclusion of the variables conversation or verb as random effects added very little information in addition to the variable speaker. Given the differences in AIC values, the best-fitting model for these data was the mixed-effects model with random intercepts for the speakers. This model assumes that the fixed effects are the same across participants, but the random intercepts allow the speakers to have a different baseline preference for variation in subject expression.

The estimated fixed and random effects of this final model are shown in Table 3, on a log-odds scale in keeping with the standard practice. Positive coefficient signs indicate a preference for zero subject; negative signs display a dispreference. In this final model, the intercept represents a constellation of properties of the syntactic units: more specifically, transitive verbs, verbs of cognition, and cases where referential distance and relative s-unit length have the value zero. Given this constellation of properties, zero subject expression was dispreferred, and was estimated as chosen with a probability of 13%.

The partial effects of this final model are visualized in Figure 1 and were back-transformed to a probability scale. The partial effects show the estimated effect of a given predictor when the other predictors are held constant (categorical predictors at their reference level, continuous predictors at their median value).

	Coefficient	*SE*	*Z*	*p* -value
Intercept	-1.896	0.2772	-6.8406	< 0.0001
Referential distance	-0.4186	0.0848	-4.9346	< 0.0001
Verb type: motion	0.6235	0.2254	2.7668	0.0057
Verb type: other	0.2778	0.2037	1.3638	0.1726
Clause type: intr	0.7755	0.1679	4.618	< 0.0001
Relative s-unit length	-0.8565	0.1216	-7.0438	< 0.0001

Table 3: Estimated coefficients of the final mixed-effects logistic regression model with random intercepts for speaker ($s^2 = 0.52$).

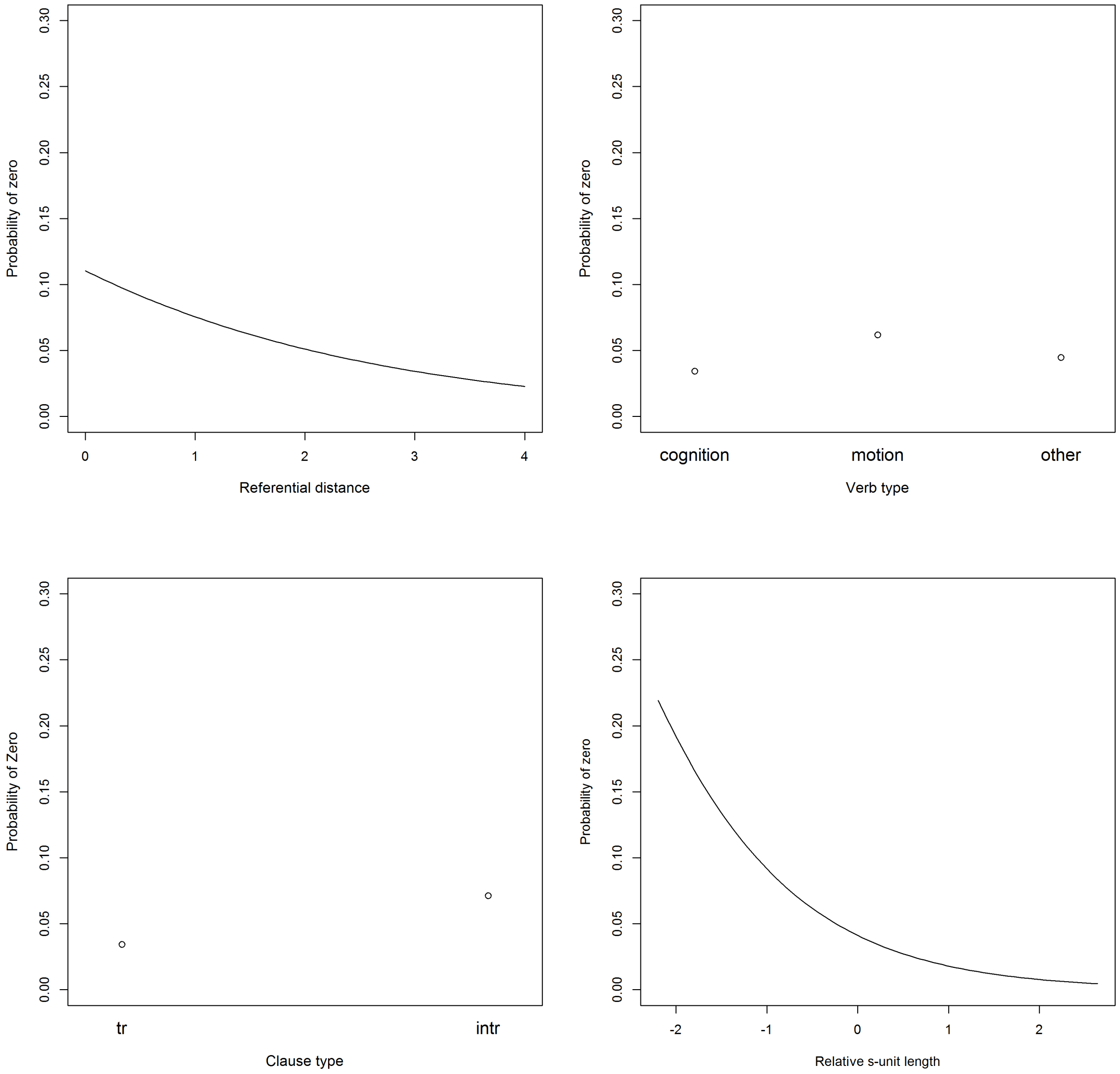

Figure 1: Estimated partial effects of the fixed effects in the final mixed-effects logistic regression.

The effect of referential distance shows that zero subjects are preferred in environments where the distance is either zero (first mention) or one (last mention in the immediately preceding syntactic unit). The estimated partial effect of this predictor is visualized in Figure 1 (upper left panel). The probability of producing a zero subject declines steadily as referential distance increases. The range of this effect, the difference between the minimum and the maximum value, was estimated as approximately 9%, indicating a modest effect size in these data.

With regard to constructional properties, the results indicate the following. The difference between verbs of cognition and motion is statistically significant (see Table 3). Zero subjects were more likely to be produced with motion verbs (observed frequencies: zero = 63 versus pronoun = 351) than with verbs of cognition (observed frequencies: zero = 48 versus pronoun = 518) when controlling for other verb types available in the data. This result is aligned with previous studies, where verbs of cognition appear to favor pronominal subjects (e.g. Travis & Torres Cacoullos 2012). However, the range was estimated as approximately 3%, indicating a small effect size (see Figure 1, upper right panel). Another factor influencing the choice of subject expression is clause type: intransitive verbs (observed frequencies: zero = 135 versus pronoun = 723) were more likely to be produced with a zero subject in the subject position than transitive verbs (observed frequencies: zero = 67 versus pronoun = 854) (see Table 3). Similar to verb type, the effect size with regard to clause type was estimated as small, approximately 4%. The estimated partial effect is visualized in Figure 1 (lower left panel).

The estimated contrasts for verb type did not factor in the fact that multiple tests were performed: the contrast between verb of cognition and verb of motion, verb of cognition and other, and verb of motion and other. On this basis, a post-hoc test was carried out to test the robustness of these contrasts using Tukey's honest significant difference test (Kramer 1956; Tukey 1994 [1953]), implemented in the R package multcomp (Hothorn et al. 2014). The test compares all pair-wise comparisons, adjusting for multiple comparisons. The results of this test are visualized in Figure 2, along with 95% confidence intervals.

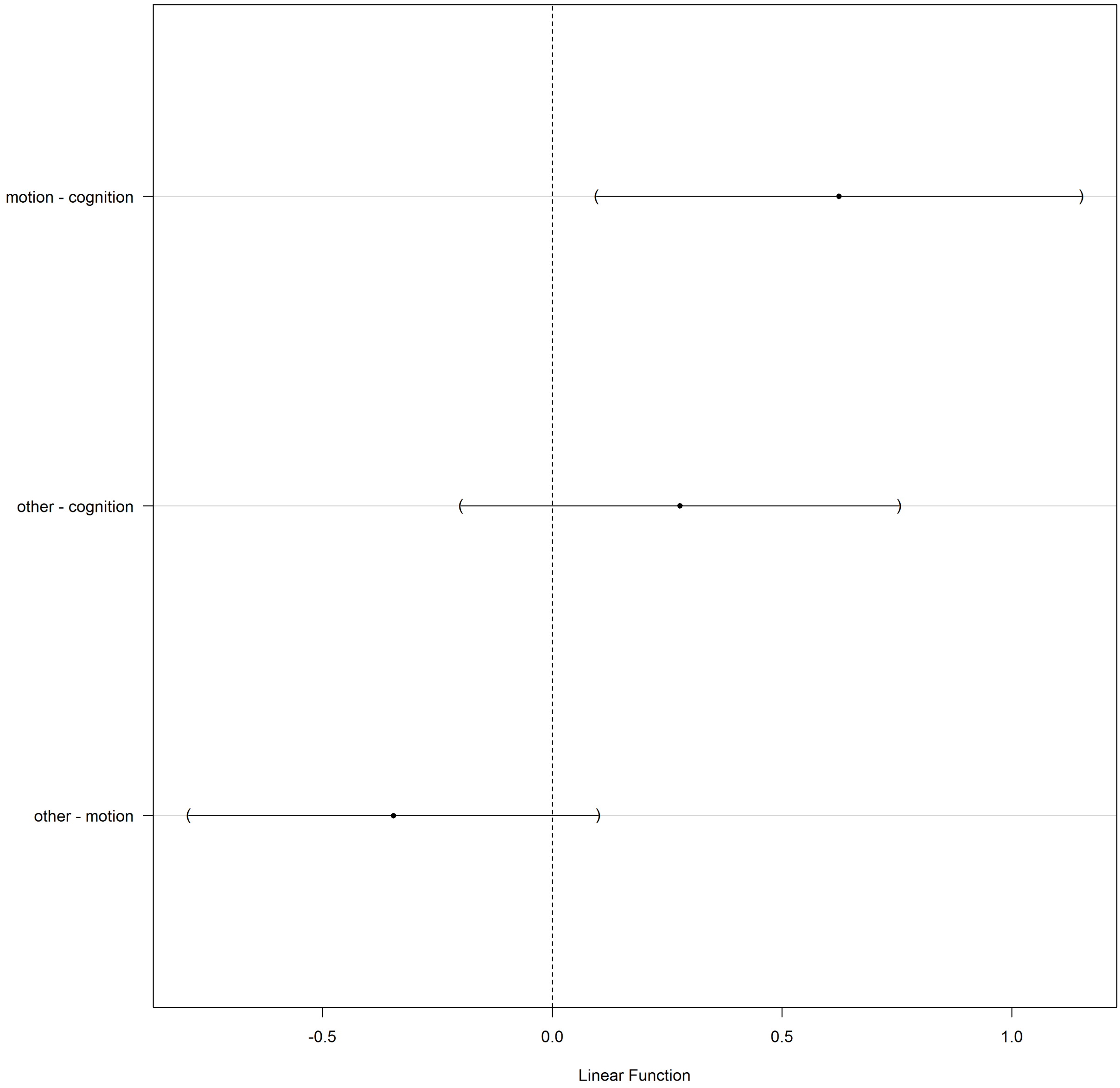

Figure 2: Post-hoc comparison for verb type with 95% confidence intervals.

After adjusting for multiple comparisons, the confidence intervals for the difference between verbs of cognition and motion do not include zero on the log-odds scale, as indicated in Figure 2 by the vertical dotted line. This shows that the estimated difference was robust. The other contrasts, however, are not statistically significant, as they include zero on the log-odds scale.

Finally, the choice of subject expression appeared to be sensitive to the relative length of the syntactic unit. The variable relative s-unit length indicated that the probability of zero subject depends on the ratio between the length of the current syntactic unit and the unit preceding it. Specifically, zero subjects are more likely when the preceding syntactic unit is longer than the current one. The range of this effect was estimated as 21%; the large effect size is visible in Figure 1 (lower right panel). The effect of this predictor is thus relatively large compared to the other predictors in the model, indicating a strong sensitivity to the relative complexity of syntactic units in discourse.

5 Discussion

Although pronominal subjects are much more common in Finnish conversational discourse in general, there appear to be certain syntactic contexts in which zero subjects are favored. We have characterized these contexts in terms of discourse and cognitive factors and constructional properties. The analysis shows that the choice of type of subject expression (pronominal vs. zero) is sensitive to the general discourse/cognitive factors and constructional properties of the syntactic unit in question. We have suggested that the choice of subject expression in conversation is sensitive to probability distributions across and within syntactic units. This suggestion is consistent with interactional studies of grammar, such as Thompson and Couper-Kuhlen (2005), Barth-Weingarten and Couper-Kuhlen 2009 and Helasvuo (2001, 2004) according to which grammar is a matter of knowing how to do things together. This knowledge is shared by the speakers. Grammar is not a monolithic matter, separate from use; rather, it emerges out of language use as on on-line process (see especially Hopper 1987, Hopper 1988, Auer 2009). The role of discourse/cognitive and constructional factors investigated in this study emphasizes the importance on the one hand of sequential organization, on the other of general formats, in the choice of subject expression.

In this study, speakers' sensitivity to the sequential organization of discourse was most strongly associated with relative syntactic complexity and referential distance. The importance of these two factors is highlighted by their associated effect sizes: 21% for relative complexity and 9% for referential distance, respectively. In less complex syntactic contexts, the zero subject is a more probable choice. This preference is probably related to production and comprehension in conversation, as both aspects have to be attended to in order to maintain successful communication. In terms of production, it is faster to initiate a less complex utterance, and single-marking probably optimizes this even further. A simple syntactic unit also contains fewer units for interlocutors to process (see Arnold et al. 2000). In these cases, it seems that single-marking is deemed sufficient for communicative purposes. Interestingly, it is single-marking that is favored in less complex syntactic contexts. Based on literature on economy in discourse one would expect that the zero subject would be chosen in more complex syntactic contexts to reduce overall complexity. In our data, however, the opposite effect is observed. Furthermore, the data showed that the number of syntactic units produced by a speaker in a single turn was not statistically significant in the final model. We suggested that this variable could be interpreted as indexing economy during production. Although the analysis presented here cannot be used to prove a null hypothesis, i.e. no effect of turn length, the results nonetheless indicate that the effect size might be small and an even larger sample would be required to obtain an effect. In addition to syntactic complexity, the choice of subject expression is influenced by referential distance. If the choice of subject expression were simply related to repetition, we would have expected that choice to be affected by the type of subject expression used in the preceding syntactic unit, i.e. persistence, but this is not supported by our findings. This is in contrast with the findings reported in Torres Cacoullos and Travis (2011) for first person singular subject expression in Spanish. However, given the imbalance between the two subject markings in the data this could simply represent a floor effect. A new data set would be required to test this effect. Importantly, our data show that zero subjects are more likely to be used when the referential distance is short. This result is aligned with the findings of Helasvuo (2014a), who describes same-subject coordination (cf. ex. 1 in section 2), list construction (ex. 2), and the latter parts of adjacency pairs (ex. 3) as typical contexts for zero subjects.

Our results regarding the variables associated with constructional factors are in line with previous studies showing that verbs of cognition primarily appear with pronominal subjects (Helasvuo 2014a: 461–462). However, previous studies have for the most part investigated only this type of association, while the role of other possible variables related to verbal semantics have received relatively little attention. The results reported here broaden the scope. It was shown that zero subjects are favored with verbs of motion. In spoken discourse, this type of verb tends to be used in a narrative function. In narrative contexts, it might be assumed that zero subjects are sufficient for communicative purposes. This suggests that there are different mechanisms influencing the choice of type of subject expression. The association between verbs of cognition and subject expression in the 1st person singular is based on certain high-frequency fixed expressions, i.e. prefabs (Helasvuo 2014b), which typically express the speaker's epistemic stance (see Helasvuo 2014b for Finnish; Kärkkäinen 2003, Kärkkäinen 2007 for English; Keevallik 2003 for Estonian). Verbs of motion, on the other hand, function very differently: they carry the discourse forward. In narrative contexts the series of events is contingent upon its component parts, with a focus on certain prominent performers (protagonists). Here, zero subject appears to be sufficient for encoding the subject in conversational Finnish.

With regard to clause type, our analysis showed that intransitive clauses are more likely to have zero subjects than transitive clauses. This finding is linked to our results concerning verb types; transitive clauses often contain verbs of cognition, more specifically certain high frequency verbs. These verbs tend to occur in prefabs, in which, interestingly, it is the 1st person form with a pronominal subject that has become crystallized. This finding is in line with Thompson and Hopper's (2001: 51) study of transitivity in American English conversation, where transitives are dominated by epistemic/evidential clauses, typically formed with verbs of cognition. Helasvuo (2014b), however, has shown that even though these verbs are in principle complement-taking predicates, they often appear without a complement and are thus low in transitivity (for the transitivity scale, see Hopper and Thompson 1980; Thompson and Hopper 2001).

In terms of constructional properties, polarity and tense were not statistically significant in the final model. We may note that Travis and Torres Cacoullos (2012) also failed to find an effect in Spanish. Our study thus offers cumulative evidence that polarity does not appear to influence the 1st person singular subject marking in conversational discourse. Regarding tense, it is possible that we simply do not have enough data to find an effect. This explanation seems plausible considering the small effect sizes associated with the constructional variables in the final model. Thus, we feel that there is no reason to speculate about the role of this variable in these data.

The results of the statistical analysis are aligned with previous empirical studies on subject expression, which have shown that the choice of subject expression cannot be reduced to a few general principles (e.g. Kibrik 2011; Travis and Torres Cacoullos 2012). Interestingly, these results appear to be consistent across Spanish and Finnish, even though the default subject marking is different in the two languages, i.e. double-marking is the default for conversational Finnish, whereas in Spanish, it is single-marking. The results presented here, nonetheless, bring forth an issue that has not been systematically investigated before, namely, the effect sizes associated with the variables used in the analysis. Specifically, Travis and Torres Cacoullos (2012) show that double-marking of 1st person singular subject is strongly associated with cognitive verbs and persistence (double-marking of the previous realization) in Spanish. Thus, the choice of subject expression appears to be associated with both cognitive/discourse and constructional factors in Spanish when a marking other than the default is chosen. In contrast, the data presented here indicate that in Finnish single-marking is most strongly associated with cognitive and discourse factors. In this respect, choosing other than the default marking in Finnish appears to be primarily driven by accumulation of recent experience in conversation.

It is worth noting that frequency of use might influence the choice of subject expression, especially in the above-mentioned contexts. After all, frequency of use is one of the best indices of accumulation of experience over time. With estimations of frequency it would be possible to differentiate the effects associated with accumulation of experience over time (constructional factors) and recent experience (cognitive/discourse factors). However, currently no data are available for a reliable estimation of frequency in conversational Finnish, or in spoken Finnish in general. Estimations based on written Finnish are problematic because of the dominant usage of single-marking (see section 2 above). This topic will have to wait until the Arkisyn database becomes available (see Helasvuo 2014e). Finally, we would also like to point out that there are a number of other possible variables that might influence the choice of the subject expression in conversation. These include for example phonological processes, such as assimilation and erosion, discussed in Helasvuo (2014b). These factors, however, are problematic to operationalize reliably for the purposes of large-scale statistical analysis.

6 Conclusion

We have presented a large-scale statistical analysis of the variation in nominative 1st person subject expression in conversational Finnish using mixed-effects logistic regression, with a focus on factors affecting the choice between pronominal and zero subjects. In the 1st person singular, the verb always carries person marking encoding the number and person of the subject. To the best of our knowledge, this is the first study to present a large-scale quantitative analysis of this type of variation in Finnish.

From a typological perspective, Finnish exhibits a mixed-type in terms of subject marking, since the 1st and 2nd person behave differently compared to 3rd person marking (cf. Dryer 2011 and Sections 1 and 2). In addition, the dominant subject-marking strategy in the 1st and 2nd person is double-marking in conversational Finnish, i.e. pronominal subject (1st or 2nd person singular pronoun) and subject marker on the verb. Indeed, the rate of zero subjects in our data was only 11% (see Section 3). The results reported here offer support for the view that the choice of subject expression can be modeled based on variables which are also relevant in explaining the variation of subject expression observed in other language types, for example in predominantly single-marking languages such as Spanish (cf. Travis and Torres Cacoullos 2012).

The statistical analysis supports the results obtained in previous empirical studies, according to which variation in subject expression cannot be reduced to a few general principles of language usage. Rather, subject expression is influenced by a number of factors (cf. e.g. Travis and Torres Cacoullos 2012, Helasvuo 2014a). The primary factors affecting subject expression were grouped in this study into two types: discourse/cognitive factors and constructional ones.

To sum up our findings concerning discourse/cognitive factors: we have shown that the choice of the 1st person nominative subject is influenced by the sequential organization of conversation. In conversational Finnish, a zero subject is more likely to be chosen when the syntactic context is relatively simple. Similarly, zero subjects are preferred when the referential distance is short, i.e. when the referent of the subject was mentioned in the preceding syntactic unit. These findings were related to clausal organization (same-subject coordination, list constructions) and sequential structure (adjacency pairs).

With regard to constructional factors, we have shown that verbs of cognition primarily appear with pronominal subjects, while zero subjects tend to co-occur with verbs of motion. The association of cognitive verbs with pronominal subjects was shown to be based on certain prefabs with high frequency. The co-occurrence of verbs of motion and zero subjects was suggested to be related to their use in narrative contexts.

The results presented here emphasize the importance of the sequential structure of the conversation, including in the case of 1st person singular subject marking. Further studies are required, however, to test whether the same principles apply to subject marking strategies in connection with other kinds of subjects. More specifically, factors influencing the choice of subject expression in the 3rd person singular might be most revealing, as the 3rd person is used to encode a wide range of different referents in discourse.

Bibliography

Akaike, Hirotugu. 1974. A new look at the statistical model identification. *IEEE transactions on automatic control* 19(6). 716–723.

Ariel, Mira. 1988. Referring and accessibility. *Journal of Linguistics* 24(1). 65–87.

Ariel, Mira. 1990. *Accessing noun-phrase antecedents*. London: Routledge.

Ariel, Mira. 2004. Accessibility marking: Discourse functions, discourse profiles, and processing cues. *Discourse Processes* 37(2). 91–116.

Arnold, Jennifer, Wasow, Thomas, Losongco, Anthony, & Ginstrom, Ryan. 2000. Heaviness vs. newness: The effects of complexity and information structure on constituent ordering. *Language* 76(1). 28–55.

Auer, Peter. 2009. On-Line Syntax: Thoughts on the Temporality of Spoken Language. *Language Sciences* 31.1: 1–13.

Baayen, R. H., Davidson, D. J., & Bates, D. M. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59(4). 390–412.

Barth-Weingarten, Dagmar and Elizabeth Couper-Kuhlen. 2011. Action, prosody and emergent constructions: The case of 'and'. In Peter Auer and Stefan Pfänder, eds., *Constructions: Emerging and emergent*, 263–292. Berlin: de Gruyter.

Bates, Douglas, Martin Maechler, Ben Bolker, Steven Walker, Rune Haubo B. Christensen, Henrik Singmann & Bin Dai. 2014. *lme4: Linear mixed-effects models using Eigen and S4*. Retrieved from http://cran.r-project.org/web/packages/lme4/index.html

Beaman, Karen. 1984. Coordination and subordination revisited: Syntactic complexity in spoken and written narrative discourse. In Deborah Tannen (ed.), *Coherence in spoken and written discourse*, 45–80. Norwood: Ablex.

Bresnan, Joan & Marilyn Ford. 2010. Predicting syntax: Processing dative constructions in American and Australian varieties of English. *Language* 86(1). 168–213.

Burnham, Kenneth P. & David R. Anderson. 2002. *Model selection and multimodel inference: A practical information-theoretic approach* (2nd edn.). New York: Springer.

Bybee, Joan L. 2010. *Language, usage and cognition*. Cambridge: Cambridge University Press.

Cohen, Jakob. 1983. The cost of dichotomization. *Applied Psychological Measurement* 7(3). 249–254.

Dahl, Östen. 1990. Standard Average European as an exotic language. In Johannes Bechert, Giuliano Bernini & Claude Buridant (eds.), *Toward a typology of European languages*, 3–8. Berlin: Mouton de Gruyter.

Dixon, R. M. W. (2005). *A new approach to English grammar on semantic principles*. Oxford: Oxford University Press.

Dryer, Matthew S. 2013. Expression of pronominal subjects. In Matthew S. Dryer & Martin Haspelmath (eds.), *The world atlas of language structures online*. Max Planck Institute for Evolutionary Anthropology. Retrieved October 28, 2014, from http://wals.info/chapter/101

Duvallon, Outi. 2006. Milloin pronominisubjekti jää pois puhutussa suomessa [When is pronominal subject is omitted in spoken Finnish]. In Anneli Pajunen & Hannu Tommola (eds.), *XXXII Kielitieteen päivät Tampereella*, 203–217. Tampere: Tampere UP.

Endo, Tomoko. 2010. Epistemic stance marker as a disagreement preface: *wo juede* 'I feel/think' in Mandarin conversation in response to assessments. *Kyoto University Linguistic Research* 29. 43–76.

Endo, Tomoko. 2013. Epistemic stance in Mandarin conversation: The positions and functions of wo juede 'I think'. In Yuling Pan and Daniel Kádár (eds.), *Chinese Discourse and Interaction: Theory and Practice*, 12–34. London: Equinox.

Erman, Britt and Beatrice Warren. 2000. The idiom principle and the open choice principle. *Text* 20(1). 29–62.

Ferreira, Fernanda. 1991. Effects of length and syntactic complexity on initiation times for prepared utterances. *Journal of Memory and Language* 30(2). 210–233.

Ferreira, Victor S. 2003. The persistence of optional complementizer production: Why saying "that" is not saying "that" at all. *Journal of Memory and Language* 48(2). 379–398.

Garnham, Alan, Jane Oakhill & Philip N. Johnson-Laird. 1982. Referential continuity and the coherence of discourse. *Cognition* 11(1). 29–46.

Givón, Talmy. 1983. Topic continuity in discourse: An introduction. In Talmy Givón (ed.), *Topic continuity in discourse: A quantitative cross-language study*, 1–43. Amsterdam: John Benjamins Publishing Company.

Givón, Talmy. 1991. Markedness in grammar: Distributional, communicative and cognitive correlates of syntactic structure. *Studies in Language* 15(2). 335–370.

Gries, Stefan Th. 2005. Syntactic priming: A corpus-based approach. *Journal of Psycholinguistic Research* 34(4). 227–294.

Hacohen, Gonen & Schegloff, Emanuel A. 2006. On the preference for minimization in referring to persons: Evidence from Hebrew conversation. *Journal of Pragmatics* 38(8). 1305–1312.

Harrell, Frank E., Jr. 2001. *Regression modeling strategies. With applications to linear models, logistic regression, and survival analysis*. New York: Springer.

Harrell, Frank E., Jr. 2014. *rms: Regression modeling strategies*. Retrieved from http://cran.r-project.org/web/packages/rms/index.html

Hawkins, John A. 2004. *Efficiency and complexity in grammars*. Oxford: Oxford University Press.

Helasvuo, Marja-Liisa. 2001. *Syntax in the making: The emergence of syntactic units in Finnish conversation*. Amsterdam: Benjamins.

Helasvuo, Marja-Liisa. 2004. Shared syntax: The grammar of co-constructions. *Journal of Pragmatics* 36(8). 1315–1336.

Helasvuo, Marja-Liisa. 2014a. Searching for motivations for grammatical patterns. *Pragmatics* 24(3). 453–476.

Helasvuo, Marja-Liisa. 2014b. Agreement or crystallization: Patterns of 1st and 2nd person subjects and verbs of cognition in Finnish conversational interaction. *Journal of Pragmatics* 63. 63–78.

Helasvuo, Marja-Liisa. 2014c. Jotta suomalaiset voisivat puhua enemmän. Puhetilanteen osallistujat tekstiviestikeskustelussa [So that Finns could speak more. Speech act participants in text message interaction]. In Marja-Liisa Helasvuo, Marjut Johansson & Sanna-Kaisa Tanskanen (eds.), *Kieli verkossa. Näkökulmia digitaaliseen* vuorovaikutukseen, 29–49. Helsinki: Finnish Literature Society.

Helasvuo, Marja-Liisa. 2014d. Subjektin ilmaiseminen verkkojuttelussa [Subject expression in chat interaction]. Paper presented at the conference Kielitieteen päivät, University of Turku, May 9, 2014.

Helasvuo, Marja-Liisa. 2014e. Arkisyn: A morphosyntactically coded database of conversational Finnish. Paper presented at the seminar on Conversational corpora. Center of Excellence on Research on Intersubjectivity, University of Helsinki, September 10, 2014.

Hopper, Paul J. 1987. Emergent grammar. *Thirteenth annual meeting, Berkeley Linguistic Society* (BLS). 139–157. Berkeley: Berkeley Linguistic Society.

Hopper, Paul J. 1988. Emergent grammar and the A Priori Grammar postulate. In Deborah Tannen (ed.), *Linguistics in Context*, 117–134. Norwood, NJ: Ablex.

Hopper, Paul J. & Sandra A. Thompson. 1980. Transitivity in grammar and discourse. *Language* 56(2). 251–299.

Hothorn, Torsten, Frank Bretz, Peter Westfall, Richard M. Heiberger & Andre Schuetzenmeister. 2014. *multcomp: Simultaneous inference in general parametric models*. Retrieved from http://cran.r-project.org/web/packages/multcomp/index.html

Huumo, Tuomas. 2010. Nominal aspect, quantity, and time: The case of the Finnish object. *Journal of Linguistics* 46(1). 83–125.

Jaeger, T. Florian. 2008. Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language* 59(4). 434–446.

Kaltenböck, Gunther. 2007. Position, prosody, and scope: The case of English comment clauses. *Vienna English Working Papers* 16(1). 3–38.

Keevallik, Leelo. 2003. *From interaction to grammar. Estonian finite verb forms in conversation*. Studia Uralica Upsaliensia,vol. 34. Uppsala: Acta Universitatis Upsaliensis.

Kibrik, Andrej A. 2011. *Reference in discourse*. Oxford: Oxford University Press.

Kärkkäinen, Elise. 2003. *Epistemic stance in English conversation*. Amsterdam: Benjamins.

Kärkkäinen, Elise. 2007. The role of *I guess* in conversational stancetaking. In Robert Englebretson (ed.), *Stancetaking in discourse: Subjectivity, evaluation, interaction*, 183–219. Amsterdam: Benjamins.

Kramer, C. Y. 1956. Extension of multiple range tests to group correlated adjusted means. *Biometrics* 13. 13–18.

Lappalainen, Hanna. 2004. *Variaatio ja sen funktiot. Erään sosiaalisen verkoston jäsenten kielellisen variaation ja vuorovaikutuksen tarkastelua* [Variation and its functions: The analysis of linguistic variation and interaction among members of a social network]. Helsinki: Finnish Literature Society.

Levinson, Stephen C. 2007. Optimizing person reference – perspectives from usage on Rossel Island. In Nick J. Enfield & Tanya Stivers (eds.), *Person reference in interaction: Linguistic, cultural, and social perspectives*, 29–72. Cambridge: Cambridge University Press.

Lindström, Liina, Mervi Kalmus, Anneliis Klaus, Liisi Bakhoff & Karl Pajusalu. 2009. Ainsuse 1. isikule viitamine eesti murretes [The first person singular reference in Estonian dialects]. *Emakeele Seltsi aastaraamat* 54. 159–185.

Meriläinen, Hanna. 2011. Yksikön ensimmäisen persoonan muotojen käyttö IRC-keskusteluissa [The use of first person singular forms in IRC chats]. Turku: University of Turku MA thesis.

Pajunen, Anneli. 2001. *Argumenttirakenne* [Argument structure]. Helsinki: Finnish Literature Society.

Posio, Pekka. 2011. Spanish subject pronoun usage and verb semantics revisited: First and second person singular subject pronouns and focusing of attention in spoken Peninsular Spanish. *Journal of Pragmatics* 43. 777–798.

Posio, Pekka. 2014. Subject expression in grammaticalizing constructions: The case of *creo* and *acho* ‘I think’ in Spanish and Portuguese. *Journal of Pragmatics* 63. 5–18.

R Development Core Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

Sacks, Harvey & Emanuel A. Schegloff. 1979. Two preferences in the organization of reference to persons in conversation and their interaction. In G. Psathas (ed.), *Everyday language: Studies in ethnomethodology*, 15–21. New York: Irvington.

Sankoff, David. 1988. Variable rules. In Ulrich Ammon, Norbert Dittmar & Klaus J. Mattheier (eds.), *Sociolinguistics: An international handbook of the science of language and society*, vol. 2, 984–997. Berlin & New York: Walter de Gruyter.

Scheibman, Joanne. 2002. *Point of view and grammar. Structural patterns of subjectivity in American English conversation*. Amsterdam: Benjamins.

Siewierska, Anna. 1999. From anaphoric pronoun to grammatical agreement marker: Why objects don't make it. *Folia Linguistica* 33(2). 225–251.

Sorjonen, Marja-Leena. 2001. Simple answers to polar questions: The case of Finnish. In Margret Selting & Elizabeth Couper-Kuhlen (eds.), *Studies in interactional linguistics*, 405–431. Amsterdam: Benjamins.

Sternberg, Saul, Stephen Monsell, Ronald L. Knoll & Charles E. Wright. 1978. The latency and duration of rapid movement sequences: Comparisons of speech and typewriting. In G. E. Stelmach (ed.), *Information processing in motor control and learning*, 117–152. New York: Academic Press.

Strellman, Urpu. 2005. *Persoonapronominin liikakäyttö: Normin synty ja muotoutuminen* [The overuse of personal pronouns: The birth and formation of a norm]. Helsinki: University of Helsinki MA thesis.

Sulkala, Helena & Merja Karjalainen. (1992). *Finnish. Descriptive grammars*. New York: Routledge.

Sun, Chao & Talmy Givón. 1985. On the so-called SOV word order in Mandarin Chinese: A quantified text study and its implications. *Language* 61(2). 329–351.

Szmrecsányi, Benedikt M. 2004. On operationalizing syntactic complexity. In Gérald Purnelle, Cédrick Fairon & Anne Dister (eds.), *7th international conference on textual data statistical analysis*, 1032–1039. Louvain-la-Neuve: Presses universitaires de Louvain.

Tagliamonte, Sali A. & R. Harald Baayen. 2012. Model, forests and trees of York English: *Was/were* variation as a case study for statistical practice. *Language Variation and Change* 24(2). 135–178.

Tao, Hongyin. 1996. *Units in Mandarin conversation*. Amsterdam: Benjamins.

Tao, Hongyin. 2001. Discovering the usual with corpora: The case of remember. In Rita C. Simpson & John M. Swales (eds.), *Corpus linguistics in North America*, 116–144. Ann Arbor: University of Michigan Press.

Thompson, Sandra A. & Elizabeth Couper-Kuhlen. 2005. The clause as a locus of grammar and interaction. *Discourse Studies* 7(4). 481–505.

Thompson, Sandra A. & Paul J. Hopper 2001. Transitivity, clause structure, and argument structure: Evidence from conversation. In Joan L. Bybee & Paul Hopper (eds.), *Frequency and the emergency of linguistic structure*, 27–60. Amsterdam: Benjamins.

Thompson, Sandra A. 1998. A discourse explanation for the cross-linguistic differences in the grammar of interrogation and negation. In Anna Siewierska & Jung Jae Song (eds.), *Case, typology and grammar: In honor of Barry J. Blake*, 309–341. Amsterdam: Benjamins.

Torres Cacoullos, Rena & Catherine E. Travis. 2011. Testing convergence via code-switching: Priming and the structure of variable subject expression. *International Journal of Bilingualism* 15(3). 241–267.

Travis, Catherine E. 2007. Genre effects on subject expression in Spanish: Priming in narrative and conversation. *Language Variation and Change* 19(2). 101–135.

Travis, Catherine E. & Rena Torres Cacoullos. 2012. What do subject pronouns do in discourse? Cognitive, mechanical and constructional factors in variation. *Cognitive Linguistics* 23(4). 711–748.

Tukey, John W. 1994 [1953]. *The problem of multiple comparisons*. Unpublished manuscript. In Henry I. Braun (ed.), The Collected Works of John W. Tukey. Multiple Comparisons: 1948–1983, vol. VIII, 1–300. New York: Chapman and Hall.

Wasow, Thomas. 2002. *Postverbal behavior*. Standford: CSLI Publications.

Zuur, Alain F., Elena N. Ieno, Neil J.Walker, Anatoly A. Saveliev, & Graham M. Smith. 2011. *Mixed effects models and extensions in ecology with R*. New York: Springer.