Jorrig Vogels and Geertje Van Bergen

# Where to place inaccessible subjects in Dutch: The role of definiteness and animacy

**Abstract:** Cross-linguistically, both subjects and topical information tend to be placed at the beginning of a sentence. Subjects are generally highly topical, causing both tendencies to converge on the same word order. However, subjects that lack prototypical topic properties may give rise to an incongruence between the preference to start a sentence with the subject and the preference to start a sentence with the most accessible information. We present a corpus study in which we investigate in what syntactic position (preverbal or postverbal) such low-accessible subjects are typically found in Dutch natural language. We examine the effects of both discourse accessibility (definiteness) and inherent accessibility (animacy). Our results show that definiteness and animacy interact in determining subject position in Dutch. Non-referential (bare) subjects are less likely to occur in preverbal position than definite subjects, and this tendency is reinforced when the subject is inanimate. This suggests that these two properties that make the subject less accessible together can 'gang up' against the subject first preference. The results support a probabilistic multifactorial account of syntactic variation.

**Keywords:** word order, subject, topic, definiteness, animacy, Dutch

**Jorrig Vogels:** Communication and Information Sciences, Tilburg University, Tilburg, Netherlands. E-mail: j.vogels@uvt.nl
**Geertje Van Bergen:** Centre for Language Studies, Radboud University Nijmegen, Nijmegen, Netherlands. E-mail: g.vanbergen@let.ru.nl

# 1 Introduction

When transforming conceptualizations into linguistic utterances, speakers have to structure these concepts in a linear order (Levelt 1989). Assuming that speech production proceeds incrementally, concepts will be transferred to a syntactic structure as soon as they become available to the production process (Kempen and Hoenkamp 1987). In some languages, this process is heavily constrained by the grammar. Which linearizations of constituents are possible in such languages is dependent on which grammatical functions, such as subject and object, the

concepts are assigned to. In a subject-prominent language such as English, for example, sentences almost always start with the subject. In line with this, the *grammatical function model* (e.g. Bock and Warren 1985) states that the most accessible concept is assigned the subject role, which is subsequently produced as the first constituent in the sentence. According to the *positional model* (e.g. Branigan and Feleki 1999), however, there is a direct relation between the accessibility of concepts and their position in the sentence: the most accessible information is produced first. For example, in languages in which word order is more free, such as German (Kempen and Harbusch 2004), Greek (Branigan and Feleki 1999), Hungarian (É. Kiss 2002), Italian and Spanish (Brunetti 2009), different permutations of the constituents in a sentence are possible, independent of grammatical function.

Accessibility is often driven by information structure (Lambrecht 1994). What is accessible is what people tend to talk about, i.e., what is topical (Reinhart 1982). Different properties can make an entity more or less topical. For example, an entity constituting given information is typically topical, because people are likely to talk about things relating to the preceding discourse (e.g. Gundel 1988). Givón (1976) proposed a universal hierarchy of topicality based on a number of conceptual properties, such as animacy, definiteness and agentivity. Each of these prototypical topic properties itself constitutes a sub-hierarchy of accessibility, entities higher in these hierarchies being more conceptually accessible.

Languages differ in the degree to which constituent linearization is constrained by grammatical rules or driven by information structure (e.g. Gundel 1988). In a subject-prominent language such as English, for instance, placing non-subject topics in the sentence-initial position is possible, but word order variation related to information structure is fairly limited. In a topic-prominent language such as Russian, the sentence-initial position is still the canonical subject position (Comrie 1989), but word order is much more free. Dutch is a language with a relatively free word order, in which grammatical function and information structure are both considered important factors in determining how to start a sentence (Bouma 2008). There is a strong tendency to start a sentence with the subject (henceforth *Subject First preference*; e.g. Bouma 2008; De Hoop and Krämer 2006; Van Tiel and Lamers 2007; Vogels and Lamers 2008), but the position directly preceding the finite verb in main clauses may contain other constituents (Bouma 2008; Haeseryn et al. 1997). An example is given in (1).

(1) *Die    hoed    draagt Piet    graag.*
    that   hat     wears Piet    gladly
    'As for that hat, Piet likes to wear it.'

In the sentence in (1), the direct object *die hoed* 'that hat' is placed sentence-initially (henceforth preverbal position), while the subject *Piet* follows the finite verb (henceforth postverbal position). This constituent order may arise because the speaker wants to talk about a particular hat, making the concept of 'hat' highly accessible, and hence the corresponding noun phrase is produced early in the sentence. The choice of the verb form (*draagt* 'wears') then dictates that this constituent be the direct object. It might also be the case that the concept expressed by the subject has certain properties that make it less accessible, and is therefore postponed to a later position in the sentence. In many cases, the Subject First preference and the tendency to place accessible information first will converge, as subjects generally express accessible information (for example, subjects are often definite or animate). Yet, subjects need not be highly accessible. When a subject is not accessible, the two forces are in conflict.

In the present study, we will investigate how this conflict is resolved in Dutch natural language production. On the basis of corpus data from the *Corpus Gesproken Nederlands* (Corpus of Spoken Dutch, henceforth CGN), we examine whether and how the Subject First preference can be overruled by accessibility effects in natural language production. We will thereby focus on effects of definiteness and animacy. Definiteness relates to how accessible a referent is in the current discourse. Animacy is an inherent property of concepts, which also influences accessibility, as suggested by findings from other studies (e.g. Brunetti 2009, van Bergen 2011, Christianson and Ferreira 2005, Prat-Sala and Branigan 2000, Van Nice and Dietrich 2003). We investigate to what degree the interplay between animacy and definiteness can explain variation in subject placement in Dutch. Definite and animate subjects are predicted to be preferred in the preverbal, canonical subject position, whereas less accessible (i.e., inanimate and/or indefinite) subjects are predicted to show a weaker tendency to occur in preverbal position. Yet, depending on the strength of the Subject First preference, accessibility effects may be overruled. In addition, different accessibility measures may have different strengths. For example, an entity's *derived* accessibility (i.e., its discourse prominence) may be a stronger competitor for the Subject First preference than its *inherent* accessibility (cf. Prat-Sala and Branigan 2000). Another question is whether effects of animacy and definiteness are absolute or relative to the properties of other entities (cf. Branigan et al. 2008). That is, are inanimate or indefinite subjects in preverbal position inherently dispreferred, or only in the presence of an animate or definite argument? To address this question, we will focus on sentences with either a preverbal subject or a preverbal PP adjunct, but not another preverbal argument.

This article is organized as follows. In Section 2, we will discuss the factors that we expect to influence word order in Dutch main clauses; in Section 3, we

will present the results of a multifactorial regression model. Section 4 provides a general discussion and conclusions are given in Section 5.

# 2 Factors affecting subject placement

In this section, we will give a description of the factors that we take into account in our corpus study, and discuss their predicted effects on subject placement in Dutch. Unless specified otherwise, examples in this section are taken from the *Corpus Gesproken Nederlands* (Corpus of Spoken Dutch, CGN).

## 2.1 Definiteness of the subject

It is commonly assumed that the accessibility of entities in the discourse is marked by the choice of a certain referential expression (e.g. Ariel 1990; Gundel et al. 1993). Definite referential expressions (including pronouns and proper nouns) often express topical concepts, because they typically mark given information, i.e. referents that are already activated in the minds of the speaker and the hearer (e.g. Epstein 2002; Givón 1976, 1979; Gundel et al. 1993). Indefinite expressions typically introduce new information that is not yet active in the minds of the speaker and the hearer, and hence do not normally express topical information. The relationship between definiteness and information status is expressed in the definiteness scale in (2) (Aissen 2003; Croft 1988; Givón 1976; Gundel et al. 1993).

(2) *Definiteness scale:*
Personal pronoun > Proper noun > Definite NP > Indefinite specific NP > Non-specific NP

At the left end of the scale, personal pronouns often refer to topical information, and therefore mark referents that are highly accessible in the discourse (e.g. Grosz et al. 1995). At the other end of the scale, non-specific indefinite NPs mark non-topical information, and hence denote inaccessible referents. It has been shown that entities with a high *derived* accessibility, i.e. accessibility determined by the discourse context, are produced early in the sentence (e.g. Bock and Irwin 1980; Ferreira and Yoshita 2003; Prat-Sala and Branigan 2000). Conversely, entities marked by indefinite expressions are predicted to be dispreferred in the sentence-initial position.

We focus on definite plural NPs, which typically express the existence of referents that are accessible from the preceding discourse, and bare plural NPs, which do not assume existence of such referents. In contrast to most singular indefinite NPs, bare plural subjects cannot get a referential reading in Dutch. Therefore, they necessarily express concepts that are inaccessible from the discourse. Thus, bare plural subjects are predicted to be less likely to occur preverbally than definite plural subjects. Following the Subject First preference, however, bare plural subjects do prefer the preverbal position, which may lead to variation in word order. This is illustrated by the sentences in (3), taken from Vogels and Lamers (2008). In (3a), the bare plural subject *kinderen* 'children' is in the preverbal position, even though it expresses inaccessible information.[1] The inaccessible subject in (3b) is in a less prominent, postverbal position, but violates the Subject First preference.

(3) a. *Kinderen   spelen in   de   tuin.*
children   play   in   the   garden
'Children play in the garden.'
   b. *In   de   tuin   spelen kinderen.*
in   the   garden play   children
'Children are playing in the garden.'

For the definite subjects in (4), by contrast, no clash between preferences is predicted. Since definite subjects express accessible information, the preverbal position (as in (4a)) will be preferred over the postverbal position (as in (4b)).

(4) a. *De   kinderen   spelen in   de   tuin.*
the   children   play   in   the   garden
'The children are playing in the garden.'

---

**1** The preverbal position may trigger a generic reading of the bare plural, in which case the noun phrase refers to a category of concepts that is accessible from world knowledge (e.g. 'penguins live on the south pole'). It is assumed that this is a marked reading for a bare plural (Krifka et al. 1995). For example, in (3a), the sentence-initial bare plural subject is likely to receive a marked, generic interpretation ('in general, children play in the garden'), while the postverbal bare plural subject in (3b) gets its unmarked, existential interpretation ('there are some unspecified children playing in the garden'). Still, in neither reading the bare plural refers to specific entities that are accessible from the preceding discourse. In addition, note that the preverbal position does not automatically result in a generic reading, as shown by example (8) in Section 2.2. For further discussion of the relation between the different interpretations of bare plural subjects and syntactic position in Dutch, see Vogels and Lamers (2008).

b. *In de tuin spelen de kinderen.*
in the garden play the children
'In the garden, the children are playing.'

Evidence that bare subjects occur less frequently in the preverbal position than definite subjects comes from both experimental (Vogels and Lamers 2008) and corpus data (Bouma 2008). In a language production experiment, Vogels and Lamers (2008) investigated word order preferences of Dutch participants. In a drag-and-drop task, participants produced short sentences such as those in (3) and (4), consisting of a definite or bare plural subject, an intransitive verb and a locative PP adjunct. They could choose one of two possible word orders: NP-Verb-PP (as in (3a) and (4a)), or PP-Verb-NP (as in (3b) and (4b)). The results showed that definite NPs were indeed produced more frequently in preverbal position than bare NPs. Vogels and Lamers (2008) argued that in sentences with bare plural subjects neither word order counts as marked. In contrast, a definite plural subject in a postverbal position was taken to be marked compared to a definite subject in its canonical position. The former would probably need a special context to be acceptable, for instance one in which *in de tuin* 'in the garden' is contrasted with another location. Thus, it was claimed that (4b) is a case of topicalization (cf. (1)), or contrastive focus (Bouma 2008; Choi 1999), while in (3b) the non-canonical word order is due to the inaccessibility of the subject. In addition, the results showed that even though bare NPs were more frequent in the postverbal position than definite NPs, in both conditions the NP was produced most frequently in the preverbal position. This suggests that the preference to start the sentence with the subject was overall more important than the preference to start the sentence with an accessible entity.

The results of the study by Vogels and Lamers (2008) support the hypothesis that bare plural subjects cause an incongruence between the Subject First preference and the tendency to start a sentence with accessible concepts, which results in word order variation. To investigate whether these results can be generalized to natural language, we analyze the position of bare and definite plural subjects in our corpus. In addition, since different kinds of bare and definite NPs exist, we employ a more fine-grained subdivision of definiteness. For definite plural NPs, we distinguish, in accordance with the Givenness Hierarchy (Gundel et al. 1993), between NPs preceded by a strong quantifier, such as *sommige* 'some', *alle* 'all', or *de meeste* 'most' (De Hoop 1992), NPs preceded by a definite article, NPs preceded by a possessive pronoun, and NPs preceded by a demonstrative pronoun. Although we consider all types of definite NP to be more topical than indefinite NPs, we predict that the more referential a definite NP is, i.e. the more unambiguously it refers to specific individuals in the discourse, the more likely it is to be in

preverbal position. For example, although strongly quantified NPs are usually referential, their referents are not uniquely identifiable. In the sentence in (5), the subject *de meeste personages* 'most characters' refers to a subset of characters from Greek tragedies, but precisely which characters are meant remains unspecified.

(5)  *Allee   bij   de   Griekse   tragedies   uhm   sterven ook   de   meeste*
      INTERJ by   the   Greek   tragedies   HESIT die   also   the   most
      *uhm   personages.*   [fv400165.32][2]
      HESIT   characters
      'In the Greek tragedies, most characters in fact die.'
      (INTERJ = interjection; HESIT = hesitation)

    NPs preceded by a demonstrative pronoun, on the other hand, are prototypically deictic expressions. As such, the use of a demonstrative pronoun indicates that the referent is known or inferable by the addressee (Clark et al. 1983). Indeed, Gundel et al. (1993) list familiarity of the hearer with the referent as a prerequisite for demonstrative reference. As a result, subjects preceded by demonstrative pronouns are predicted to be more topical than other types of definites.
    Within the category of bare plurals, we distinguish between modified and unmodified bare NPs. Modified bare plurals can be considered more specific than unmodified bare plurals, because the referent is narrowed down to a more restricted set (e.g. Bosque 2001; Chierchia 1998; Longobardi 2001). For example, in (6), 'Dutch Muslims' refers to a more specific group than 'Muslims'. Therefore, modified bare subjects are considered more topical.

(6)  *Nederlandse   moslims   werden   op   de   site   aangespoord   om*
      Dutch         Muslims   became   on   the   site   urged         to
      *zich         militair   te   trainen bij   schietclubs   en   de*
      themselves   militarily   to   train   at   shooting.clubs and   the
      *landmacht*   [fn001692.3].
      land.forces
      'Dutch Muslims were urged on the website to get a military training at shooting clubs and the land forces.'

---

**2**  The code between square brackets is the name of the session within the CGN from which the utterance is taken. Sessions of which the name starts with *fv* are recorded in Belgium; sessions of which the name starts with *fn* are recorded in the Netherlands.

In sum, if the different degrees of definiteness as defined above give an indication of an entity's derived accessibility, we should find that the higher the degree of definiteness of the subject, the more likely it will be that a subject occurs in its canonical, preverbal position.

## 2.2  Animacy of the subject

The accessibility of a referent is not only determined by its givenness or familiarity in the discourse, but also by its intrinsic semantic properties (e.g. Bock and Warren 1985). For example, as people tend to talk about human or animate entities, these are more topical than inanimate entities (Givón 1979). Thus, human referents are considered more *inherently* accessible than other animate referents (e.g. animals), which are more accessible than inanimate referents. This is expressed in the prominence scale for animacy (Aissen 2003; Givón 1976), as shown in (7).

(7) *Animacy scale:*
    Human > Animate > Inanimate

It has been shown in a wide range of studies (e.g. Van Bergen 2011; Bresnan et al. 2007; Bresnan and Hay 2008; Dahl and Fraurud 1996; Hendriks et al. 2005; Lamers 2007; Prat-Sala and Branigan 2000; Rosenbach 2005; Van Nice and Dietrich 2003) that animacy affects constituent order: There is a universal tendency for animate NPs to be placed early in the sentence. For example, it has been shown for a number of languages that there is an increase in passive structures in sentences with inanimate agents (e.g. McDonald et al. 1993; Prat-Sala and Branigan 2000; Van Nice and Dietrich 2003). For Dutch, psycholinguistic experiments have shown that in constructions that have an inanimate subject, and in which there is another constituent that is higher in animacy, this other constituent is likely to occupy the sentence-initial position (e.g. *De atleet beviel de foto* 'the picture pleased the athlete' (lit. 'the athlete pleased the picture'), Lamers 2007).

We investigate whether animacy also plays a role in the position of subjects in Dutch natural language, and whether this effect is independent of the animacy of competing entities. We predict that even in the absence of a competing entity the preference for the sentence-initial position is less strong for inanimate subjects, resulting in more postverbal subjects compared to sentences with animate subjects. More importantly, we investigate how animacy interacts with definiteness. If bare plural subjects cause variation in word order due to a conflict between being a subject and being inaccessible at the same time, being animate may

increase a subject's accessibility and moderate the conflict in favour of the pre-verbal position, cf. the sentence in (8). Being inanimate, on the other hand, may decrease accessibility and reinforce the conflict, resulting in more postverbal subjects.

(8) *Mensen    liepen   hier    rond    met    DAT-recorders.*    [fn000058.4]
      people     walked here   around with   DAT-recorders
      'People were walking around here with DAT-recorders.'

If derived accessibility has precedence over inherent accessibility (cf. Prat-Sala and Branigan, 2000) in the placement of subjects in Dutch, we would not expect to find an effect of animacy for definite subjects, because they are pre-dicted to have a strong preference for the preverbal position. An effect of animacy is expected for bare plural subjects, for which this preference is not as pro-nounced. Thus, we predict that the preference to place inanimate subjects in postverbal position is stronger for bare plurals than for definite plurals, since definite plural subjects already have a preference for the preverbal position irre-spective of their animacy.

## 2.3  Type of PP

As mentioned above, we investigate sentences in which the preverbal position is occupied by either the subject or a PP adjunct. We expect the properties of the subject to influence its preference for the preverbal position. When the subject is not accessible, the PP is predicted to be more likely to occupy this position. Yet, this likelihood may depend on the type of PP adjunct. For instance, spatiotemporal PP adjuncts provide a situational setting for the rest of the sentence (e.g. Crasborn et al. 2009; Davison 1984; Van Oosten 1986). Therefore, these types of adjuncts might cause a higher frequency of preverbal PPs than would have occurred on the basis of the properties of the subject alone. PP type also relates to the degree of accessibility or predictability of the subject constituent. Concrete locative PPs can project predictions about upcoming constituents, because many events are bound by certain locations in which they are likely to occur (Grondelaers and Speelman 2007; Grondelaers et al. 2009). As a consequence, locative adjuncts are often fronted, especially when introducing unpredictable events or theme shifts (Vonk et al. 1992). For example, the occurrence of the subject *sterren* 'stars' in (9) is fairly predictable after the locative PP *Aan de hemel . . .* 'In the sky . . .', because the number of possibilities is confined to objects that normally appear in the sky. The set is further narrowed down to shiny objects by the verb *schitteren*

'shine'. In (10), by contrast, the subject *tijdschriften* 'magazines' is not predictable by the abstract PP adjunct *met grote regelmaat* 'with great regularity' and the verb *verdwijnen* 'disappear', since many things can disappear with great regularity.

(9) *Aan de hemel schitteren sterren.* [fn001001.22]
on the sky twinkle stars
'Stars are shining in the sky.'

(10) *Met grote regelmaat verdwijnen tijdschriften* [. . .] [fv600272.3]
with great regularity disappear magazines
'Magazines are disappearing with great regularity.'

In sum, we predict that in sentences with concrete spatiotemporal PP adjuncts, subjects occur more frequently in postverbal position than in sentences with abstract PP adjuncts.

## 2.4 Tense and aspect

The accessibility of an entity may be influenced by properties of the predicate. We restrict ourselves here to possible effects of tense and aspect, because these properties are morphologically visible on the predicate, and hence easy to annotate. Subjects of present tense and imperfective aspect verbs may be more accessible than subjects of past tense and perfective aspect verbs, because what happened in the past is less in focus (e.g. Carreiras et al. 1997). In addition, tense and aspect may affect the interpretation of bare plural subjects (e.g. Carlson 1977, 1982). Verbs with past tense or perfective aspect are more likely to have an episodic reading than verbs with present tense or imperfective aspect. When the predicate expresses an episodic event (stage-level predicate; e.g. 'to be running'), a bare plural subject is likely to get an existential reading. A predicate that expresses a habitual or general statement (individual-level predicate; e.g. 'to be intelligent') is likely to trigger a generic reading. In Dutch, an existential reading of bare plural subjects is often associated with a postverbal position, while a generic reading is associated with the preverbal position (De Hoop 1992). Thus, the predicate might have an influence on the placement of bare plural subjects. For example, the verb *komen* 'come' in (11) has perfective aspect, which may induce an existential interpretation of the bare plural subject *nieuwe feiten* 'new facts'. If the verb would have been imperfective, a generic reading ('in general, new facts come to light during the survey') would also have been possible.

(11) *Tijdens de enquête zijn nieuwe feiten aan 't licht gekomen.*
     during   the   survey   are   new   facts   on   the   light   come
     [fn007226.20]
     'During the survey, new facts came to light.'

Thus, if the tense or aspect of the predicate influences the accessibility of subjects or the interpretation of bare plurals, we expect more postverbal subjects with past and perfective verbs.

## 2.5  Voice

Another prototypical topic property is agentivity. Agents are also often subjects. However, in passive constructions, the agent is either omitted or demoted to an adjunct phrase, and the patient becomes the subject. Hence, subjects of passive verbs are expected to be less topical (Brunetti 2009), by which the Subject First preference and the tendency to start a sentence with accessible concepts are not in accordance with each other. For example, the subject in (12) is *de gasten* 'the guests', which should be preferred in the preverbal position according to the Subject First preference. However, because it is not an agent, it is not a prototypical topic, and this might keep it from this position.

(12) *Via originele uitnodigingen werden de gasten enthousiast*
     through   original   invitations   became   the   guests enthusiastic
     *gemaakt om te komen* [fn001411.9]
     made   to   to   come
     'Through original invitations, the guests were made enthusiastic about coming over.'

Thus, we predict that subjects of passive sentences occur postverbally more often than subjects of active sentences.

## 2.6  Length

It is known that longer constituents tend to be produced later in the sentence (e.g. Bresnan et al. 2007; Wasow 1997). To control for this, we counted the number of words in both the subject and the PP adjunct. We also calculated relative length, by subtracting the length of the PP from the length of the subject.

# 3 Corpus study

## 3.1 Methods

For this study, we used the *Corpus Gesproken Nederlands* (Corpus of Spoken Dutch, CGN). This is a corpus containing spontaneous and prepared speech collected from speakers of Dutch in the Netherlands and Belgium. It consists of about 9 million words. All data has been transcribed orthographically and phonetically, and has been POS-tagged and lemmatized. In addition, about 10 percent of the speech has been annotated syntactically. From this syntactically annotated subcorpus, we extracted all main clauses that had both a plural subject and a PP adjunct, using TIGERSearch 2.1 (König et al. 2003). Next, we searched the part that was not syntactically annotated for sentences with a plural NP and a PP. From the first 1000 hits, we manually selected all main clauses in which the plural NP had the function of subject and the PP was an adjunct. The two resulting data sets were further manually refined as follows: First, we omitted all sentences in which the initial position was occupied by a constituent other than the subject or the PP adjunct. Second, we excluded sentences with subjects modified by relative clauses or preceded by indefinite quantifiers such as *veel* 'many' and *enkele* 'some'. We also excluded sentences with conjoined subjects and subject complements (e.g. *Mensen zijn in de winter depressief, 's zomers niet* 'People are depressed in winter, not in summer' [fn007120.11]). Furthermore, sentences with more than one adjunct or with directional PPs (e.g. *naar de rechter* 'to court'), PP complements or other PPs closely connected to the predicate (e.g. *luisteren naar klassieke muziek* 'listen to classical music') were excluded. Finally, we excluded sentences that contained a postverbal occurrence of existential *er* 'there' (e.g. *In Amerika zijn er dus zomerkampen* 'So in America there are summer camps').[3]

Our final selection contained 541 sentences, which were manually annotated for the nine factors described in the previous section. The coding scheme employed is presented in Table 1. Annotation was performed by two independent

---

**3** According to Grondelaers and Speelman (2007), existential *er* 'there' signals an upcoming inaccessible subject. Since inaccessible subjects are preferably postponed to a non-prominent postverbal position, this would make the presence of existential *er* a good predictor of subject position. However, this predictor would be dependent on factors that make the subject inaccessible in the first place. Indeed, all sixteen sentences with postverbal *er* that we encountered in our initial corpus selection had a bare subject, and all but two had an inanimate subject. Thus, because the presence of *er* cannot be said to be independent from the specificity and animacy of the subject, we excluded these sixteen instances from our selection.

**Table 1:** Operationalization of the factors

| Factors | Values | Cohen's κ |
|---|---|---|
| Definiteness of the subject | [ref]:<br>  [dem] [def]<br>[non-ref]:<br>  [mod bare] [bare] | .98 |
| Animacy of the subject | [anim] [inan] | .96 |
| Type of PP | [loc] [temp] [abs] | .90 |
| Tense | [pres] [past] | .98 |
| Aspect | [imperf] [perf] | .94 |
| Voice | [act] [pass] | .92 |
| Length of the subject | 1–25 words (log transformed) | – |
| Length of the PP | 2–28 words (log transformed) | – |
| Relative length | −21–22 words (log transformed with sign preservation) | – |

annotators. The value of Cohen's κ for inter-annotator agreement is given in the third column. Cases of disagreement were resolved through discussion.

For the factor Definiteness of the subject, definite NPs were coded as *referential* [ref] and bare plurals as *non-referential* [non-ref]. In addition, a more fine-grained division of specificity was applied. Definite plurals were coded as *demonstrative* [dem] (NPs preceded by a demonstrative pronoun), or *definite* [def], in which we collapsed NPs preceded by a definite article, NPs preceded by a strong quantifier, and NPs preceded by a possessive pronoun. Bare plurals were coded as *modified bare* [mod bare] or *non-modified bare* [bare], referring to whether they were preceded or followed by a modifier or not. For the factor Animacy of the subject, we coded subjects as *animate* [anim] or *inanimate* [inan]. Because the category non-human animates (animals, but also NPs referring to companies or organizations with a collective voice that are treated as if they are sentient beings) turned out to have a low number of tokens compared to the other two categories, *human* and *animate* were collapsed into a single category *animate*. PPs were coded as *locative* [loc] ('in the garden'), *temporal* [temp] ('during the survey') or *abstract* [abs] ('in the Greek tragedies'), and finite verbs were coded for tense (*present* [pres] or *past* [past]), aspect (*imperfective* [imperf] or *perfective* [perf])

and voice (*active* [act] or *passive* [pass]). The length of the subject and the PP in number of words was log transformed to reduce the influence of outliers. For their relative length, a constant was added first to preserve the sign. Next, the sentences in the selection were coded for the dependent variable, i.e. whether they had the subject in preverbal position and the PP adjunct in postverbal position (*S-V-PP*), or the PP adjunct in preverbal position and the subject in postverbal position (*PP-V-S*). We determined the frequency with which the two word orders occurred for each level of the nine factors. Frequency differences were tested for significance using Chi-square tests and univariable logistic regression in the case of Length.

## 3.2  Statistical exploration per factor

In our selection, 258 of the 541 sentences (47.7%) contained a preverbal subject, against 283 sentences (52.3%) with a postverbal subject. This was not different from chance (binom(541, 0.5); $p = 0.3021$): subjects appeared equally often in both positions. Hence, we found no clear subject first preference, confirming that there is variation in how to start a sentence in Dutch main clauses with a PP adjunct. In the next sections, the individual effects of the factors introduced above on the subject position will be discussed.

### 3.2.1  Definiteness of the subject

Bare plural subjects in our corpus behaved differently than definite plural subjects with respect to their position in the sentence. Bare plural subjects occurred more often in postverbal position than in the preverbal position, whereas definite plural subjects were more frequent in the preverbal position than in postverbal position. This difference was significant, $\chi^2(1) = 10.4$; $p < .01$. For the four-level distinction of definiteness, the frequency distributions of the two word orders were also significantly different, $\chi^2(3) = 11.9$; $p < .01$. Figure 1 shows a trend towards a gradual increase of subject-initial word orders as the subject's definiteness increases. For unmodified bare plural subjects, the postverbal position was most frequent. Modification caused the proportion of preverbal bare subjects to increase. Subjects preceded by a demonstrative pronoun were most frequent in preverbal position, while other definite subjects occurred about as much preverbally as postverbally.
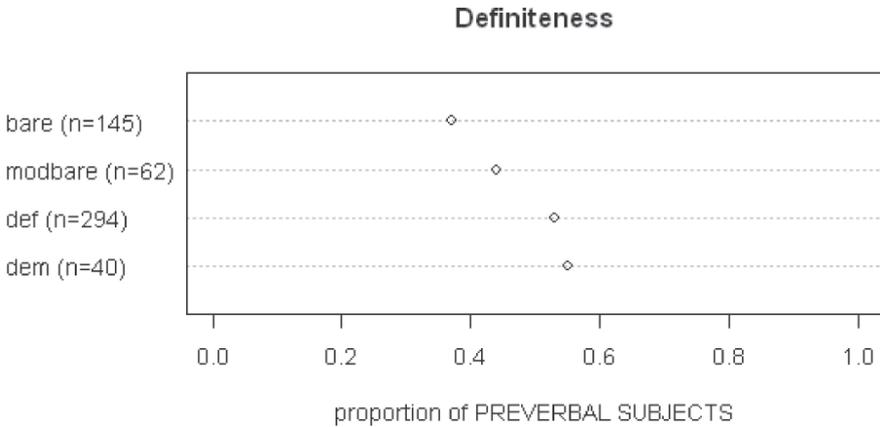
**Fig. 1:** Proportion of preverbal subjects as a function of Definiteness of the subject (*dem* = NP preceded by demonstrative pronoun; *def* = NP preceded by definite article, possessive pronoun or strong quantifier; *modbare* = bare plural NP preceded or followed by a modifier; *bare* = bare plural NP)

### 3.2.2 Animacy of the subject

Inanimate subjects were slightly more frequent in postverbal position than animate subjects, but this difference was not significant, $\chi^2(1) < 1$; $p = .585$. When we examined the effect of animacy for each level of definiteness separately, however, the two factors appeared to interact: Figure 2 suggests that animacy has an effect on the position of demonstrative and bare subjects, the two ends of the definiteness hierarchy. For demonstrative subjects, being animate seems to boost their preference for the preverbal position; for bare subjects, being inanimate seems to boost their preference for the postverbal position, while they occur about as often in preverbal as in postverbal position when they are animate. We will investigate this apparent interaction further when we take into account all factors simultaneously (Section 3.3).

### 3.2.3 Type of PP

Figure 3 shows the frequency of a preverbal subject for each of the three types of PP. There appears to be a difference between spatiotemporal PP adjuncts on the one hand and abstract PP adjuncts on the other: The former are more likely to occur in the preverbal position, with a postverbal subject, while the latter are
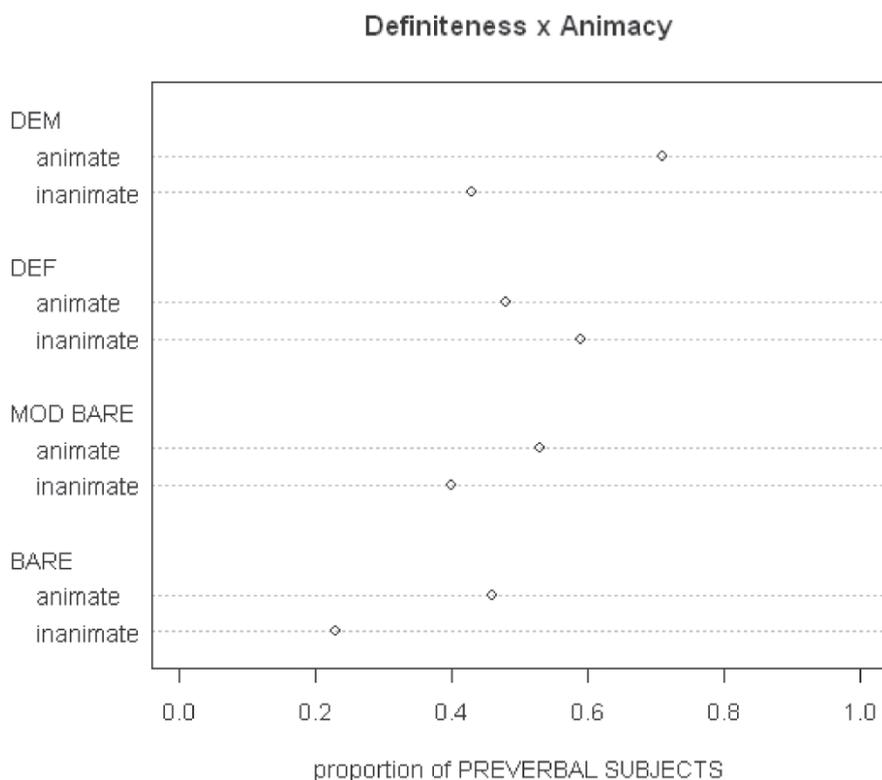
**Fig. 2:** Proportion of preverbal subjects as a function of Animacy of the subject, for each level of Definiteness of the subject.
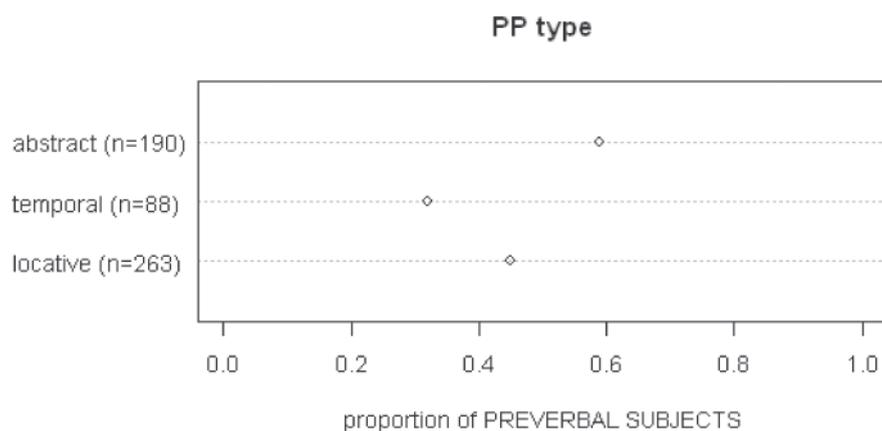


**Fig. 3:** Proportion of preverbal subjects as a function of Type of PP

more likely to occur in postverbal position, with a preverbal subject. There also seems to be a difference between locative and temporal PPs, the latter occurring more frequently in the preverbal position. The association between Type of PP and subject position was significant, $\chi^2(2) = 19.4$; $p < .001$.

### 3.2.4 Tense, Aspect and Voice

The frequency distributions of the two word orders for the three factors that concern properties of the predicate, Tense, Aspect and Voice, showed no significant differences between subjects of present and past tense verbs, verbs with imperfective and perfective aspect and active and passive verbs (Tense: $\chi^2(1) < 1$; $p = .395$; Aspect: $\chi^2(1) = 2.2$; $p = .138$; Voice: $\chi^2(1) < 1$; $p = .796$). This suggests that subject position is not affected by the nature of the predicate (at least for these three properties).

### 3.2.5 Length of subject and PP

In order to determine the correlation between constituent order and subject length, PP length and relative length, we included each variable as a predictor in a univariate logistic regression model (cf. Hosmer and Lemeshow 2000). None of these models showed a significant correlation (subject length $\chi^2(1) < 1$, $p > .9$; PP length $\chi^2(1) = 1.94$, $p = .16$; relative length $\chi^2(1) = 0.87$, $p = .35$), suggesting that longer constituents in our corpus do not seem to be more likely to be placed postverbally. Figure 4 shows the frequency distribution over the two word orders by log-transformed subject length (upper plot) and PP length (lower plot).

## 3.3 Multifactorial analysis

Table 2 summarizes the individual effects of the nine factors investigated in the previous section. The factors Definiteness of the subject (both the distinction between bare and definite subjects and the more fine-grained classification) and Type of PP yielded a significant effect on subject position. Animacy of the subject, Tense, Aspect, Voice, and Length did not significantly affect subject position. However, in natural language all these factors are at work at the same time. To investigate the combined effects of animacy and definiteness simultaneously while controlling for the other factors, we performed a stepwise backward mixed logit regression analysis. This type of analysis extends ordinary logistic regression in
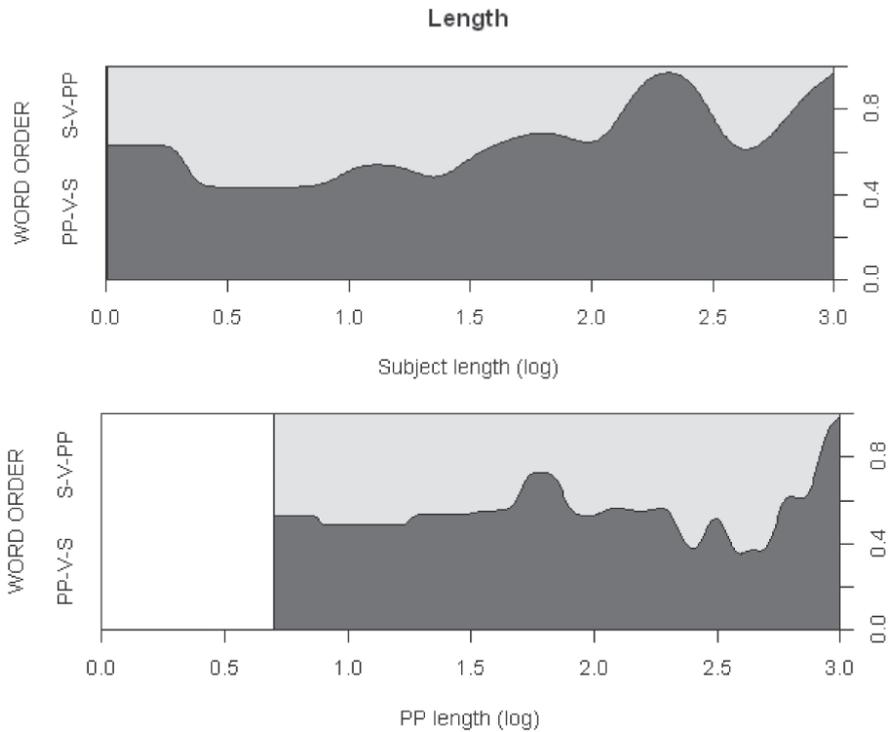
**Fig. 4:** Distribution over the two word order variants (S-V-PP = light grey area, PP-V-S = dark grey area) by log-transformed subject length (top) and PP length (bottom)

**Table 2:** Summary of the individual effects of the factors

| Factors | $\chi^2$ | df | p |
|---|---|---|---|
| Definiteness of the subject | | | |
|   bare vs. definite | 10.4 | 1 | <.01 |
|   four-level distinction | 11.9 | 3 | <.01 |
| Animacy of the subject | <1 | 1 | .585 |
| Type of PP | 19.4 | 2 | <.001 |
| Tense | <1 | 1 | .395 |
| Aspect | 2.2 | 1 | .138 |
| Voice | <1 | 1 | .796 |
| | *log-lik* $\chi^2$ | df | p |
| Length of the subject | 0 | 1 | >.9 |
| Length of the PP | 1.94 | 1 | .163 |
| Relative length | .87 | 1 | .35 |

that it allows for the inclusion of random factors that can control for dependencies across observations (e.g. multiple observations from the same speaker; Jaeger 2011). A mixed logit model allows us to determine the relative importance of our fixed factors and any possible interactions between them, while accounting for variance resulting from random differences between speakers and utterances. As in ordinary regression, the model gives an indication of how much variation (i.e. in word order that was found in the data) can be explained by these factors, and how much variation is not yet covered. In other words, it shows how well our selected factors can predict the position of the subject in the data set.

We built a mixed logit model using the lmer function in R (R Development Core Team 2008). All factors listed in Table 2 were included as fixed factors. Definiteness of the subject was included as an ordinal predictor, i.e., as a scale with fixed intervals, from 1 (demonstrative) to 4 (bare). This predictor was Helmert coded. That is, the highest level was compared to the three lower levels, the second level to the two lower levels, and the third level to the fourth level. The categorical predictors Animacy, Tense, Aspect, and Voice, and the log-transformed continuous predictors Length of the subject, Length of the PP and Relative length were centered, i.e. the mean was subtracted from these factors, causing their values to be centered around 0. Centering reduces collinearity between predictors, which is especially a concern in unbalanced designs such as the present study (e.g. Jaeger 2008). The three-level predictor PP type was contrast coded into two new predictors (e.g. Wendorf 2004). The recoded predictors represent the difference between abstract PPs and the grand mean, and the difference between temporal PPs and the grand mean, respectively.

We included two random intercepts in the model: one accounting for variation between conversations, and one accounting for variation between specific subject nouns. Since different conversations in the corpus are likely to involve different speakers, the factor Conversation is assumed to capture some between-speaker variance. For the second factor, we took the lemma of the head noun of each grammatical subject to account for idiosyncrasies in the position of specific subject nouns.

We started out with a model that included all fixed factors and their possible interactions, and successively removed predictors that did not significantly contribute to the overall fit of the model. The final model includes 4 main effects and one interaction effect. A summary of the model is given in Table 3. It has a model likelihood of –338.36, which is significantly better than a model with only random effects. $\chi^2(10, N = 541) = 66.69$; $p < .001$; Nagelkerke's $R^2 = 0.155$, indicating that adding the fixed effects is justified. The index of concordance suggests that the model has a good predictive ability ($C = 0.816$; Somers' $D_{xy} = 0.63$; e.g. Baayen 2008).

**Table 3:** Mixed logit model

| Random factors | $s^2$ | $s$ | | | |
|---|---|---|---|---|---|
| CONVERSATION | 0.000 | 0.000 | | | |
| SUBJECT NOUN | 0.465 | 0.682 | | | |

| Fixed factors | $\beta$ | SE | 95% C.I. | | $p$ |
|---|---|---|---|---|---|
| *Intercept* | −0.480 | 0.147 | −0.768 | −0.192 | <.01 |
| DEFINITENESS | | | | | |
|   dem vs. [def, modbare, bare] | 1.169 | 0.478 | 0.232 | 2.106 | <.05 |
|   def vs. [modbare, bare] | 1.601 | 0.332 | 0.951 | 2.251 | <.001 |
|   modbare vs. bare | 2.265 | 0.511 | 1.263 | 3.267 | <.001 |
| ANIMACY | | | | | |
|   animate vs. inanimate | 0.455 | 0.260 | −0.055 | 0.965 | .08 |
| PP TYPE | | | | | |
|   abstract vs. Grand Mean | 0.696 | 0.147 | 0.407 | 0.985 | <.001 |
|   temporal vs. Grand Mean | −0.851 | 0.189 | −1.220 | −0.481 | <.001 |
| SUBJECT LENGTH | −1.259 | 0.312 | −1.870 | −0.648 | <.001 |
| ANIMACY * DEFINITENESS | | | | | |
|   animate * dem vs. [def, modbare, bare] | 1.034 | 0.942 | −0.813 | 2.880 | .27 |
|   animate * def vs. [modbare, bare] | −1.266 | 0.548 | −2.341 | −0.192 | <.05 |
|   animate * modbare vs. bare | −0.322 | 0.822 | −1.933 | 1.290 | .70 |

In Table 3, positive values of the estimates correspond to an increase in the probability of a preverbal subject, while negative values correspond to a higher probability of a postverbal subject. For the factor Definiteness, the estimates show that the higher a subject is on the definiteness hierarchy, the more likely it is to occur in sentence-initial position. The main effect of Animacy of the subject, which was not significant in the univariate analysis, almost reaches significance in the regression model: Animate subjects tend to be more likely to be placed in sentence-initial position than inanimate subjects. However, this effect is qualified by a significant interaction with definiteness: the effect of animacy is significantly different for definite subjects than for (modified and unmodified) bare subjects, the latter being more likely to occur in the preverbal position when they are animate than when they are inanimate. Although being animate also increases the probability of a preverbal demonstrative subject, this turns out to be non-significant. The model furthermore shows that in sentences with abstract PPs,

subjects are significantly more likely to be placed in preverbal position than on average, while subjects in sentences containing temporal PPs are significantly more likely to occur postverbally. Finally, the model shows a significant effect of the length of the subject, where we did not find this before: the longer the subject, the more likely it is to be placed postverbally. As for the random factors, there is almost no variation in subject position across different conversations, probably due to the fact that there are only few conversations of which we have more than one observation. However, there is some variation between subject nouns. The partial effects of each predictor in our model are represented visually in Figure 5.

# 4 Discussion

We found variation in the position of plural subjects in Dutch main clauses that also contain a PP adjunct. About half of the subjects occurred in the preverbal position, the other half in postverbal position. This suggests that, as expected, the Subject First preference leaves room for other preferences. We have shown that part of the variation can be explained by properties of the subject, in line with findings from other languages with (relatively) free word order that inanimate and indefinite subjects are more likely to occur further down the sentence than animate and definite subjects (e.g. Branigan and Feleki 1999; Kempen and Harbusch 2004; Øvrelid 2004; Prat-Sala and Branigan 2000; Van Nice and Dietrich 2003). We specifically investigated bare plural subjects, which typically mark inaccessible entities due to their non-referentiality. Our finding that in Dutch bare plural subjects are more likely to be placed in postverbal position than definite plural subjects is in accordance with Vogels and Lamers (2008) and Bouma (2008), who found that definite subjects are more likely to be fronted than bare subjects. Extending these findings, our data provide support for a more fine-grained analysis of definiteness. There was a gradual increase in the probability of a subject to be in postverbal position as it was lower on the definiteness scale, from demonstrative to definite to modified bare to bare. This suggests that definiteness is not a discrete property: Bare and definite NPs can be considered part of a continuum (cf. Gundel et al. 1993), with the probability of a subject being placed in the preverbal position gradually increasing as it moves up along this scale.

More importantly, we have shown that there is a complex interplay between definiteness and animacy. These two properties can be conflicting. For example, when a subject is definite but inanimate or animate but a bare noun, the two properties are indecisive in determining whether the subject is accessible enough to be placed in the preverbal position. How this conflict is resolved depends on
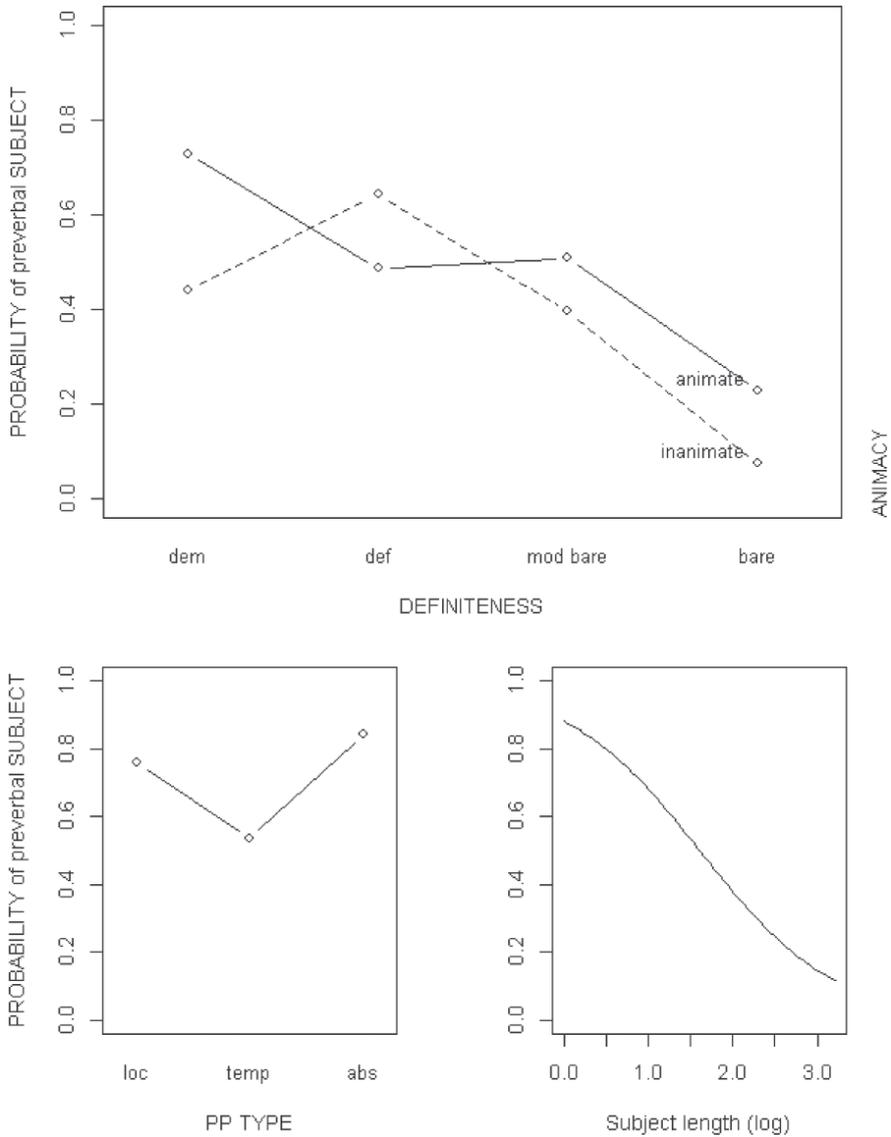
**Fig. 5:** Predicted probabilities for a preverbal subject; partial effects are adjusted for the reference categories of the other factors (anim, dem, loc) and for the median value of Subject length.

the relative importance of the topic properties. Our findings showed that bare subjects (modified or unmodified) were more likely to be produced postverbally when they were inanimate, whereas animate bare plural subjects were about as frequent in preverbal as in postverbal position. For definite plural subjects, being animate did not increase the probability of preverbal occurrences. Thus, animacy only has an effect on subject position when the subject is a bare plural. This suggests that definiteness is a stronger cue for subject position, in line with the idea that derived accessibility can overrule inherent accessibility (Prat-Sala and Branigan, 2000). In sum, despite the general preference in Dutch to start a sentence with the subject, inaccessible subjects are less likely to be placed preverbally. Subjects with a high derived accessibility (i.e. definite) are likely to be placed in the preverbal position, coinciding with the Subject First preference. Given the convergence of these two strong factors, a weaker factor such as the subject's inherent accessibility (animacy) will not affect its position. However, when subjects are not accessible from the discourse (as with bare plurals), the effect of inherent accessibility reveals itself. Thus, animacy only comes into play when definiteness is indecisive: when a subject is both inanimate and bare, these two properties may 'gang up' against the Subject First preference, which results in the balance tipping in favour of the postverbal position. Our results show that such an interplay between different factors is not only found in a controlled experimental environment, but also arise in normal natural language production (cf. Van Bergen 2011).

To our knowledge, the interaction between animacy and definiteness has not been found before in the position of the subject in Dutch main clauses. Whereas it has been observed that there is a preference in Dutch for animate and definite entities to be placed first in the sentence (e.g. Bouma, 2008; De Hoop and Lamers 2006), such an interaction has only been found in the Dutch middlefield (*Mittelfeld*), i.e. the postverbal domain. Van Bergen and De Swart (2010) found in a corpus study that the order of direct objects and adverbs in the Dutch middlefield (*scrambling*) was almost completely determined by the definiteness of the object: Pronominal objects almost categorically occurred before the adverb, whereas full noun phrases almost always followed the adverb. They furthermore found that proper nouns, which are neither high nor low on the definiteness scale, occurred about as often in scrambled as in unscrambled position. Part of this variation was explained by animacy, while this factor did not affect scrambling behaviour of other noun phrases. Our results show that a similar interaction between definiteness and animacy plays a role in the preverbal domain in Dutch.

Our results also provide evidence for the hypothesis that the effects of definiteness and animacy on subject position are absolute rather than relative to the properties of other arguments. Since in our data the preverbal position was either

occupied by the subject or by a PP adjunct, the fact that subjects were sometimes produced postverbally could not be due to the presence of another argument higher in definiteness or animacy claiming the preverbal position. For example, whereas studies on transitive sentences have suggested that direct objects are more likely to be fronted when they are higher in animacy than the subject, but not when subject and object have the same animacy (e.g. Van Bergen 2011; Lamers and De Hoop 2008; Van Nice and Dietrich 2003), the present study suggests that inanimate subjects are more likely to be placed in postverbal position, independently of the animacy of other arguments.

On the other hand, we did find that in sentences with PP adjuncts expressing a spatio-temporal location, the subject was more likely to occupy the postverbal position, while in sentences with abstract PPs, the subject was more likely to occur preverbally. One explanation is that PP adjuncts expressing a concrete location provide information about the upcoming subject (Grondelaers and Speelman 2007; Grondelaers et al. 2009). For example, in the context of a fishing event, a sentence starting with *On the hook was* . . . is likely to be continued with *a worm*. In other words, when a speaker begins his utterance with a locative adjunct, the hearer can start making *predictive inferences* about what will come next, based on semantic and conceptual information that the location evokes (cf. Altmann and Kamide 1999). It is less clear why temporal PP adjuncts were highly frequent in preverbal position, even more so than locative PPs. It is not likely that this high frequency is due to a function of enhancing predictions about upcoming subjects (Grondelaers et al. 2009). Temporal expressions do not project predictive inferences on upcoming constituents as much as locative expressions do, and they are less restrictive as to what kinds of subjects are compatible with it than spatial locations (Grondelaers and Speelman 2007). The reason for the preference of temporal PP adjuncts for the preverbal position should thus be sought elsewhere. Perhaps it is related to the fact that temporal adjuncts can refer to the reference time of an utterance, while locatives only have scope over an event (Frey 2003). Another explanation might be found in the structure of the corpus: About 6% of the CGN consists of news reports broadcasted on radio and TV. These types of texts may be more likely to contain temporal information (when did the reported events happen) in prominent syntactic positions. By selecting only sentences that contained PP adjuncts, we might have extracted a comparatively large part of news reports, which could have influenced the number of sentence-initial temporal PPs. Indeed, 17% of our data originated from news reports. However, in this subset of the data we did not find a clear preference for temporal PPs to occur in sentence-initial position.

Given the effect of PP Type on word order, one could argue that inaccessible subjects are only preferred in postverbal position when the PP adjunct is more

accessible than the subject. However, the fact that we did not find an interaction between the factors PP Type and Definiteness or Animacy of the subject in the regression model indicates that the high frequency of postverbal bare and inanimate subjects arose independently of the accessibility of the PP adjunct. In light of this, it would be interesting to look at existential constructions, in which the preverbal position is occupied by expletive *er* 'there', and hence such sentences are essentially topic-less. In Dutch, subjects of existential sentences are almost exclusively non-specific (Haeseryn et al. 1997). We would predict that they would also be more likely to be inanimate (cf. Beaver et al. 2005). We leave this for future research.

　　The results of the present study at least support a moderately incremental view on sentence production (e.g. Ferreira and Swets 2002). Whereas a radically incremental view would predict that constituents are produced as soon as they become available, our results show that inaccessible subjects are sometimes placed in the preverbal position in Dutch, suggesting that subjects prefer the preverbal position independently of accessibility. Still, subjects are more likely to be postponed when they have a low accessibility, suggesting that the order in which entities become available for processing at least partly affects word order in Dutch main clauses (cf. Kempen and Hoenkamp 1987; Levelt 1989; see also Van Bergen 2011). Our findings are also compatible with a probabilistic account of syntactic variation, in which multiple constraints simultaneously affect the likelihood that a constituent is placed in a certain position (e.g. Bates and MacWhinney 1989; Bresnan and Hay 2008; Bresnan et al. 2007; Gries 2003). In this paper, we have focused on the interplay between definiteness and animacy. Although we controlled for a number of potentially relevant factors in this study (e.g. length, properties of the predicate), it is clear that more factors may influence subject placement in Dutch, given the relatively low overall model quality (i.e., the amount of explained variance in the data as indicated by the concordance index $C$ and Somers' $D_{xy}$). When we look at those sentences in the corpus in which the subject is highly accessible (both definite and animate) and short (2 words), and in which the PP adjunct is abstract, still 41% has the subject in postverbal position. This percentage needs to be accounted for. It is likely that a large role is reserved for contextual factors. For example, although we have taken definiteness as a proxy for derived accessibility, this is only a rough estimation, since not all definite NPs convey given information (e.g. Epstein 2002). Thus, we might be missing important contextual factors, although these are hard to code for in a corpus of natural language. In addition, for a more thorough investigation of the influence of the predicate, more properties of the predicate should be taken into account, such as verb semantics, genericity and thematic role selection. Given effects of definiteness and animacy reported in earlier studies on other types of

word order variation, our findings add to the evidence that there is a complex interplay between different sources of accessibility in determining subject position in Dutch.

# 5 Conclusion

In this article, we have presented a corpus study of the interplay between definiteness and animacy in the placement of subjects in Dutch main clauses. Our results support the hypothesis that an interaction between definiteness and animacy affects the probability of a preverbal subject. Definite subjects mark referents that are accessible from the discourse, as a result of which the subject is likely to occur in preverbal position. Bare subjects are typically inaccessible, which causes an incongruence with the Subject First preference, resulting in word order variation. When the subject is also inanimate, the two properties 'gang up' against the Subject First preference. Our multifactorial analysis supports the view that subject position is determined by multiple possibly conflicting forces in a probabilistic manner. Although more factors than we have taken into account may have an impact on subject position, the present study adds to the existing evidence that also in natural language there is a complex interplay between multiple sources of accessibility that determine word order variation.

## Acknowledgements

## References

Aissen, Judith. 2003. Differential object marking: Iconicity vs. economy. *Natural Language and Linguistic Theory* 21(3). 435–483.
Altmann, Gerry T. M. and Yuki Kamide. 1999. Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition* 73(3). 247–264.
Ariel, Mira. 1990. *Accessing noun-phrase antecedents*. London: Routledge.
Baayen, R. Harald. 2008. *Analyzing linguistic data. A practical introduction to statistics using R*. Cambridge: University Press.

Bates, Elizabeth and Brian MacWhinney. 1989. Functionalism and the Competition Model. In Elizabeth Bates and Brian MacWhinney (eds.), *The crosslinguistic study of sentence processing*, 3–73. Cambridge, UK: Cambridge University Press.

Beaver, David, Itamar Francez and Dmitry Levinson. 2005. Bad subjects: (Non-)canonicality and NP distribution in existentials. In Effi Georgala and Jonathan Howell (eds.), *Proceedings of Semantics and Linguistic Theory XV*, 19–43. Ithaca, NY: CLC Publications.

van Bergen, Geertje. 2011. *Who's first and what's next. Animacy and word order variation in Dutch language production*. Nijmegen: Radboud University Nijmegen dissertation.

van Bergen, Geertje and Peter de Swart. 2010. Scrambling in spoken Dutch: Definiteness versus weight as determinants of word order variation. *Corpus Linguistics and Linguistic Theory* 6(2). 267–295.

Bock, J. Kathryn and David E. Irwin. 1980. Syntactic effects of information availability in sentence production. *Journal of Verbal Learning and Verbal Behavior* 19(4). 467–484.

Bock, J. Kathryn and Richard K. Warren. 1985. Conceptual accessibility and syntactic structure in sentence formulation. *Cognition* 21(1). 47–67.

Bosque, Ignacio. 2001. Adjective position and the interpretation of indefinites. In Javier Gutiérrez-Rexach and Luis Silva-Villar (eds.), *Current Issues in Spanish Syntax and Semantics*, 17–37. Berlin: Mouton de Gruyter.

Bouma, Gerlof Johannes. 2008. *Starting a sentence in Dutch: a corpus study of subject- and object-fronting*. Groningen: Rijksuniversiteit Groningen dissertation.

Branigan, Holly P. and Eleonora Feleki. 1999. Conceptual accessibility and serial order in Greek speech production. In Martin Hahn and Scott C. Stoness (eds.), *Proceedings of the Twenty First Annual Conference of the Cognitive Science Society*, 96–101. Mahwah, NJ: Lawrence Erlbaum.

Branigan, Holly P., Martin J. Pickering and Mikihiro Tanaka. 2008. Contributions of animacy to grammatical function assignment and word order during production. *Lingua* 118(2). 172–189.

Bresnan, Joan, Anna Cueni, Tatiana Nikitina and R. Harald Baayen. 2007. Predicting the dative alternation. In Gerlof Bouma, Irene Krämer and Joost Zwarts (eds.), *Cognitive Foundations of Interpretation*, 69–94. Amsterdam: Royal Netherlands Academy of Science.

Bresnan, Joan and Jennifer Hay. 2008. Gradient grammar: An effect of animacy on the syntax of *give* in New Zealand and American English. *Lingua* 118(2). 245–259.

Brunetti, Lisa. 2009. On the semantic and contextual factors that determine topic selection in Italian and Spanish. *The Linguistic Review* 26(2/3). 261–289.

Carlson, Gregory N. 1977. A unified analysis of the English bare plural. *Linguistics and philosophy* 1. 413–456.

Carlson, Gregory N. 1982. Generic terms and generic sentences. *Journal of Philosophical Logic* 11. 145–181.

Carreiras, Manuel, Núria Carriedo, María Angeles Alonso and Angel Fernández. 1997. The role of verb tense and verb aspect in the foregrounding of information during reading. *Memory and Cognition* 25(4). 438–446.

Christianson, Kiel and Fernanda Ferreira. 2005. Conceptual accessibility and sentence production in a free word order language (Odawa). *Cognition* 98(2). 105–135.

Chierchia, Gennaro. 1998. Reference to kinds across language. *Natural Language Semantics* 6(4). 339–405.

Choi, Hye-Won. 1999. *Optimizing structure in context: Scrambling and information structure*. CSLI Publications dissertation.

Clark, Herbert H., Robert Schreuder and Samuel Buttrick, S. 1983. Common ground and the understanding of demonstrative reference. *Journal of Verbal Learning and Verbal Behavior* 22(2). 245–258.

Comrie, Bernard. 1989. *Language universals and linguistic typology* (2nd edition). Chicago: University of Chicago Press.

Crasborn, Onno, Els van der Kooij, Johan Ros and Helen de Hoop. 2009. Topic agreement in NGT (Sign Language of the Netherlands). *The Linguistic Review* 26(2/3). 355–370.

Croft, William. 1988. Agreement vs. case marking and direct objects. In Michael Barlow and Charles A. Ferguson (eds.), *Agreement in Natural Language*, 159–179. Stanford: CSLI.

Dahl, Östen and Kari Fraurud. 1996. Animacy in grammar and discourse. In Thorstein Fretheim and Jeanette Gundel (eds.), *Reference and Referent Accessibility*, 47–64. Amsterdam: John Benjamins.

Davison, Alice. 1984. Syntactic markedness and the definition of sentence topic. *Language* 60(4). 797–846.

É. Kiss, Katalin. 2002. *The syntax of Hungarian*. Cambridge: Cambridge University Press.

Epstein, Richard. 2002. The definite article, accessibility and the construction of discourse referents. *Cognitive Linguistics* 12(4). 333–378.

Ferreira, Fernanda and Benjamin Swets. 2002. How incremental is language production? Evidence from the production of utterances requiring the computation of arithmetic sums. *Journal of Memory and Language* 46(1). 57–84.

Ferreira, Victor S. and Hiromi Yoshita. 2003. Given-new ordering effects on the production of scrambled sentences in Japanese. *Journal of Psycholinguistic Research* 32(6). 669–692.

Frey, Werner. 2003. Syntactic conditions on adjunct classes. In Ewald Lang, Claudia Maienborn and Cathrine Fabricius-Hansen (eds.), *Modifying adjuncts*, 163–209. Berlin: Mouton de Gruyter.

Givón, Talmy. 1976. Topic, pronoun, and grammatical agreement. In Charles N. Li (ed.), *Subject and Topic*, 149–188. New York/San Francisco/London: Academic Press.

Givón, Talmy. 1979. *On understanding grammar. Perspectives in neurolinguistics and psycholinguistics*. New York/San Francisco/London: Academic Press.

Gries, Stefan Th. 2003. *Multifactorial analysis in corpus linguistics: A study of particle placement*. London/New York: Continuum.

Grondelaers, Stefan and Dirk Speelman. 2007. A variationist account of constituent ordering in presentative sentences in Belgian Dutch. *Corpus Linguistics and Linguistic Theory* 3(2). 161–193.

Grondelaers, Stefan, Dirk Speelman, Denis Drieghe, Marc Brysbaert and Dirk Geeraerts. 2009. Introducing a new entity into discourse: Comprehension and production evidence for the status of Dutch *er* "there" as a higher-level expectancy monitor. *Acta Psychologica* 130(2). 153–160.

Grosz, Barbara J., Aravind K. Joshi and Scott Weinstein. 1995. Centering: A framework for modelling the local coherence of discourse. *Computational Linguistics* 21(2). 203–225.

Gundel, Jeanette K. 1988. Universals of topic-comment structure. In Michael Hammond, Edith A. Moravcsik and Jessica R. Wirth (eds.), *Studies in Syntactic Typology*, 209–239. Amsterdam: John Benjamins.

Gundel, Jeanette K., Nancy Hedberg and Ron Zacharski. 1993. Cognitive status and the form of anaphoric expressions in discourse. *Language* 69(2). 274–307.

Haeseryn, Walter, K. Romijn, G. Geerts, J. de Rooij and M.C. van den Toorn. 1997. *Algemene Nederlandse Spraakkunst*. Groningen: Martinus Nijhoff.

Hendriks, Petra, Helen de Hoop and Monique J.A. Lamers. 2005. Asymmetries in language use reveal asymmetries in the grammar. In P. Dekker and M. Franke (eds.), *Proceedings of the 15th Amsterdam Colloquium*, 113–118. Amsterdam: ILLC, University of Amsterdam.

de Hoop, Helen. 1992. *Case Configuration and NP Interpretation*. Groningen: University of Groningen dissertation.

de Hoop, Helen and Irene Krämer. 2006. Children's optimal interpretations of indefinite subjects and objects. *Language Acquisition* 13(2). 103–123.

de Hoop, Helen and Monique J.A. Lamers. 2006. Incremental distinguishability of subject and object. In Leonid Kulikov, Andrej Malchukov and Peter de Swart (eds.), *Case, Valency and Transitivity*, 269–287. Amsterdam/Philadelphia: John Benjamins.

Hosmer, D.W. and S. Lemeshow. 2000. Applied logistic regression (2nd edition). New York: Wiley.

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.

Jaeger, T. Florian. 2011. Corpus-based research on language production: Information density and reducible subject relatives. In Emily M. Bender and Jennifer E. Arnold (eds.), *Language from a cognitive perspective: Grammar, usage, and processing*, 161–197. Stanford: CSLI.

Kempen, Gerard and Edward Hoenkamp. 1987. An incremental procedural grammar for sentence formulation. *Cognitive Science* 11(2). 201–258.

Kempen, Gerard and Karin Harbusch. 2004. A corpus study into word order variation in German subordinate clauses: Animacy affects linearization independently of grammatical function assignment. In Thomas Pechmann and Christoph Habel (eds.), *Multidisciplinary approaches to language production*, 173–181. Berlin: Mouton De Gruyter.

König, Esther, Wolfgang Lezius and Holger Voormann. 2003. *TIGERSearch 2.1 User's Manual*. IMS, University of Stuttgart.

Krifka, Manfred, et al. 1995. Genericity: An introduction. In Gregory N. Carlson and Francis Jeffry Pelletier (eds.), *The Generic Book*, 1–124. Chicago: The University of Chicago Press.

Lambrecht, Knud. 1994. *Information structure and sentence form. Topic, focus, and the mental representations of discourse referents*. Cambridge: Cambridge University Press.

Lamers, Monique J.A. 2007. Verb type, animacy and definiteness in grammatical function disambiguation. In Bettelou Los and Marjo van Koppen (eds.), *Linguistics in the Netherlands 2007*, 125–137. Amsterdam: John Benjamins.

Lamers, Monique J.A. and Helen de Hoop. 2008. Object fronting in Dutch. Unpublished manuscript, Radboud University Nijmegen.

Levelt, Willem J.M. 1989. *Speaking: From intention to articulation*. Cambridge, MA: the MIT Press.

Longobardi, Giuseppe. 2001. How comparative is semantics? A unified parametric theory of bare nouns and proper names. *Natural Language Semantics* 9(4). 335–369.

McDonald, Janet, Kathryn Bock and Michael H. Kelly. 1993. Word and world order: Semantic, phonological and metrical determinants of serial position. *Cognitive Psychology* 25(2). 188–230.

van Nice, Kathy Y. and Rainer Dietrich. 2003. Animacy effects in language production: From mental model to formulator. In Holden Härtl and Heike Tappe (eds.), *Mediating between Concepts and Grammar*, 101–117. Berlin/New York: Mouton de Gruyter.

van Oosten, Jeanne. 1986. *The nature of subjects, topics and agents: A cognitive explanation*. Indiana University Linguistics Club dissertation.

Øvrelid, Lilja. 2004. Disambiguation of grammatical functions in Norwegian: Modeling variation in word order interpretations conditioned by animacy and definiteness. In F. Karlson (ed.),

proceedings of the 20th Scandinavian Conference of Linguistics. University of Helsinki: Department of General Linguistics.

Prat-Sala, Mercè and Holly P. Branigan. 2000. Discourse constraints on syntactic processing in language production: a cross-linguistic study in English and Spanish. *Journal of Memory and Language* 42(2). 168–182.

R Development Core Team. 2008. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL: http://www.R-project.org.

Reinhart, Tanya. 1982. Pragmatics and linguistics: An analysis of sentence topics. *Philosophica* 27(1). 53–94.

Rosenbach, Anette. 2005. Animacy versus weight as determinants of grammatical variation in English. *Language* 81(3). 613–644.

van Tiel, Bob and Monique J.A. Lamers. 2007. Animacy in verschillende teksttypes. *Tabu: Bulletin voor Taalwetenschap* 36. 19–38.

Vogels, Jorrig and Monique J. A. Lamers. 2008. The placement of bare plural subjects in Dutch. In Marjo van Koppen and Bert Botma (eds.), *Linguistics in the Netherlands* 2008, 169–80. Amsterdam: John Benjamins.

Vonk, Wietske, Lettica G.M.M. Hustinx and Wim H.G. Simons. 1992. The use of referential expressions in structuring discourse. *Language and Cognitive Processes* 7(3/4). 301–333.

Wasow, Thomas. 1997. Remarks on grammatical weight. *Language Variation and Change* 9(1). 81–105.

Wendorf, Craig A. 2004. Primer on multiple regression coding: Common forms and the additional case of repeated contrasts. *Understanding Statistics* 3(1). 47–57.

# Bionotes

*Jorrig Vogels* is a Ph.D. student at Tilburg center for Cognition and Communication (TiCC), one of the research centres of the Tilburg School of Humanities at Tilburg University, the Netherlands. The research presented in this article is an extension of the author's MA thesis project at Radboud University Nijmegen. The author is currently working on a project on the production of referring expressions, and is interested in psycholinguistic approaches to both word order choice and choice of referential forms.

*Geertje Van Bergen* received her Ph.D. in 2011 from the Radboud University Nijmegen, with a dissertation on the role of animacy on different types of word order variation in Dutch language production. From February to September 2011, she worked as a postdoctoral researcher at the University of Louvain-la-Neuve (Belgium), where she investigated the meaning and use of discourse markers in Dutch. Since October 2011, she holds a postdoctoral position at the Centre for Language Studies at Radboud University Nijmegen, in the NWO-project 'Varieties of Initial Learners in Language Acquisition'. In this project she investigates the gradual development of an L2 lexicon from the very first exposure.