Gathering Corpus Evidence of Word Order Freezing in Dutch

Gerlof Bouma

Rijksuniversiteit Groningen / Universität Potsdam

g.j.bouma@rug.nl

1 Introduction

In languages that allow the order or position of argument constituents to vary, it is possible that one cannot reliably assign grammatical function (or thematic role) to a constituent purely on the basis of its position in a sentence. Thus, in those languages, grammatical function assignment is not necessarily recoverable from word order alone. For several free word order languages it has been claimed that word order freedom is contingent upon the word order independent recoverability of grammatical function assignment. The resort to strict canonical word order that some free word order languages show when there is not enough word order independent information to correctly interpret argument constituents is known as *word order freezing*. Typical examples of word order independent information sources are case and verbal agreement. However, information that is not expressed formally, such as animacy, may also help to assign the correct grammatical function to a constituent. In this paper, we will argue that Dutch shows word order freezing as a tendency, and support this claim by providing novel corpus evidence of freezing in spoken Dutch.

2 Freezing in Dutch

Dutch can be described as a verb-second, verb-final language: In a main clause, the finite verb occupies second position, and any other, non-finite verbs cluster towards the end of the sentence. Apart from these fixed verbal positions, Dutch allows for a moderate amount of word order variation. For instance, the first, directly preverbal position ⊟known as the *Vorfeld* ⊟can be occupied by many different kinds of material. In (1a), the Vorfeld is occupied by the subject, whereas in (1b) the Vorfeld is occupied by the direct object. This illustrates only two of the many possibilities.

(1)	a.	**Ik**	begrijp dat	niet.	
		I	understand	that	not

b. **Dat** begrijp ik niet.
That understand I not
`I don't understand that.'

In terms of corpus frequency, the subject is the default Vorfeld occupant (about 70% of the subjects is in the Vorfeld). In addition, Vorfeld subjects are not information structurally restricted. Subjects may thus be considered the unmarked Vorfeld occupant. The pair in (1) shows that subject and object can occupy the same pre- and post-verbal positions in Dutch and that the object may precede the subject in a sentence. As a result, word order is not necessarily a reliable source of information for a hearer or parser, who has to assign grammatical function to constituents. In both (1a) and (1b), subjecthood of *ik* `I' can be established on the grounds of its form (*ik* is in subject form). Also, the demonstrative *dat* `that' is very likely to have an abstract referent, which makes it a good object for the verb *begrijp* `understand'. That is, there is formally encoded as well as not formally encoded word order independent information in (1) that allows grammatical function assignment to be recovered.

What happens in Dutch when there is no word order independent information present in a sentence to determine grammatical function assignment? One might expect that such sentences are ambiguous between subject-before-object and object-before-subject, since we have seen in (1) that both orders are possible. However, as (2) shows, this ambiguity is not clearly observed. The object-initial interpretation is dispreferred to the point of unavailability.

(2) Ella groet Gerald.
Ella greets Gerald
`Ella greets Gerald.'
Not*, or at least strongly dispreferred:* `Gerald greets Ella.'

The object-initial interpretation of (2) becomes available when the Vorfeld occupant receives focus accent and the rest of the sentence is deaccented, or when the sentence is placed in a context that triggers the expectation that Gerald is the subject. The subject-initial interpretation needs no such help, however. Therefore, we may say that Dutch shows word order freezing: With respect to the Vorfeld, word order freezes to subject-before-object when there is not enough information present to recover non-canonical word order. On the basis of intuition data, Dutch can be grouped amongst languages like Hindi, Russian and Japanese, as a free word order language that shows word order freezing, although the freezing effect in Dutch seems to be a tendency rather than an absolute effect.

3 Corpus Investigation

The claim that word order freedom in a language is contingent upon the presence of word order independent information about the grammatical function assignment is a claim about production. In order to find empirical validation for this intuition-databased claim, we have investigated word order freezing in the Spoken Dutch Corpus (CGN), which contains approximately 1 mln words of syntactic annotation for spoken Dutch. The syntactic annotation includes annotation of dependency structure, which means that grammatical function information can be directly obtained from the CGN. Since case is only marked in parts of the pronominal paradigm in Dutch, we have focussed on information that is not expressed formally in our corpus investigation of freezing. In particular, we investigate recoverability on the basis of relative definiteness and relative animacy (explained below). The tested *freezing hypothesis* is that object-initial realization of transitive sentences is more common when grammatical function can be recovered using relative definiteness or relative animacy.

Relative Definiteness We can look at definiteness as a scale: pronoun - definite full NP - indefinite full NP. In general it has been observed that subjects have the tendency to be on the high end of this scale, whereas (direct) objects tend to appear at the bottom end. In the CGN, too, subjects are associated with high definiteness, and objects with low definiteness. When the subject is higher on the definiteness scale than the object, the correct grammatical function assignment can be recovered by means of relative definiteness: The unmarked interpretation− the most definite NP is the subject− yields the correct grammatical function assignment. Under the freezing hypothesis, object-initial realization of transitive sentences should be more frequent when the subject is higher on the definiteness scale than the object. Logistic regression modelling of word order in 16146 transitive clauses (object-initial vs canonical realization) confirms this. The model predicts object-fronting from definiteness (NP form) and length of the subject and object, which are factors in Dutch word order that have been established in earlier research. The additional, three-valued factor relative definiteness indicates were the subject is on the definiteness scale relative to the object. The model shows that when the subject is higher on the definiteness scale, and thus when grammatical function assignment is recoverable through relative definiteness, the odds of object-initial realization increase by at least 50%. There is no evidence for an extra decrease in object-initial realization when the subject is less definite than the object.

Relative Animacy Subjecthood is also associated with animacy, and likewise objecthood with inanimacy. Under the freezing hypothesis, object-initial realization should be more frequent when the subject is animate and the object inanimate, since in that case relative animacy allows one to recover grammatical function assignment. Investigation of a random selection of 2345 transitive sentences from the CGN that was manually annotated for animacy of the subject and object NPs provides preliminary confirmation of this prediction. The proportion of object-initial sentences rises when the subject is animate and the object inanimate (from 9% to 20%). Logistic regression modelling shows that this effect can not be fully attributed to other factors in word order, like object definiteness, the positive effect of relative definiteness, or the independent contributions of subject and object animacy. These results should be considered as preliminary, however, since the dataset is too small to support a logistic regression that includes all of the (highly correlated) factors that have been previously found relevant in predicting object fronting.

4 Conclusion

Word order freezing refers to the resort to canonical word order in an otherwise free word order language that occurs when there is not enough word order independent information to recover the correct grammatical function assignment. We have presented corpus evidence for the existence of word order freezing as a trend in spoken Dutch. First, non-canonical (that is, object-initial) word order is more frequent when grammatical function of the NPs in a sentence can be recovered on the basis of relative definiteness of the NPs. Secondly, there is preliminary evidence that non-canonical word order is more frequent when grammatical function is recoverable on the basis of relative animacy.

The fact that recoverability has an influence on word order frequencies in a corpus suggests that speakers are sensitive to the chances of communicative success. After all, recoverability refers to the question of whether information is sufficiently encoded in an utterance to be correctly understood when the utterance is used. Speakers of Dutch rely more on canonical word order when there is an increased chance that grammatical function assignment will be misconstrued by a hearer, and thus when there is an increased chance that the message will be misunderstood.

More work is needed to confirm the preliminary findings with respect to relative animacy. It would also be interesting to repeat the corpus experiments on other languages. For this purpose, German would be an interesting candidate, as it is very similar to Dutch, but with more general case marking and more word order freedom. Investigating German would allow us to compare the relevance of not formally encoded grammatical function information, such as relative definiteness and animacy, with the influence of distinguishing case marking on word order variation.