Om-omission in Dutch verbal complements

Gosse Bouma
Information Science
University of Groningen
g.bouma@rug.nl

August 30, 2013

Abstract

The complementizer *om*, which heads *to*-infinitival clauses in Dutch, is optional if the clause it introduces is a complement. We investigate which linguistic features influence the distribution of *om* in this construction, using data collected from an automatically parsed corpus of Dutch. A large part of the variation in the distribution of *om* is accounted for by the governing verb. In addition to this, features that account for syntactic or processing complexity play a significant role as well as features that characterize typical complement clauses for a given governor and typical purpose or goal modifier clauses. The fact that lexical variation plays a dominant role in our model motivates our choice for a mixed effects logistic regression model, where verbs are used as random effects.

1 Introduction

Dutch *to*-infinitival complement clauses (ICs) can be optionally introduced by the complementizer *om*. We find such ICs as dependents of verbs, nouns, adjectives, and prepositions (the element that selects the complement clause is indicated in bold):

(1) a. De Indiërs **aarzelen** (om) te investeren in Uganda
The Indians hesitate (COMP) to invest in Uganda
The Indians hesitate to invest in Uganda

b. Ik ben niet **vrij** (om) daarover te spreken
I am not free (COMP) about-that to speak
I am not free to speak about that

c. Huurders krijgen het **recht** (om) mee te praten
tenants obtain the right (COMP) with to talk
Tenants obtain the right to have a say

d. Ik hou er niet **van** (om) Beverly Hills af te kammen
I like there not of (COMP) Bevery Hills PRT to disrespect
I do not like to criticize Berverly Hills

It seems highly unlikely that the presence or absence of *om* in examples like these in actual language use is totally random. For one thing, the governor (i.e. *aarzelen* in (1-a)) has a very strong effect on the probability that the IC is introduced by *om*.

Lexical properties of the governor can only express the tendency for *om* to be present given a certain *governor*. There might be other properties that play a role in the choice for *om* in a given sentence. In particular, one might argue that syntactic (processing) complexity and semantic properties play a role.

Processing complexity, for instance, can be reduced by eliminating (local) ambiguity. The complementizer *om* explicitly marks the start of an IC. Therefore, one potential reason to use*om* is to disambiguate situations where the start of the IC is unclear. In (2) for instance, word order makes it unclear whether the adverb *regelmatig* is part of the matrix clause (2-b) or the IC (2-c).

(2) a. De inspectie verzoekt regelmatig alle objecten te inspecteren
The inspection asks regularly all objects to inspect

b. *The inspection regularly asks to inspect all objects*
c. *The inspection asks to inspect all objects regularly*

More in general, we might expect *om* to be used more often in sentences that are ’complex’ in one way or another. Long sentences containing material that could be part of either the matrix clause or the IC, with many words intervening between the verbal governor and the vebal head of the IC, might contain *om* more often than ’simple’, short, sentences.

An alternative, semantic, explanation might point to the fact that in (purpose or goal) modifier clauses, *om* is obligatory:

(3) Omstanders duwden hem in een vijver om af te koelen
Bystanders pushed him in a pond COMP PRT to cool
Bystanders pushed him into a pond to cool off

Historically, the use of *om* as a complementizer in modifier clauses precedes that of its use as complement marker. If this historical origin is still reflected in the current use of *om*, one expects *om* to be present especially in those ICs that bear some resemblance to purpose and goal modifier clauses. We investigate the role of two features that might be used to distinguish between typical complements of a verb and typical modifier clauses.

The existing literature on *om* as complementizer does not address the issue what controls the choice for *om* in individual sentences. IJbema (2002) describes the historical development of *om* from being used exclusively as preposition to an element that is also used as complementizer in purpose modifier clauses, and, more recently, in complement clauses.

Jansen (1987) discusses the fact that prescriptive grammars until recently disapproved of the use of *om* in complement clauses, and also provides some corpus evidence for the fact that *om* is used more often in spoken (informal) language, suggesting that register and genre might play a role. However, the results of Jansen (1987) are based on a very small data-set, and only contrasts spoken and written language. There has not been any investigation into the distribution of *om* based on a large data-set that describe the features that influence the presence or absence of *om* in individual sentences.

This is in strong contrast with similar constructions in English, i.e. the optional presence of *that* in finite complement clauses (4) and relatives (5), which has been the subject of numerous studies (see, among others, Ferreira and Dell (2000), Hawkins (2002), and Wasow, Jaeger, and Orr (2011)).

(4) a. The athlete realized (that) her goals would be difficult to achieve
b. Will explained (that) Trump was more important to the process than Huntsman

(5) a. That is certainly one reason (why/that) crime has increased
b. I think that the last movie (which/that) I saw was *Mysery*
c. They have all the water (that) they want

In particular, Roland, Elman, and Ferreira (2006) observe that the strongest predictor for complementizer presence is the governing verb. Jaeger (2010) extends this result by showing that this effect can to a large extent be contributed to subcategorization frequency, in particular, the likelihood that a governing verb occurs with a complement clause.

In this paper, we collect statistics on *om* presence in ICs from a large, automatically parsed, corpus of modern written Dutch and a smaller corpus of spoken Dutch. We restrict ourselves to ICs where *om* is truly optional, i.e.both the utterance with and without *om* are considered to be grammatical.

In particular, we will show that:

1. Nominal governors have a stronger preference for the presence of *om* than adjectival governors, which in turn have stronger preference for *om* than verbal governors.

2. Preference for *om* is stronger in spoken language than in written text.

3. Apart from lexical preferences, both syntactic and semantic properties play a significant role in predicting the presence of *om* in individual sentences.

Below, we first give a brief overview of the use of *om* as complementizer in Dutch. Next, we discuss why language users might add *om* in those cases where it is optional. Next, we describe the corpora we used and our method for collecting data.

In section 5 we compare the distribution of *om* for verbal, adjectival, and nominal governors, and in written vs. spoken data. In section 7 we discuss various features that we used to model syntactic complexity and the semantic distinction between typical complements and typical (purpose or goal) modifiers. Next, we present logistic regression models that incorporate various combinations of these features. The verbal governor of an IC has a strong influence on the distribution of *om.* Using the verbal governor as a random effect, we show that various features that account for syntactic complexity also play a statistically significant role. Semantic features modelling the distinction between typical complements and modifiers play a role as well, and a combination of both features leads to the best model.

2 The Use of Om as Optional Complementizer

In modern Dutch, we can distinguish configurations where the use of *om* is obligatory, cases where it is optional, and cases where *om* cannot be used. The complementizer *om* is obligatory if the infinitival clause functions as a modifier (typically expressing a goal or purpose) (6-a) or is in predicative position (6-b):[1]

(6) a. Het slachtoffer was naar Groningen gekomen om er de jaarwisseling te vieren
The victim was to Groningen come COMP there the newyear to celebrate
The victim had come to Groningen to celebrate new year
b. dat was niet om aan te zien
that was not COMP at to see
that was horrible to look at

As illustrated in (1), the complementizer *om* generally is optional for verbs, nouns, adjectives, and prepositions that take a *to*-infinitive as complement. *To*-infinitival clauses can also be subjects, in which case *om* can be optionally present as well:

(7) Het is **onmogelijk** (om) te wachten tot alle afschriften binnen zijn
It is impossible to wait COMP until all receipts in are
It is impossible to wait until all receipts are in

Subject complement sentences are usually extraposed, with the expletive pronoun *het* taking the position of the subject, but *te* and *om te* infinitives can also, although rarely, occur in sentence initial position:

(8) a. Om mijn show als parodie te omschrijven gaat te **ver**
COMP my show as parody to describe goes to far
To characterize my show as parody is an exaggeration
b. Feynman te lezen over fysische onderwerpen is even **boeiend als begrijpelijk**
Feynman to read about physics topics is equally capturing as understandable
To read Feynman on physics topics is equally capturing as understandable

The complementizer *om* cannot be used if the governing verb is a so-called *crossing dependency* (or *verb raising*) verb:

(9) Het was voor het eerst dat de thuisclub (*om) de ontmoeting wist te winnen
It was for the first that the homeclub COMPL the match knew to win
It was for the first time that the homeclub managed to win the match

In (9), the object of *te winnen*, *de ontmoeting*, precedes the governing verb, *wist*, while *te winnen* follows it. This pattern characterizes crossing dependency verbs. Governors that require such crossing dependency word orders never allow *om.*

IJbema (2002) points out that there also are semantic restrictions on verbal governors that allow *om*: "om can optionally appear in infinitival complement clauses with irrealis modality such as (10-a). Om is excluded in propositional (10-b), factive (10-c), and implicative infinitival complements (10-d)[2]:"

[1] See Geerts et al. (1984), section 19.3.3., on reduced subordinate clauses, for a more extensive overview.

[2] In our corpus, we found one counterexample to the claim that *beginnen* does not take *om*: *Ze zijn begonnen om de 50 cc-klasse af te schaffen* (*they have started to eliminate the 50 CC class*) (AD19940623-0108-6-2).

(10) a. Jan belooft/ besluit/ dwingt mij/ raad mij aan/ weigert (om) een boek te lezen
John promises/ decides/ forces me/ advises me PART/ refuses for a book to read
John promises/decides/forces me/tries/advises me/refuses to read a book

b. Jan beweert/ zegt (*om) erg intelligent te zijn
John claims/ says for very intelligent to be
John claims/says that he is very intelligent

c. Jan beseft/ realiseert zich (*om) erg intelligent te zijn
John realizes/ realizes himself for very intelligent to be
John realizes that he is very intelligent

d. Jan begint (*om) een boek te lezen
John begins for a book to read
John begins to read a book

In our data we noticed the near synonym verbs *bestaan* and *presteren* (*to manage to*), and the verb *ophouden* (*to stop*) that are factive and do not describe a future event but still allow *om*:

(11) a. Parijs **bestaat** het om aan te kondigen dat van Mururoa een vakantieparadijs gemaakt zal worden
Paris manages it COMP PRT to announce that of Mururoa a resort made will be
Paris manages it to announce that Mururoa will be made into a resort

b. Eltingh **presteert** het om vier games te winnen
Eltingh manages it COMP four games to win
Eltingh manages it to win four games

c. Maar men moet eens **ophouden** om tegen Maywood aan te schoppen
but one must once stop COMP against Maywood PRT to kick
But one should stop now with criticizing Maywood

3 Why add om?

There are two considerations that might explain why language users do sometimes include *om*, and sometimes do prefer to omit it: processing complexity and semantics. We discuss both potential explanations below.

A complementizer explicitly marks the beginning of an embedded infinitival clause, and as such can help to reduce processing complexity. Processing complexity has been used to explain the distribution of *that* in English finite complement clauses:

(12) a. I believe (that) we've pretty much summed it up
b. I know (that) the expectation for them, uhm, was to have sex...

Roland, Elman, and Ferreira (2006) observe that the verb governing the complement clause (CC) is important for predicting *that.* This in turn can be explained in terms of the probability that the governing verb selects for a CC: if a governing verb occurs with a CC often (i.e. of all occurrences of the verb, a high proportion is with a CC), the complementizer *that* will be omitted more easily (i.e. a small portion of all CC occurrences starts with *that*). Jaeger (2010) introduces the notion of *uniform information density*: "speakers prefer utterances that distribute information uniformly across the signal". The (Shannon) information of a word, I(word), is defined as its log transformed inverse probability, $-log\ p(word)$. Jaeger argues that speakers are sensitive to information density when choosing between producing a CC with or without *that.* The relevant information score for presence of a CC is approximated as $-log\ p(\text{CC}|matrix\ verb\ lemma)$, i.e. the probability that a given verbal governor occurs with a CC. As *that* helps to reduce processing complexity, uniform information density suggests that *that* will be used more often with verbs that rarely occur with a CC. Note that this type of explanation assumes that speakers have access to information about the frequency of words and their subcategorization frequencies.

One might argue that choice for the complementizer *om* in Dutch can be explained in a similar way. This explanation is plausible if *om* indeed helps to reduce (local) ambiguity. As *om* marks the beginning of a complement clause, it seems that in general, as in English, *om* at least helps to reduce local ambiguity. Jansen (1987) presents the following example, illustrating that sometimes *om* helps to resolve a global ambiguity:

(13) a. Ik beloof haar te opereren
I promise her to operate
I promise her to operate/ to operate her
b. Ik beloof om haar te opereren
I promise her to operate
c. Ik beloof haar om te opereren
I promise to operate her

In (13-a), the object pronoun could be a dependent of either the governing verb or the embedded verb. By adding *om* this ambiguity is resolved.

On the other hand, the fact that *om* can also be a preposition or particle, and as a complementizer can be used to introduce goal and purpose modifiers, can sometimes lead to the introduction of (local) ambiguity as well. Eerkens (2011), for instance, presents examples such as the following

(14) a. Hij belooft om de toren te beklimmen
He promises COMP the tower to climb
He promises to climb the tower
b. Hij belooft om de toren te wandelen
He promises around the tower to walk
He promises to walk around the tower

In (14-a), *om* is used as complementizer, whereas in (14-b) it is used as preposition.

Om is also a separable verbal prefix. In cases where the verb heading the TI occurs with *om* as prefix, an ambiguity may arise:

(15) a. De toren staat op het punt om te vallen
The tower stand on the point PREFIX/COMP to fall
b. *The tower is about to fall over*
c. *The tower is about to fall*

The verb *vallen* (*to fall*) may occur with *om* as verbal prefix, and the verbal governor *op het punt staan* (*be about to*) selects for an IC optionally introduced by *om*. Therefore, example (15-a) is ambiguous.

Finally, adding *om* potentially introduces an ambiguity between interpreting the clause as goal or purpose modifier clause or as complement.

(16) a. Ik raad mensen nooit aan om meteen te gaan slapen
I advise people never PART COMP immediately to go sleep
b. De arts raadt zonnebrandcrème aan om verbranding tegen te gaan
The doctor advises sun-screen PART COMP sunburn against to go

In (16-a), the verb *aanraden* (*to advise*) occurs with an indirect object and an OTI complement. In (16-b), the verb *aanraden* occurs with a direct object, and the OTI clause is a modifier. Note also that *om* is optional in (16-a) but not in (16-b).

The effect of *om* as disambiguator is therefore only limited. The situation is similar in English, as *that* can be used as complementizer but also as determiner or demonstrative. Thus, the complementizer in both languages does not by itself unambiguously mark the beginning of a complement clause in every conceivable context.

It is tempting to draw an analogy between English *that*-omission and Dutch *om*-omission, as both involve a complement clause governed optionally introduced by a complementizer. It should be noted, however, that there are also differences. First of all, finite complements have an explicit subject, but infinite complements do not. Thus, the complexity of the subject in the complement clause (shown to play a role for *that*-omission in Jaeger (2010)) will not play a role in the choice for *om*.

Second, the distance between the governor and the start of the complement clause has been shown to play a role in *that*-omission. While the same might be true for *om*-omission, it should be noted that in Dutch the position of the governing verb within the clause varies according to clause type. Finite verbs heading main clauses appear in first or second positio, whereas in all other cases, i.e. finite and infinite subordinate clauses, the verb occurs in clause final position. In particular, in subordinate clauses the verb follows NP-complements and adverbial modifiers, while PP-constituents may either precede or follow the verb. In main clauses it is therefore much more likely that other material intervenes between the verbal governor and the IC than in cases where the governor heads a subordinate clause.

It is hard to say how this interacts with the preference for using *om*. Subordinate clauses are a sign of syntactic complexity and thus might increase the preference for *om* as compared to sentences with a finite verbal governor in a main clause. On the other hand, in subordinate clauses, the IC always immediately follows the 'verbal complex' of the matrix clause. Consequently, one might argue that the (local) syntactic ambiguity at this point is low, and thus the preference for *om* should go down. We do not know which of the two suggestions is correct, but for Dutch it seems reasonable to include the type of the matrix clause (i.e. main or subordinate) as a factor as well.

If reducing syntactic complexity is the driving force for choosing *om*, we expect factors such as length of the IC, distance (in words) between governor and IC, matrix clause type (i.e. verb final or not), and the presence of other complements to play a role.

One might also argue for a semantic account. Purpose and goal infinitival modifier clauses obligatorily are introduced by the complementizer *om*. Some verbs that take *om* as complement express a meaning that makes the complement clause very close in meaning to a purpose or goal clause:

(17) a. De EU zal de komende jaren alles in het werk stellen om Oost-Europa te
The EU will the coming years everything in the work put COMP Eastern-Europe to
helpen
help
The EU will do everything it can in the coming year to help Eastern-Europe

b. Daarmee geef je mensen de tijd om psychisch aan het idee te wennen
With-that give you people the time COMP psychologically at the idea to get-used
That way you are giving people the time to get used to the idea psychologically

c. Smit zag zijn kans schoon te verzelfstandigen
Smit saw his chance clean to become-independent
Smith saw an opportunity to become independent

A semantic account would suggest that complement clauses that are close in meaning to a goal or purpose clause, will more likely be introduced by *om*. Conversely, ICs that have a meaning that is typical for complements of a given matrix verb, will usually not be introduced by *om*.

It is hard to determine whether an IC expresses a purpose or goal other than by manually classifying sentences, and even then, it might be hard for annotators to do this consistently. As an approximation we therefore estimate how much an IC is similar to a purpose clause by looking at the main verb in the IC. We assume that all OTI modifier clauses in our corpus express a goal or purpose. We use a statistical measure (pointwise mutual information) to find verbs that are typically used in purpose modifier clauses. If such a typical purpose verb occurs in an IC, we assume that this IC is similar to a purpose clause.

Another rough approximation of the 'complement'-hood of an IC is to determine whether the infinitival verb is typical for complements occurring with a given governor. In this case, we compute how strongly an embedded verb is associated with a governor. In general, we expect higher association scores between governors and verbs heading complements than between governors and verbs heading a modifier clause. This is because complements are 'selected' by the governor. Governors may impose certain semantic constraints on the kind of complements they can occur with. Such restrictions in general do not hold for modifiers. Thus, if a governor and a verb heading an IC are strongly associated, the complement-hood of this IC is strong.

4 Data

As corpus, we used an 80 million word subset[3] of the Twente Newspaper corpus (Ordelman et al., 2007). For comparison with spoken language, we used the Corpus of Spoken Dutch (Oostdijk, 2000). For computing association scores, we used the full Twente Newspaper corpus (500 million words).

[3] consisting of material from *Algemeen Dagblad* and *NRC Handelsblad*, 1994 and 1995

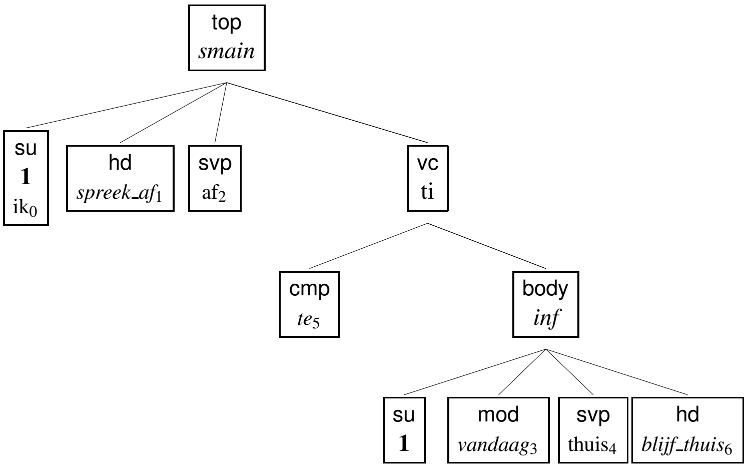

Figure 1: Dependency tree for *Ik spreek af vandaag thuis te blijven* (*I arrange to stay at home today*)

All corpora were parsed automatically using Alpino (van Noord, 2006). Alpino is the state-of-the art parser for Dutch, which produces highly accurate dependency trees. Using automatically parsed data has the advantage that it allows us to collect a large number of relevant examples quickly, including several features that might be relevant for predicting the distribution of TI vs. OTI ICs. Nevertheless, automatically parsed data tends to be more noisy than manually annotated data, due to parse errors. We took several measures to ensure that the amount of noise is kept to a minimum.

Initially, we selected all sentences containing a TI or OTI functioning as verbal complement, i.e. with grammatical relation label VC in the dependency graph output by the parser.[4] An example of a dependency tree for a VC/TI and VC/OTI is given in figure 1 and 2, respectively.

In our experiments, we used data for all and only governors that occur at least 10 times with a TI and at least 10 times with an OTI. We imposed this restriction to make sure that we are indeed considering examples where both forms are possible. We also filtered all cases where the governor (also) had a use as *cross-serial dependency* verb. An example is the verb *besluiten* (*to decide*):

(18) a. ...waarna hij zich blijvend in de VS **besloot** te vestigen
after-which he himself permanent in the US decided to stay
...after-which he decided to stay in the US permanently
b. ...waarna hij **besloot** (om) zich blijvend in de VS te vestigen

Example (18-a) exhibits cross-serial dependency word order where *om* is not possible. In (18-b), the IC is extraposed and *om* is possible. As the dependency structure of both cases is identical, it is hard to detect cross-serial cases automatically. To avoid confusion about the actual number of (extraposed, non cross-serial) TI cases, we decided not to include cases where the governor allows both word orders. We manually compiled a list of verbs that have this property.

Finally, we checked samples of extracted sentences for all remaining verbal governors. In those cases where the number of 'false positives' was too high, we discarded the data for this governor as well. False hits occur relatively often with certain verbs that do select for an IC but do not allow *om*. As the lexical specifications for verbs selecting an IC do not impose the requirement that an IC exclusively has to be TI or OTI, occurrences of such verbs with purpose modifier clauses are easily misanalysed as occurrences of an OTI complement.

A governor is a verbal lexical item, for which the Alpino dependency graph provides both a `stem` attribute and a `sense` attribute. The `stem` is the morphological base form of the verb. The value of `sense` is the base form plus subcategorized-for lexical material, such as prepositions that head a prepositional complement and fixed expressions.[5] We used `sense` as it allows us to distinguish more accurately the

[4] Subject TI and OTI clauses are rare and were ignored.

[5] Although the `sense` feature was introduced to make it easier to distinghuish between different *meanings* of a lexical item, it is at the moment only a placeholder for a system that would do proper word sense disambiguation.

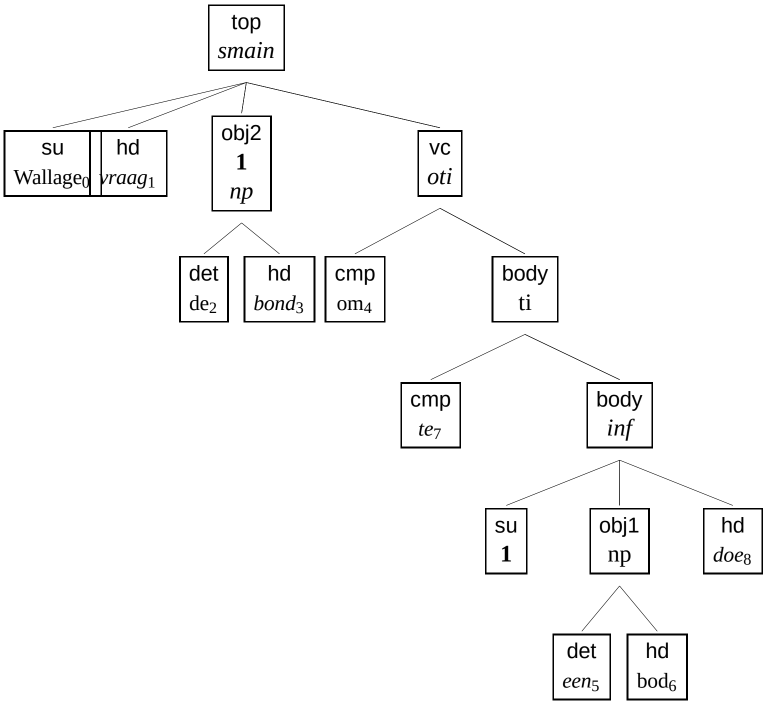

Figure 2: Dependency tree for *Wallage vroeg de bonden om een bod te doen* (*Wallage asked the unions to make a bid*)

use of a verb. For instance, the auxiliary verb *ben* (*to be*) is not included in our counts, but an idiomatic expression like *van plan ben* (*be planning to*) is.

From the newspaper corpus, we collected 49,077 relevant sentences, containing an IC and a verbal governor that met the frequency and grammatical properties described above. 11,682 cases contain *om* (23%). 95 different verbal governors occur in the data, with a zipfian frequency distribution, ranging from 9,287 (*besluit*, *decide*) to 26 (*beschouw*, *consider*).

5 Register and category of the governor

Before investigating our main research question, i.e. which factors influence the choice for *om* in constructions where the governor is a verb, we briefly address the question whether *om* as complementizer is indeed more frequent in spoken language, and the question how verbal governors compare with adjectival and nominal governors.

Jansen (1987) observes that *om* is used more frequently in spoken language than in (formal) written language. He compares a small data-set of spoken Dutch (containing 200 relevant ICs) with a manually collected set of examples in written language (containing 568 relevant cases). He finds that in written text 43% of the ICs are OTIs, whereas this is as high as 80% in spoken language.

In figure 3, we give the percentage of OTIs for a number of verbal governors occurring with an IC. We computed percentages for two corpora, the Corpus of Spoken Dutch (CGN, 10 million words) and our newspaper corpus (CLEF, 80 million words). The table only contains results for those verbs that occurred at least 10 times with both forms in both corpora for reasons explained in the previous section. The CGN percentage for the presence of *om* per verb is consistently higher than that for written, newspaper, text. This confirms the claim that *om* is more frequent in spoken language. However, it seems the contrast between spoken and written language is much less strong than in the samples studied by Jansen (1987).

Table 1 gives an overview of *ic* occurrences with verbs, adjectives, and nouns as governor. It shows that with verbal governors, *om* is omitted most often, while for nominal governors, this tendency is least strong. The histograms in figure 4 illustrate that these differences are not due to a few high frequent outliers. The majority of verbal governors has a preference for omitting *om* when occurring with an IC

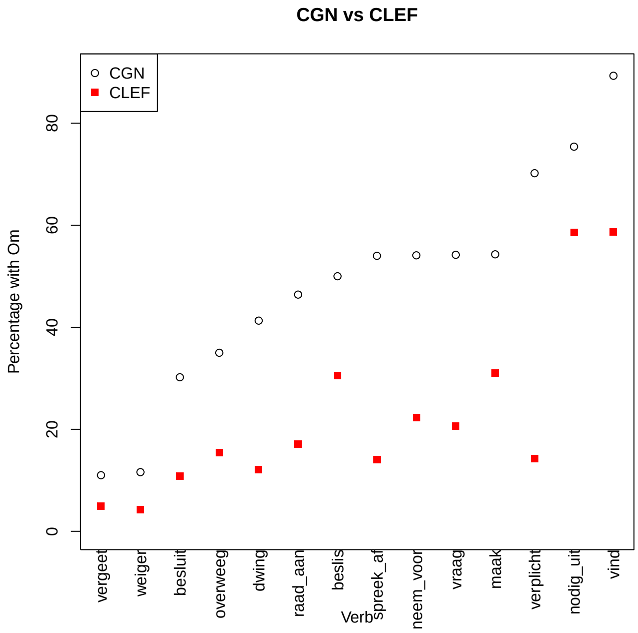

Figure 3: Relative percentage of OTI occurrences for verbs that occur at least 10 times with both an TI and OTI IC in CGN and CLEF.

Category	governors	count	%OTI
Verb	98	50,351	0.23
Adjective	24	16,370	0.41
Noun	123	64,816	0.61

Table 1: Number of different governors, total count, and percentage of OTI occurrences of verbal, adjectival, and nominal governors occuring with an IC.

while exactly the opposite is true for nominal governors. The histogram for adjectives is less informative because of the low number of selected adjectival governors (24).

6 Model

Given a sentence containing an IC, we want to predict whether the IC is realized as an OTI or TI on the basis of several features that might play a role in the choice between TI and OTI. Regression modelling is a technique that allows various predictors (both categorical or continuous) to be combined to make a prediction. Normally, the prediction of a regression model is continuous. Logistic regression is a technique for dealing with situations where the outcome is categorical, as is the case in predicting whether an IC will be realized as OTI or not.

The features $F_1..F_n$ that we consider relevant to predicting OTI are combined into a linear model:

$$\eta = \beta_0 + \beta_1 F_1 + ...\beta_n F_n$$

Note that depending on the variable values $F_1...F_n$ and the choice of weights $\beta_0...\beta_n$ the outcome of the predictor, η, can in principle be any real number. This outcome is compared to the actual proportion of the data with the given variable values that has outcome 1 by means of the logit link function:

$$\eta = logit(\textsc{oti}) = ln\frac{p(\textsc{oti})}{1 - p(\textsc{oti})} = ln\frac{p(\textsc{oti})}{p(\textsc{ti})}$$

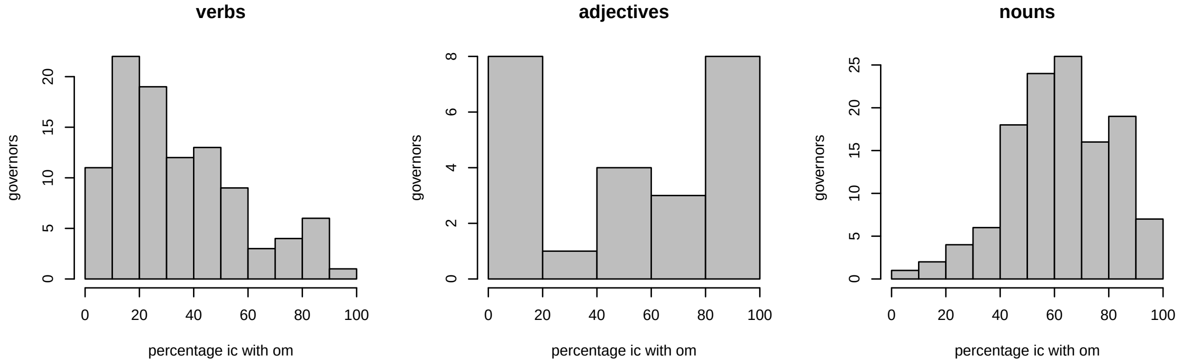

Figure 4: Histogram showing the number of governors per percentage of OTI complements for verbs, adjectives, and nouns.

Again, the outcome of this function can be any real number. See Levy (2011) (chapter 6) and Baayen (2008) (chapter 6) for more discussion.

As we will argue below, the distribution of OTIs versus TIs is strongly influenced by the governing verb. There are many different verbs selecting for ICs that optionally allow for *om*, and the probability that *om* appears differs strongly per verb. Furthermore, we are not dealing with a balanced set, but with data which shows the usual Zipfian distribution, i.e. the data is dominated by a few verbal governors that occur with high frequency and there is a long tail of verbal governors that occur less frequently. In such a situation it is typically hard to determine what the role of other properties of the sentence are that might play a role in the realization of the IC. Mixed-effects regression modeling is a technique that has recently become popular to deal with such situations (see Baayen (2008), chapter 7). Here we assume that the governing verbs form a random selection from a larger population (of IC taking verbs) and that these can influence both the intercept and slope of the fixed effects in the regression model.

7 Variables for predicting TI vs. OTI

In this section we present the various variables that we extract from the data to predict whether *om* is present in a particular sentence containing an IC.

7.1 The verbal governor as random effect

The histogram for verbs in figure 4 (left pane) suggests that there are large differences between individual verbal governors in the percentage of ICs that occurwith *om*. Clearly, the governor plays an important role in predicting *om*.

In our 80M word newspaper corpus, we find almost 50K sentences containing a IC governed by a verb that optionally allows *om*-insertion. 23% of the sentences contains *om*. There are 95 different verbal governors, 23 of which have a preference for OTI over TI, i.e. where the percentage of OTI of all IC occurrences is over 50%. 11 verbal governors occur with an OTI less than 10% of the time.

Although we consider the difference in preference for *om* as a random effect in the models below, one might still wonder whether these preferences do not follow from other properties of the lexical item. It is well known that frequency of lexical items can have an effect on processing. If *om* is used to reduce processing complexity, we expect OTIs to occur relatively more often with low frequent governors than with high frequent governors. Figure 5 (left pane) illustrates that such a correlation indeed exists. The y-axis represents the log frequency of the verbal stem in our 80M newspaper corpus, and the x-axis represents the ratio of OTI against TI occurrences with this verbal stem as governor in our dataset. It shows that verbs that occur with *om* relatively often, tend to be low frequent.

Another property that might play a role is the likelyhood that a given verbal governor occurs with an IC in the first place. Roland, Elman, and Ferreira (2006) and Jaeger (2010) observe a strong correlation

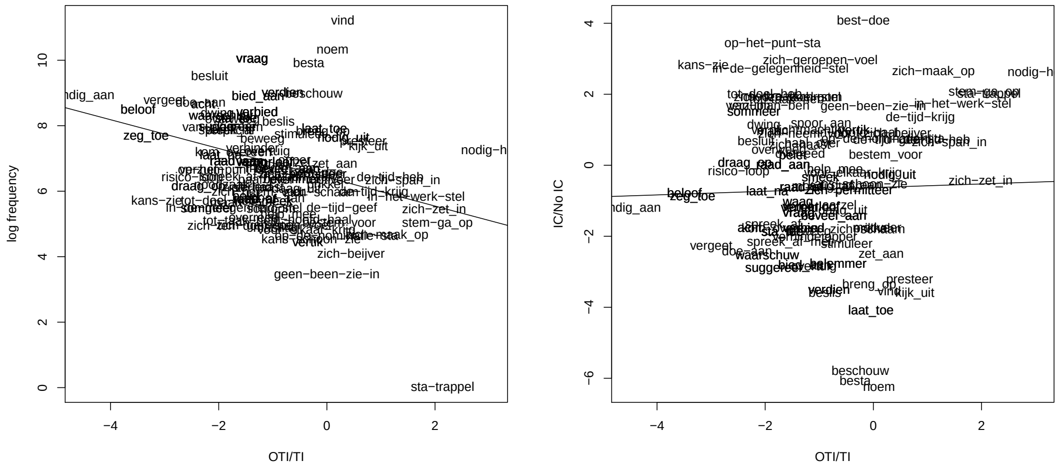

Figure 5: Overall log frequency of a verbal governor against ratio of OTI occurrence (left pane) and Ratio of overall IC over non IC occurrence of verbal governors against ratio of OTI occurrence (right pane)

between the possibility of *that* deletion in English complement clauses and the probability that a governor occurs with a finite clausal complement. That is, if a verbal governor tends to occur with a finite complement, the possibility of *that* deletion increases. If *om* reduces processing cost, we may expect a similar correlation between preference for an IC and preference for TI. Figure 5 (right pane) shows that such a correlation does not hold for Dutch *om* omission.

As described in section 4, we use the value of the SENSE attribute in the corpus to identify verbal governors. This has the advantage that it allows us to distinguish idiomatic uses of a verb, as collocational verbal expressions are usually reflected in the SENSE attribute. However, verbs can also have multiple valence frames. For instance, the verb *achten* (*to consider*) occurs with a direct object or with a predicate and direct object:

(19) a. Een burgemeester wordt geacht boven de partijen te staan
A mayor is considered above the parties to stand
A mayer is supposed to be neutral
b. Feyenoord acht het niet nodig een nieuwe directeur aan te stellen
Feyenoord considers it not necessary a new director PRT to appoint
Feyenoord considers it unnecessary to appoint a new director

Preferences for *om* can differ considerably between different uses of a verb. For the verb *achten*, for instance, this is 5% for occurences with just a direct object and 25% for cases where a predicate is also present. The annotation in the treebank registers the valence frame of the verb that was used in the parse for a given sentence. Thus, different uses of verbal governors can be further distinguished by their valence frame. If verbal governors are identified by a combination of the value of SENSE and SUBCAT (the attribute used to identify valence frames), 125 different verbal governors can be distinguished. Below, we will give results for both methods of identifying verbal governors.

7.2 Features for predicting complexity

In this section we describe a number of features of individual utterances containing an IC that can be seen as indicators of syntactic or processing complexity, and thus potentially play a role in the choice for inserting *om*.

7.2.1 Properties of the IC

The complementizer *om* marks the beginning of the IC. One might expect *om* to show up especially in those cases where the start of the IC is hard to recognize (locally) or where the sentence is just complex. Several features can be used as predictors for such a situation:

- Length of the IC. We count the number of words in the IC. (Obviously, the complementizer *om*, if present, is not included in this count).
- Relative position of the *te*-infinitive from the start of the IC (excluding *om*). We count the number of words.
- Syntactic category of the first constituent of the IC. We use four categories: *nominal, adverbial, verbal, other.*

Note that although we include length of the IC as an indicator of complexity, it is not clear that this is directly relevant for the decision to insert *om*, as length by itself does not make it harder to detect the start of an IC. The relative position of the head of the IC, i.e. the *te*-infinitival form, might be a better predictor for the choice between OTI and TI. As *te* is a clear morphological marker of an IC, it helps to reduce syntactic complexity if it is near the beginning of the IC. Finally, the complexity of the subject in the embedded clause has been shown to play a role in *that*-deletion in English. As we are dealing with infinitival clauses, we look at the first constituent instead. Also, instead of measuring complexity (as the number of words), we look at the category of the first constituent. The intuition is that adverbs and PPs in particular (categorized as *other*) tend to introduce local ambiguity that can be resolved by inserting *om*.

7.2.2 Adjacency and distance

The position of the verb and its complements is due to some variation in Dutch. Dutch is a verb-final language, where the verb occurs in first or second position only if it is finite and heads a main clause or a yes/no question. In all other cases, the verb is part of the so-called *verb cluster*, which occupies a position at the end of the sentence, following the subject, direct object, predicative complements, fixed expressions, and adverbials. Prepositional complements may either precede or follow the verb cluster. Verbal complements, such as ICs, always follow the verb cluster. Thus, there are a number of situations, especially in subordinate clauses and in cases where the governor is a non-finite verb, where the governor and beginning of the IC are close and the beginning of the IC is relatively clear.

Table 2a lists the percentage of ICs for various distances between the governor and IC. ICs immediately following the governor have *om* in only in 20% of the cases, whereas for ICs at least two words away, the percentage of *om* is 28% or higher. Surprisingly, the lowest percentage of IC use is found with a distance of 1, i.e. with a single word intervening between the governor and the IC. We speculate that this is due to some peculiarities of Dutch word order, but at the moment have no clear explanation for this fact.

As an alternative for distance, we can also look at the category of the clause headed by the verbal governor directly. If this is a finite main clause, we expect the percentage of *om* to be higher. Table 2b shows that our expectations are confirmed only to a certain extent. The highest percentage of OTIs is indeed found in main clauses, but it is only slightly higher than that for cases where the governor is infinitival or heading a (finite) subordinate clause. The lowest percentage of OTIs is found with participial verbal governors.

We used the following features for distance, adjacency, and clause type:

- Distance. The log of the number of words between the verbal governor and the start of the IC plus one (to avoid distances of 0).
- Adjacency. A binary feature that is set to 1 if the beginning of the IC is at most 1 word away from the verbal governor.
- Clause Type. A categorical feature for the clause type of the governor, with values SMAIN, SV1, INF, SSUB, and PPART.

distance	TI	OTI	% OTI
0	23,382	5,780	19.8
1	4,843	752	13.4
2	3,437	1,387	28.7
3	2,884	1,196	29.3
4	1,367	868	38.8
5	846	526	38.3
6	512	342	40.0
7	344	219	38.8
8	244	139	36.2
9	167	109	39.4
10	121	73	37.6
≥ 11	427	226	34.6

(a) %OTI for various distances between between governor and IC

Clause type	TI	OTI	% OTI
SMAIN	14,941	5,357	26.4
INF	4,342	1,552	26.3
SSUB	5,045	1,609	24.2
SV1	523	152	22.6
PPART	13,723	2,947	17.7
average	38,574	11,617	23.2

(b) %OTI for different clause types

Table 2: Percentage OTI

7.2.3 Presence of other complements

If the governing verb selects other complements besides the IC, it might be harder to detect the start of the IC, and thus the probability of *om* could be higher. Table 3 gives results for the percentage of OTIs given the presence of various types of other complements. For direct objects (OBJ1), we distinguish between pronominal and other (full NP) objects. For indirect objects (OBJ2), we distinguish between NP and PP indirect objects. The other cases refer to the presence of an inherent reflexive (REFL), a predicative complement (PREDC), expletive *it*, or a fixed phrase (including seperable verb particles) (FIXED).

The probability of OTI goes up rather strongly if an inherent reflexive, predicative complement, or expletive is present. Expletives are interesting, as they can be seen as placeholder for the IC. The majority of these cases occur with the governor *vinden*, which also selects for a predicative complement.

(20) a. Weinig onderdanen durven het aan om haar een succes te noemen
few nationals dare it PRT CMP her a success to call
Few citizens dare to call her a success

b. Hij vindt het spannend om te zien dat je openstaat voor je eigen impulsen
he finds it exciting CMP to see that you are-open to your own impulses
he finds it exiting to see that you are open to your own impulses

Note that their is a rather strong correlation between using the valence frame of the governor for identifying the governor and looking at the presence of certain complements in the sentence. In models that use the verbal stem and valence to identify governors, we will therefore not use these features. On the other hand, in models that do not use the valence frame, an advantage of looking at individual complements might be that they are less grammar dependent than the various subcategorization frames used by Alpino, and also, that they register the presence of individual complements instead of their combination.

In our experiments, we included categorical features for the complements listed in table 3, with the given values.

7.3 Semantic Features

Given the fact that OTI clauses can both function as complements and as modifiers expressing a goal or purpose, one might expect that *om* is present more often in complement clauses that are similar to such modifying purpose clauses. We use two features that are inspired by this idea.

7.3.1 Association strength between verbs

It is well known that the distribution of words and phrases in a corpus is not random, but follows certain patterns. In particular, words and phrases that share some semantic property, tend to co-occur

complement	TI	OTI	% OTI	
obj1	pron	919	258	21.9
	np	8,089	2,462	23.3
	no	29,566	8,897	23.1
obj2	np	5,372	1,305	19.5
	pp	143	47	24.7
	no	33,059	10,265	23.6
refl	yes	928	781	45.6
	no	37,646	10,836	22.3
predc	yes	1,477	2,607	63.8
	no	37,097	9,010	19.5
expletive	yes	1,931	1,955	50.3
	no	36,643	9,662	20.8
fixed	yes	6,170	2,623	29.8
	no	32,404	8,994	21.7
average	38,574	11,617	23.2	

Table 3: Counts and percentages of TI and OTI occurrence given the presence of various other complements.

more often than the frequency of the individual words or phrases would suggest. Distributional models of semantics determine the association strength between pairs of words, stems, phrases, and other linguistic units by means of statistical measures based on the relative frequency of occurrence of the individual units. For instance, the verb *eat* will occur relatively often with a subject that denotes an animate entity, and with an object that is edible.

We can use this technique also to measure how much a verbal governor is associated with the verbal head of its IC. The assumption is that, if the two are strongly associated, (the event described by) the IC is typical for this governor, and thus a more likely complement than if the association strength between the two is low. In such cases, the need to use *om* might be less.

The association score between a governor and the verbal head of its IC is computed as the pointwise mutual information (Church and Hanks, 1990) between the two (where f(W) is the relative frequency of W in the corpus:

$$\text{pmi(Gov,IC-head)} = ln(\frac{f(\text{Governor,IC-head})}{f(\text{Governor}) * f(\text{IC-head})}$$

7.3.2 Association strength between verb and purpose clause

Some verbs will occur in modifier OTI purpose clauses much more often than others. Such verbs express an event that is typical for a goal or purpose. If an IC is headed by such a verb, its semantics shares some resemblance with a purpose clause. We expect the probability of *om* to go up in such cases.

Again, we use pointwise mutual information to measure the association between the modifier purpose clause and the verbal head:

$$\text{pmi(PurposeClause,Head)} = ln(\frac{f(\text{PurposeClause,Head})}{f(\text{PurposeClause}) * f(\text{Head})}$$

To obtain the relevant statistics, we assume that all OTI constituents in the corpus that have the dependency relation MOD are indeed purpose or goal clauses.

Verbs and verbal expressions that are ranked high according to this measure are for instance: *kracht bij zetten (to emphasize), erger voorkomen (to limit the damage), het hoofd bieden (aan) (to cope with), voorkomen (to prevent), promoten (to promote), beschermen tegen (to protect against).*

$$Model = outcome \sim dist + length + te + het(1 + dist + length + te + het|stem)$$

	effect	std. err	significance
(Intercept)	-0.90	0.20	***
dist	0.13	0.05	*
length	-0.13	0.03	***
te	0.27	0.04	***
het	0.38	0.19	*

Table 4: Best model using verbal stem of the governor as random effect and various syntactic complexity features as fixed effects.

8 Experiments

In this section, we describe various experiments to determine which properties influence the choice for *om*, and how these properties interact. We used R (R Development Core Team, 2011) and *lme4* (Bates, Maechler, and Bolker, 2011) to perform a linear mixed effects analysis, where verbs are random effects.

8.1 Complexity features in main clauses

We start with the situation that is perhaps most similar to English *that*-deletion, i.e. the distribution of *om* where the governing verb is finite and heading a main clause. In such cases, the governing verb is in second position in the sentence, while the IC is clause final. There are 19.862 relevant cases in our dataset, containing 94 different governors. Features that have been shown to play a role in English *that*-deletion are likely to play a role in this case as well.

We use the verb as random effect, where a verb is identified by its stem. As fixed effects, we used various features that might be indicators of syntactic or processing complexity.

The best model according to these assumptions (given in Table 4) includes distance between governor and IC (*dist*), length of the TI, distance between start of the TI and the *te*-infinitive verb (*te*), length of the first constituent inside the TI, presence of expletive *het* to signal the presence of a TI (*het*). Numeric features were log-normalized and centered.

The negative intercept follows from the fact that the majority of cases do not have *om*. Longer distances between governor and IC, and between the start of the IC and the *te*-infinitive verb, as well as the presence of expletive *het* all increase the likelihood of *om*. The overall length of the IC has a small negative effect.

An anova test shows that the model improves significantly over a baseline model using only sense as random effect (Model AIC[6] = 15,716, Baseline AIC = 16,001, $\chi^2 = 288.35, p < 0.001$). Addition of various other potential features such as length and syntactic category of the first constituent of the IC, frequency of the head of the TI, and presence of other syntactic dependents in the matrix clause (direct object, predicative phrase, reflexive, prepositional complement) did not improve the model significantly.

8.2 Valence

Eventhough we found that including features for the presence of various dependents of the governing verb as fixed effects did not improve the model, one might still argue that valence can help to disambiguate the specific use of the governing verb in a given sentence. Thus, instead of using the stem of the verbal governor to identify governors, one might also try a model where a combination of stem and valence frame is used to identify the governor. We used the same model as above, except that the random effect is now the combination of stem and valence frame of the governor. Whereas there are 94 different stems used in the previous model, there are now 125 different combinations of stem and frame. We obtain a slightly higher AIC score than for the model using stem only, and an anova test shows that the model

[6] The Akaike Information Criterion is a measure for model fit based on Information Theory. Lower values indicate better model fit.

	effect	std. err	significance
(Intercept)	-0.98	0.13	***
dist	0.15	0.02	***
length	-0.10	0.02	***
te	0.20	0.02	***
het	0.49	0.12	***

Table 5: Fixed effects for the complete dataset using syntactic features.

without valence frames performs slightly better. We conclude that adding valence information, either as fixed effects for the individual syntactic dependents or as random effect, does not significantly contribute to the task of predicting the presence of *om*.

Out of curiousity, we also investigated a model using only the frame to distinguish between governing verbs. In such a model, all verbs selecting for, say, a direct object and a TI, are grouped. On the other hand, the data for verbs that occur with different valence frames, will be distributed over these frames. There are 20 different frames in our dataset. This model performs considerably worse than the model from the previous section (frame model AIC = 18,897, $\chi^2 = 3,180, p < 0.001$). We conclude that lexical properties of the governor are essential for predicting presence of *om* and that these lexical properties do not follow from valence information only.

8.3 Word Order

In this section, we consider the complete dataset, i.e. also including cases where the governing verb is nonfinite or where the governor heads a subordinate clause. There are 49,077 cases in this set and 95 different verbal governors.

Using the same model as for main clauses, we get the result given in table 5. The model outperforms the baseline (using only the governing verb as random effect) significantly ($\chi^2 = 549.88, p < 0.001$, Model AIC = 40,087, baseline AIC = 40,601).

As we are now dealing with cases where the governing clause can be either a main clause (with the governing verb in first or second position) or a subordinate or non-finite clause (with the verb in a position following most other dependents and seperated from the IC only by other verbs in the so-called verb cluster), one might expect a feature signalling clause type to be significant. However, in our experiments, we found that including a categorical feature for clause type was in general not significant as soon as the feature measuring distance between governing verb and IC was also included. Of course, the two will be highly correlated (with longer distances being exclusively found in main clauses) and clause type by itself apparently does not contribute over the information encoded in the distance feature.

8.4 Semantics

The historical development of *te*-infinitive complements headed by *om* suggests that these might occur more often with complements that are semantically similar to purpose or goal modifier clauses. On the other hand, typical complements of a verb are not confused easily with modifiers, and thus might have less preference for *om*. To measure these intuitions, we use two features based on pointwise mutual information, as explained in section 7.3.

A model for all data that uses only these two features as fixed effect is given in table 6. We see that the model confirms our expectation. If a TI is headed by a verb that typically occurs in purpose/goal modifier clauses, the likelihood of *om* goes up, whereas if the verb heading the TI co-occurs with the given governor often, the likelyhood of *om* goes down. The model outperforms the baseline (using only the random effect) significantly ($\chi^2 = 181.64, p < 0.001$, Model AIC = 40,433, baseline AIC = 40,601).

The model does not perform as well as the model using features inspired by syntactic and processing complexity considerations. Thus, complexity seems to play a more dominant role in the choice for *om* than semantics. ($\chi^2 =, p < 0.001$, semantic model AIC = 40,433, complexity model AIC = 40,087).

A model using both complexity features and semantic features does perform better than the model using complexity features only.($\chi^2 = 144.27, p < 0.001$, complexity + semantics model AIC = 39,973). The integrated model has a concordance (C) score of 0.809, which indicates that the model has modest

$$Model = outcome \sim complement + purpose + (1 + complement + purpose|stem)$$

	effect	std. err	significance
(Intercept)	-0.85	0.13	***
complement	- 0.07	0.02	***
purpose	0.11	0.02	***

Table 6: Model and fixed effects for the complete dataset using semantic features.

predictive qualities.[7] We conclude that complexity and semantic factors both influence the choice for *om*.

9 Conclusions

In this paper, we have investigated the distribution of the complementizer *om* in *te*-infinitive complement clauses in Dutch. Using a large automatically parsed corpus, we collected almost 50.000 instances of such clauses. It is clear from our data that the verb that selects for the *te*-infinitive influences the likelihood of *om* significantly. Given the strong influence of this lexical feature, we decided to use a mixed effects model, where the verb is used as random effect.

Following similar investigations on the distribution of the optional complementizer *that* in English complement clauses, we found that features that are indicative of processing or syntactic complexity do play a significant role. Features that directly register the presence of certain other dependents in the matrix clause do not play a significant role. This is true both in models that add categorical features for individual dependents as fixed effects, as well as in models where valence frames (in combination with the stem of the verb) are used as random effect.

We also found that semantic features that measure the similarity of the *te*-infinitive to typical complements for the given governor and to typical purpose or goal modifer clauses, play a significant role, although their effect is smaller than the 'complexity' features. A combination of 'complexity' and 'semantic' features gives rise to the best model.

We see a number of ways in which this work could be extended in future work. First of all, it would be interesting to use data from manually corrected treebanks. Although in general the accuracy of the automatically parsed data is quite high, and we took several measures to remove problematic cases, it might still be the case that manually corrected data gives rise to clearer models. On the other hand, manually corrected treebanks will always be an order of magnitude smaller than what is available as automatically annotated data, so careful combination of statistics from automatic and manually corrected treebanks may be required.

Another important question is to what extent these results are genre and medium dependent. Initial experiments showed that in our data, even the source of data (i.e. the newspaper from which it originated) gives small but significant effects. This shows that edited material may not be the most ideal source for this kind of research. Data from spoken language is especially interesting in this respect, as we already showed that *om* is more frequent in spoken language, and spoken language in general is more spontaneous than (edited) written material.

Finally, we would like to investigate the relationship between processing complexity and the use of optional function words such as *om* more directly. For instance, given a formal grammar for the data, as implemented in our automatic parser for Dutch, one could systematically investigate the number of parses for sentences with and without *om* to determine whether *om* indeed helps to reduce global ambiguity. In addition, one could investigate local ambiguity, i.e. the number of partial parses at the point where the IC starts, and see whether adding *om* helps to reduce local ambiguity.

[7] The concordance scores measures for all pairs of a negative (TI) outcome and a positive (OTI), how often the model predicts a hihger log-odds for the positive case. We used the function `somers2` from the Hmisc package.

References

Baayen, R.H. 2008. *Analyzing Linguistic Data.* Cambridge University Press.

Bates, Douglas, Martin Maechler, and Ben Bolker, 2011. *lme4: Linear mixed-effects models using S4 classes.* R package version 0.999375-39.

Church, Kenneth Ward and Patrick Hanks. 1990. Word association norms, mutual information & lexicography. *Computational Linguistics*, 16(1):22–29.

Eerkens, Muriël. 2011. Ombiguity: the effects of priming a structure on the subsequent processing of a different structure. BA Thesis, Utrecht University.

Ferreira, V.S. and G.S. Dell. 2000. Effect of ambiguity and lexical availability on syntactic and lexical production. *Cognitive Psychology*, 40(4):296–340.

Geerts, G., W. Haeseryn, J. de Rooij, and M.C. van den Toorn. 1984. *Algemene Nederlandse Spraakkunst.* Groningen: Wolters-Noordhoff.

Hawkins, J.A. 2002. Symmetries and asymmetries: their grammar, typology and parsing. *Theoretical Linguistics*, 28(2):95–150.

IJbema, Aniek. 2002. *Grammaticalization and Infinitival Complements in Dutch.* Ph.D. thesis, Leiden University.

Jaeger, Florian. 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, 61:23–62.

Jansen, F. 1987. Omtrent de om-trend. *Spektator*, 17:83–98.

Levy, Roger. 2011. Probabilistic models in the study of language.

Oostdijk, Nelleke. 2000. The Spoken Dutch Corpus: Overview and first evaluation. In *Proceedings of LREC 2000*, pages 887–894.

Ordelman, Roeland, Franciska de Jong, Arjan van Hessen, and Hendri Hondorp. 2007. Twnc: a multifaceted Dutch news corpus. *ELRA Newsletter*, 12(3/4):4–7.

R Development Core Team, 2011. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Roland, D., J.L. Elman, and V.S. Ferreira. 2006. Why is that? structural prediction and ambiguity resolution in a very large corpus of English sentences. *Cognition*, 98(3):245–272.

van Noord, Gertjan. 2006. At last parsing is now operational. In Piet Mertens, Cedrick Fairon, Anne Dister, and Patrick Watrin, editors, *TALN06. Verbum Ex Machina. Actes de la 13e conference sur le traitement automatique des langues naturelles.* pages 20–42.

Wasow, T., T.F. Jaeger, and D.M. Orr. 2011. Lexical variation in relativizer frequency. In H. Wiese and H. Simon, editors, *Proceedings of the workshop on Expecting the Unexpected: Exceptions in Grammar.* Mouton De Gruyter.